



future internet

Data Science and Knowledge Discovery

Edited by
Filipe Portela

Printed Edition of the Special Issue Published in *Future Internet*

Data Science and Knowledge Discovery

Data Science and Knowledge Discovery

Editor

Filipe Portela

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin



Editor

Filipe Portela
Information Systems
University of Minho
Guimarães
Portugal

Editorial Office

MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Future Internet* (ISSN 1999-5903) (available at: www.mdpi.com/journal/futureinternet/special_issues/DS_KD).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, Volume Number, Page Range.

ISBN 978-3-0365-4316-1 (Hbk)

ISBN 978-3-0365-4315-4 (PDF)

Cover image courtesy of Filipe Portela

© 2022 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons license CC BY-NC-ND.

Contents

About the Editor	vii
Preface to "Data Science and Knowledge Discovery"	ix
Filipe Portela	
Data Science and Knowledge Discovery	
Reprinted from: <i>Future Internet</i> 2021 , 13, 178, doi:10.3390/fi13070178	1
Ana Teresa Ferreira, Carlos Fernandes, José Vieira and Filipe Portela	
Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients	
Reprinted from: <i>Future Internet</i> 2021 , 13, 102, doi:10.3390/fi13040102	5
Albert Weichselbraun, Philipp Kuntschik, Vincenzo Francolino, Mirco Saner, Urs Dahinden and Vinzenz Wyss	
Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities	
Reprinted from: <i>Future Internet</i> 2021 , 13, 59, doi:10.3390/fi13030059	21
Nuno Marques da Costa, Nelson Mileu and André Alves	
Dashboard COMPRI_MCOMPRI_MOV: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal	
Reprinted from: <i>Future Internet</i> 2021 , 13, 45, doi:10.3390/fi13020045	43
Aleksandr Romanov, Anna Kurtukova, Alexander Shelupanov, Anastasia Fedotova and Valery Goncharov	
Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks	
Reprinted from: <i>Future Internet</i> 2020 , 13, 3, doi:10.3390/fi13010003	61
Jing Wang, ZhongCheng Wu, Fang Li and Jun Zhang	
A Data Augmentation Approach to Distracted Driving Detection	
Reprinted from: <i>Future Internet</i> 2020 , 13, 1, doi:10.3390/fi13010001	77
Laith T. Khrais	
Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce	
Reprinted from: <i>Future Internet</i> 2020 , 12, 226, doi:10.3390/fi12120226	89
Piotr Artiemjew, Lada Rudikova and Oleg Myslivets	
About Rule-Based Systems: Single Database Queries for Decision Making	
Reprinted from: <i>Future Internet</i> 2020 , 12, 212, doi:10.3390/fi12120212	103
Sook-Ling Chua, Lee Kien Foo and Hans W. Guesgen	
Predicting Activities of Daily Living with Spatio-Temporal Information	
Reprinted from: <i>Future Internet</i> 2020 , 12, 214, doi:10.3390/fi12120214	117
José Paulo Lousado and Sandra Antunes	
Monitoring and Support for Elderly People Using LoRa Communication Technologies: IoT Concepts and Applications	
Reprinted from: <i>Future Internet</i> 2020 , 12, 206, doi:10.3390/fi12110206	131
Mikolaj Karpinski, Svitlana Kuznichenko, Nadiia Kazakova, Oleksii Frazee-Frazenko and Daniel Jancarczyk	
Geospatial Assessment of the Territorial Road Network by Fractal Method	
Reprinted from: <i>Future Internet</i> 2020 , 12, 201, doi:10.3390/fi12110201	161

Alan Ponce and Raul Alberto Ponce Rodriguez An Analysis of the Supply of Open Government Data Reprinted from: <i>Future Internet</i> 2020 , 12, 186, doi:10.3390/fi12110186	175
Antonio Maria Rinaldi, Cristiano Russo and Cristian Tommasino A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features Reprinted from: <i>Future Internet</i> 2020 , 12, 183, doi:10.3390/fi12110183	193
Christian Scheel, Francesca Fallucchi and Ernesto William De Luca Visualization, Interaction and Analysis of Heterogeneous Textbook Resources Reprinted from: <i>Future Internet</i> 2020 , 12, 176, doi:10.3390/fi12100176	213
Theodora A. Maniou and Andreas Veglis Employing a Chatbot for News Dissemination during Crisis: Design, Implementation and Evaluation Reprinted from: <i>Future Internet</i> 2020 , 12, 109, doi:10.3390/fi12070109	229

About the Editor

Filipe Portela

Filipe Portela holds a PhD in Information Systems and Technologies. He is an integrated researcher at Research Centre ALGORITMI, where he developed his post-doctoral research work on the topic “Pervasive Intelligent Decision Support Systems”. His research started in the INTCare R&D project (Intensive Medicine area) and then extended to other areas like education, public administration, industry and smart cities. He already has many relevant indexed publications on the main research topics: Knowledge Discovery, Data Science, Gamification, Intelligent Systems and Pervasive Data. He is also (co) organiser of several conferences and workshops, (co) editor of journals and books and reviewer of many indexed journals, books and conferences on these topics. Currently, he is also an Invited Assistant Professor in the Information Systems Department, School of Engineering, University of Minho, Portugal, where he supervises several master’s students in the areas mentioned above. Filipe Portela founded IOTech - Innovation on Technology in 2018, where he is the Chief Executive Officer (CEO) and Chief Innovation and Research Officer (CIRO). He is transferring and applying their scientific knowledge to benefit the citizens and companies.

Preface to “Data Science and Knowledge Discovery”

This book shows a set of emerging topics in Data Science and Knowledge Discovery. This book also presents works using different datasets like Covid-19, e-commerce, text, driving or spatial. This book is essential for anyone (students, professors, researchers, decision-makers) who want to know more about this area, see new findings, and see how to use data science to support the decision process. It can be helpful to open new windows of knowledge or research opportunities in an even more significant area.

Filipe Portela

Editor

Data Science and Knowledge Discovery

Filipe Portela ^{1,2} 

¹ Algoritmi Research Centre, University of Minho, 4800-058 Guimarães, Portugal; cfp@dsi.uminho.pt

² IOTECHE—Innovation on Technology, 4785-588 Trofa, Portugal

Abstract: Nowadays, Data Science (DS) is gaining a relevant impact on the community. The most recent developments in Computer Science, such as advances in Machine and Deep Learning, Big Data, Knowledge Discovery, and Data Analytics, have triggered the development of several innovative solutions (e.g., approaches, methods, models, or paradigms). It is a trending topic with many application possibilities and motivates the researcher to conduct experiments in these most diverse areas. This issue created an opportunity to expose some of the most relevant achievements in the Knowledge Discovery and Data Science field and contribute to such subjects as Health, Smart Homes, Social Humanities, Government, among others. The relevance of this field can be easily observed by its current achieved numbers: thirteen research articles, one technical note, and forty-six authors from fifteen nationalities.

1. Introduction

The importance and impact of Data Science (DS) in the decision process are significantly increasing. DS is an interdisciplinary field that combines various areas, including Computer Science, Machine Learning, Math and Statistics, domain/business knowledge, software development, and traditional research. DS applies scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data as a research topic.

Knowledge Discovery (KD) is the basis of Data Science and consists of creating knowledge from structured and unstructured sources (e.g., text, data, and images). The output needs to be in a readable and interpretable format. It must represent knowledge in a manner that facilitates inferencing. This new trend is being explored in several areas, such as education, health, accounting, energy, and public administration. In this context, this Special Issue arises as an excellent opportunity to provide scientific knowledge and disseminate the findings and achievements through several communities.

This Special Issue discusses this trending topic and presents innovative solutions to show the importance of Data Science and Knowledge Discovery to researchers, managers, industry, society, and other communities. Through invited and open call submissions, a total of fourteen excellent articles have been accepted, following a rigorous review process that required a minimum of three reviews and at least one revision round for each paper.

2. Contributions

The first paper, written by Theodora A. Maniou and Andreas Veglis [1] and entitled Employing a Chatbot for News Dissemination during Crisis: Design, Implementation, and Evaluation, presents some benefits of using chatbots by journalists and media professionals. It shows the advantages of implementing chatbots in news platforms during a crisis when the audience's need for timely and accurate information rapidly increases. This study was evaluated using two metrics: the technical effort of creating a functional and robust news chatbot and, the second, users' perception regarding the appropriation of this news chatbot. The participants involved in the case study agreed that the COVINFO Reporter's accessibility was very good, and they experienced no problems navigating the chatbot.

Citation: Portela, F. Data Science and Knowledge Discovery. *Future Internet* **2021**, *13*, 178. <https://doi.org/10.3390/fi13070178>

Received: 27 May 2021

Accepted: 5 July 2021

Published: 7 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

The second paper, entitled Visualization, Interaction and Analysis of Heterogeneous Textbook Resources, and written by Christian Scheel, Francesca Fallucchi, and Ernesto William De Luca [2], proposes a Component Metadata Infrastructure (CMDI)-based approach. The authors used this approach for data rescue and reuse, where data is retroactively joined into one repository, minimizing future research projects' implementation efforts. While the data is precious, it cannot be used by any service—except the prepared tool. With this approach, the authors want to increase data understanding, sustainability and reusability, and reduce data silos.

Antonio Maria Rinaldi, Cristiano Russo, and Cristian Tommasino wrote the third paper of this issue: Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features [3]. In this work, the authors analyzed studies in the semantic research field based on ontologies. They considered that, although modern search engines provide visual queries, it is not easy to find systems that allow searching from a particular domain of interest and that perform such searches by combining text and visual questions. Then, the authors proposed a novel approach for semantic image retrieval that included a possible combination with multimedia document analysis. This paper presents the method developed and several results to show its performance compared with the state of the art.

Alan Ponce and Raul Alberto Ponce Rodriguez wrote the fourth paper, an Analysis of the Supply of Open Government Data [4]. This work presents an analysis based on the index of the release of open government data, published in 2016 by the Open Knowledge Foundation, which shows a significant variability in the country's supply of open data. The authors used several linear regression models to explain the cross-country differences. This work provides evidence that the country's civil liberties, government transparency, quality of democracy, efficiency of government intervention, and economies of scale in the provision of public goods, as well as the size of the economy, are the most statistically important reasons for differences in the supply of open government data.

The following paper, the fifth one, is entitled Geospatial Assessment of the Territorial Road Network by Fractal Method [5] and was written by Mikolaj Karpinski, Svitlana Kuznichenko, Nadiia Kazakova, Oleksii Frazee-Frazenko, and Daniel Jancarczyk. This paper proposes an approach to the geospatial assessment of a territorial road network based on the fractal theory. The method allows calculation of the fractal dimension based on a combination of box-counting and GIS analysis. The authors created a geoprocessing script tool for the GIS software system ESRI ArcGIS 10.7 using the spatial pattern of the transport network of the Ukraine territory and other countries of the world. The study results help to better understand the different aspects of the development of transport networks, their changes over time, and the impact on the socioeconomic indicators of urban development.

The sixth paper was written by José Paulo Lousado and Sandra Antunes and is entitled S. Monitoring and Support for Elderly People Using LoRa Communication Technologies: IoT Concepts and Applications [6]. This paper is the second article motivated by the SARS-CoV-2 virus (COVID-19) and aims to show an implementation of low-cost technologies, which make it possible to answer a fundamental question: how can near real-time monitoring and follow-up of the elderly and their health conditions, as well as their homes, especially for those living in isolated and remote areas, be provided within their care and protect them from risky events? The proposed system uses low-cost devices for communication and data processing, supported by Long-Range (LoRa) technology, and incorporates various sensors, both personal and in residence. It allows family members, neighbors, and authorized entities (including security forces) to have access to the health condition of system users and the habitability of their homes, as well as their urgent needs. This article shows that it is possible to implement sensor networks to monitor the elderly using the LoRa gateway and other low-cost infrastructures.

The seventh publication, entitled About Rule-Based Systems: Single Database Queries for Decision Making, is a technical note written by Piotr Artiemjew, Lada Rudikova, and Oleg Myslivets [7]. It explores the implementation of artificial intelligence systems for

manipulating data and the surrounding world in a more complex way. In this work, the authors addressed the possibility of placing the rule-based learned model of decision support in a SQL database environment. They propose a universal solution for any IF-THEN rule induction algorithm to place the previously trained model in the database and apply it by employing single queries.

Sook-Ling Chua, Lee Kien Foo, and Hans W. Guesgen wrote the eighth paper—Predicting Activities of Daily Living with Spatio-Temporal Information [8]. This paper is framed in smart homes and shows the importance of having spatial and temporal information for reasoning. The authors created a method for predicting user activities given the spatial and temporal information and explained how it could be represented for activity recognition. The method was evaluated using three publicly available smart-home datasets and achieved an average accuracy of more than 81%.

The following article, the ninth one, addresses the Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce and was written by Laith T. Khrais [9]. This article explores the use of AI in e-commerce, where ethical soundness is a contentious issue, especially regarding the concept of explainability. The study adopted the use of word cloud analysis, voyance analysis, and concordance analysis to gain a detailed understanding of the idea of explainability as has been utilized by researchers in the context of AI. Motivated by a corpus analysis, the authors formulated the Explainable Artificial Intelligence (XAI) model that provides insights into the decision points, variables, and data used to produce recommendations. This study also suggests that the Machine Learning models should be improved by making them interpretable and comprehensible, and allowing them to deploy explainable XAI systems.

Jing Wang, Zhong Cheng Wu, Fang Li, and Jun Zhang present the tenth article entitled Data Augmentation approach to Distracted Driving Detection [10]. This work addresses the behavior problem associated with distracted driving that leads to vehicle crashes. To address this problem, the authors proposed a Data Augmentation method based on the driving operation area using the convolutional neural network classification model. The classification's result achieved a 96.97% accuracy using the distracted driving dataset. This method is helpful to detect drivers in actual application scenarios and identify dangerous driving behaviors. It helps to give an early warning of unsafe driving behaviors and avoid accidents.

The eleventh paper, entitled Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks was written by Aleksandr Romanov, Anna Kurtukova, Alexander Shelupanov, Anastasia Fedotova, and Valery Goncharov [11]. The authors explored the advantages and disadvantages of various approaches that can determine the author of a natural language text. Some of the examples found were used to identify authors of suicide notes, conduct forensic exams, and detect plagiarism. This article describes the process of identifying the author of Russian-language texts using support vector machine (SVM) and deep neural network architectures, as well as convolutional neural networks (CNN) with attention networks and transformers. The results show that all the considered algorithms are suitable for solving the authorship identification problem, but SVM offers the best accuracy. The average accuracy of SVM reaches 96%.

Nuno Marques da Costa, Nelson Mileu, and André Alves present article number twelve, entitled Dashboard COMPRIME_COMPRI_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal [12]. This article, the third in this Special Issue about the pandemic, shows a set of dashboards to disseminate information and multi-scale knowledge of COVID-19. As a result, the authors developed a system for monitoring the evolution of the pandemic. The constructed platform dynamically and interactively brings together a diverse set of variables and indicators that reflects the evolutionary behavior of the pandemic from a multi-scale perspective in Portugal. The authors mention that this approach proves to be crucial to guarantee everyone's access to information while simultaneously emerging as an epidemiological surveillance tool.

This tool can assist public authorities in terms of ensuring competent decision-making by defining control policies and fighting the spread of new coronavirus strains.

The article Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities, the thirteenth in this Special Issue, was written by Albert Weichselbraun, Philipp Kuntschik, Vincenzo Francolino, Mirco Saner, Urs Dahinden, and Vinzenz Wyss [13], and addressed the application of data-driven research methods to disciplines such as the Social Sciences and Humanities. The authors presented a case study that demonstrates the potential of the proposed method in the domain of Communication Science by creating approaches that aid domain experts in locating, tracking, analyzing, and finally, better understanding the dynamics of media criticism. The paper shows that data-driven research approaches require a tighter integration with the methodological framework of the target discipline to provide a significant impact on the target discipline.

The last article is the fourth study about covid19. Ana Teresa Ferreira, Carlos Fernandes, José Vieira, and Filipe Portela present the study Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients [14]. The authors induced intelligent models capable of predicting and supporting clinical decisions to predict if the patient will die or recover from COVID-19. The best scenario is composed of all comorbidities, symptoms, and ages. The best model achieved a sensitivity of 95.20%, accuracy of 90.67%, and specificity of 86.08%. The models were deployed as a service and are part of a clinical decision support system named ioCOVID19, which is available for authorized users anywhere and anytime.

Funding: This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.


Conflicts of Interest: The authors declare no conflict of interest.

References

- Maniou, T.; Veglis, A. Employing a Chatbot for News Dissemination during Crisis: Design, Implementation and Evaluation. *Future Internet* **2020**, *12*, 109. [CrossRef]
- Scheel, C.; Fallucchi, F.; De Luca, E. Visualization, Interaction and Analysis of Heterogeneous Textbook Resources. *Future Internet* **2020**, *12*, 176. [CrossRef]
- Rinaldi, A.; Russo, C.; Tommasino, C. A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features. *Future Internet* **2020**, *12*, 183. [CrossRef]
- Ponce, A.; Ponce Rodriguez, R. An Analysis of the Supply of Open Government Data. *Future Internet* **2020**, *12*, 186. [CrossRef]
- Karpinski, M.; Kuznichenko, S.; Kazakova, N.; Frazee-Frazenko, O.; Jancarczyk, D. Geospatial Assessment of the Territorial Road Network by Fractal Method. *Future Internet* **2020**, *12*, 201. [CrossRef]
- Lousado, J.; Antunes, S. Monitoring and Support for Elderly People Using LoRa Communication Technologies: IoT Concepts and Applications. *Future Internet* **2020**, *12*, 206. [CrossRef]
- Artiemjew, P.; Rudikova, L.; Myslivets, O. About Rule-Based Systems: Single Database Queries for Decision Making. *Future Internet* **2020**, *12*, 212. [CrossRef]
- Chua, S.; Foo, L.; Guesgen, H. Predicting Activities of Daily Living with Spatio-Temporal Information. *Future Internet* **2020**, *12*, 214. [CrossRef]
- Khrais, L. Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce. *Future Internet* **2020**, *12*, 226. [CrossRef]
- Wang, J.; Wu, Z.; Li, F.; Zhang, J. A Data Augmentation Approach to Distracted Driving Detection. *Future Internet* **2021**, *13*, 1. [CrossRef]
- Romanov, A.; Kurtukova, A.; Shelupanov, A.; Fedotova, A.; Goncharov, V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* **2021**, *13*, 3. [CrossRef]
- Marques da Costa, N.; Mileu, N.; Alves, A. Dashboard COMPRIME_COMPRI_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal. *Future Internet* **2021**, *13*, 45. [CrossRef]
- Weichselbraun, A.; Kuntschik, P.; Francolino, V.; Saner, M.; Dahinden, U.; Wyss, V. Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities. *Future Internet* **2021**, *13*, 59. [CrossRef]
- Ferreira, A.; Fernandes, C.; Vieira, J.; Portela, F. Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients. *Future Internet* **2021**, *13*, 102. [CrossRef]

Article

Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients

Ana Teresa Ferreira ¹, Carlos Fernandes ², José Vieira ² and Filipe Portela ^{1,2,*} 

¹ Algoritmi Research Centre, University of Minho, 4800-058 Guimarães, Portugal; a80702@alunos.uminho.pt

² IOTECH—Innovation on Technology, 4785-588 Trofa, Portugal; carlosfernandes@iotech.pt (C.F.); josevieira@iotech.pt (J.V.)

* Correspondence: cfp@dsi.uminho.pt

Abstract: Nowadays, there is an increasing need to understand the behavior of COVID-19. After the Directorate-General of Health of Portugal made available the infected patient's data, it became possible to analyze it and gather some conclusions, obtaining a better understanding of the matter. In this context, the project developed—ioCOVID19—Intelligent Decision Support Platform aims to identify patterns and develop intelligent models to predict and support clinical decisions. This article explores which typologies are associated with different outcomes to help clinicians fight the virus with a decision support system. So, to achieve this purpose, classification algorithms were used, and one target was studied—Patients outcome, that is, to predict if the patient will die or recover. Regarding the obtained results, the model that stood out is composed of scenario s4 (composed of all comorbidities, symptoms, and age), the decision tree algorithm, and the oversampling sampling method. The obtained results by the studied metrics were (in order of importance): Sensitivity of 95.20%, Accuracy of 90.67%, and Specificity of 86.08%. The models were deployed as a service, and they are part of a clinical decision support system that is available for authorized users anywhere and anytime.

Citation: Ferreira, A.T.; Fernandes, C.; Vieira, J.; Portela, F. Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients. *Future Internet* **2021**, *13*, 102. <https://doi.org/10.3390/fi13040102>

Academic Editor: Paolo Bellavista

Received: 5 March 2021

Accepted: 17 April 2021

Published: 20 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: COVID-19; classification; information systems; public health; data mining; ioCOVID19

1. Introduction

Every day, the world population is faced with an increasing number of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases. Portugal is one of the European Countries where Coronavirus has a significant impact, and the number of cases and deaths is increasing every day. As a way to provide some inputs to the decision process, a research project was released: ioCOVID19 aims to develop an intelligent decision support platform that allows the prediction of the evolution of the disease in a specific patient to support clinicians in the fight against Coronavirus disease 2019 (COVID-19).

Since the provided data only refers to the Portuguese population, it was attempted to seek information regarding the most likely outcome of a specific patient based on his characteristics—such as the comorbidities he presents, symptoms, age, and gender. This article was developed as an integral part of a research project. It is a continuation of the work already demonstrated in the article “The clinical reality of COVID19 in Portugal—A clustering analysis” [1]. It intends to expose the obtained results concerning the analysis of the data carried out to the moment of writing the document. Therefore, the development of this phase of the project went through the following steps: Data preparation, since there was a need to replicate some of the data under study; Modeling, where the classification process was carried out and Evaluation of the results obtained in the last moment. The study aims to help health professionals in the moment of crucial decision-making. Previous records of infected patients by COVID-19 were used to predict their outcome. Thus, based on the classification processes, it is attainable to indicate whether the patient will need to be hospitalized or need specific medical support. It is also possible to understand the patient's

outcome, that is, to predict whether the patient will recover or die. The data found in this article goes back to February of 2021 regarding newly infected patients.

The article presents the following structure to expose all the results, techniques, and conclusions in the most organized and detailed way: first, to situate the reader in the theme and problem addressed, a short introduction to the subject is presented. The article's main themes are then detailed in the Background section, where it is described what type of data mining technique is used in the study. It is then portrayed in more detail which materials and methods were used in the project's development, such as which data was used and which methodologies were adopted. Regarding the Case Study point, the Cross-industry Standard process for data mining (CRISP-DM) and Design Science Research (DSR) methodologies are exposed in more detail, always establishing a connection with the project to understand the development achieved in each phase. In the section referring to the Results and Discussion, all relevant results and information obtained during the classification are exposed, discussed, and evaluated in detail.

2. Background

This section presents the article's relevant topics, showing Portugal's current situation and mentioning related works.

2.1. COVID-19

COVID-19 is the World Health Organization (WHO) official name for the disease caused by the new coronavirus SARS-COV-2 (Severe acute respiratory syndrome coronavirus 2), which can cause severe respiratory infections such as Pneumonia. This virus was first identified in humans in the Chinese city of Wuhan, Hubei province, at the end of 2019. The main symptoms associated with the COVID-19 infection are fever (body temperature above 38 degrees Celsius), cough, and difficulty breathing, such as shortness of breath. Some fewer common symptoms associated with the disease are sore throat, runny nose, headaches, muscle aches, and tiredness. In more extreme cases, it can also result in severe Pneumonia with acute respiratory failure, kidney and other organs failure, and, eventually, death. The contagion period is currently considered 14 days; however, transmission by asymptomatic people is still under investigation [2].

2.2. Portuguese Reality of COVID-19

When writing this article (6 April 2020), the scenario in which Portugal found itself concerning cases of COVID-19 was 824,368 infected and 16,887 deaths. This information was provided by the Directorate-General of Health of Portugal (DGS) [3].

To understand the pandemic's effect in Portugal, it is essential to know its mortality rate concerning the deadliest diseases. According to the available records for the year 2019, the relative circulatory system's diseases represent 29.90% of deaths in the country, being the leading cause of death. This is followed by malignant tumors, representing 25.50% of deaths and respiratory system diseases with 10.90% [4]. Therefore, through comparison, it is possible to understand the Coronavirus's impact—however, considering that the virus is a recent phenomenon, the used data may not provide a correct representation of its actual effects. So, taking this into account, in March 2021, the mortality rate in Portugal stands at 2.0%. [5].

2.3. Project ioCOVID-19

This article is linked to the project developed—ioCOVID19—Intelligent Decision Support Platform NORTE-01-02B7-FEDER-048344. The article represents the second phase of the developed project. It aims to create an essential platform for clinicians to combat COVID-19. Its primary goals are to analyze the available data referring to those infected by Coronavirus in Portugal and predict the evolution of a given patient's disease from a set of predictive models. Using open data accessible online and made available by the SNS (Portuguese National Health Service) and DGS, it is possible to categorize patients and

assess the impact that each variable has on the disease's course and predict the kind of patient discharge. A Web/Mobile platform—ioCOVID19—was also developed, aiming to allow physicians and/or nurses to access a set of essential data for decision-making. The models here depicted are part of the inference engine of the application conceived.

2.4. Data Mining

For the depicted project, Data Mining (DM) techniques were applied, in order to extract only valuable and relevant information from the data. The main objective is to find non-evident relationships and patterns between data or, in other words, it is the process of discovering knowledge from the data [6,7]. To this end, the used techniques allowed the identification of different categories, patterns detection, and the forecast of different scenarios—in this case, the outcome of the disease on a given patient. Nevertheless, the methodologies used to achieve the proposed objective were Classification Analysis and Neural Networks.

To assess and compare the models obtained, confusion matrix and Receiver Operating Characteristic (ROC) curve were considered [8].

2.5. Classification

Classification is a supervised machine learning technique used in DM. Briefly, the classification algorithms learn from the data input provided and then use the knowledge obtained to classify new observations [7]. From the available classification algorithms, only five were considered for this project (a brief explanation of the algorithms is presented in Section 4 of this paper): Logistic Regression, Naive Bayes Classifier, Decision Tree, and Deep Learning. For this project, a classifier was used to predict the categorical labels. From this method, it was possible to understand, for example, the outcome (dead or recovered) of a given patient taking into account their characteristics. To assess and compare the models obtained, a set of metrics were used: Confusion Matrix and Receiver Operating Characteristic (ROC) curve (more specifically, the measures calculated were Accuracy, Sensitivity, and Specificity) [8].

The Confusion Matrix presented in Table 1 is used to measure the performance of a classification algorithm in terms of True Positive (the classifier predicts positive and it is correct); True Negative (the classifier predicts negative, and it is correct); False Positive (the classifier predicts positive, and it is incorrect), and False Negative (the classifier predicts negative, and it is incorrect) [9].

Table 1. Confusion Matrix.

		Actual Values	
		True	False
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

The metric Accuracy exposes how often the classifier was correct. Sensitivity represents the positive data correctly predicted the total number of positive samples. Specificity represents the ratio of samples correctly classified as negative to the total number of negative samples [10]. The metrics mentioned are calculated as follows [10]:

$$\begin{aligned}
 \text{ACCURACY} &= (TP + TN) / (TP + TN + FP) \\
 \text{SENSITIVITY} &= TP / (TP + FN) \\
 \text{SPECIFICITY} &= TN / (TN + FP)
 \end{aligned}$$

2.6. Similar Works

Due to the impact of COVID-19 on today's society, an extensive set of studies was carried out. However, since the data under analysis exclusively refers to Portugal, no

works using data mining referring to the Portuguese situation in the face of the pandemic were found. Therefore, after conducting research and analyzing the obtained results, it was possible to realize that no study was being carried out along the lines of the project exposed in this article. Nonetheless, a set of works was carried out in the area in question using data from Portuguese institutions, which allows assessing this project's feasibility. Regarding the studies carried out at national level in this area of support systems in the medical sector, the following examples will be considered:

- An intelligent decision support system is applied to Intensive Care Unit (ICU) based on several models, such as patient's vital signs, critical events, medical scores in ICU, and the data mining models. The main goal of this project concerns the hourly forecast of organ failures and the result [11];
- The application of domain knowledge in order to improve an intelligent decision support system related to the study of bacteriological ingestions. The goal is to make the decision-making process more efficient about which antibiotic is the most appropriate for a given situation, based on specialists' knowledge in the field [12].

However, for projects related to COVID-19, the majority of studies are carried out outside Portugal. As an example, two studies in this field are as follows:

- The prediction of early mortality risk based on patients infected with covid-19—For this purpose, several machine learning models were used, which revealed factors such as age, c-reactive protein sensitivity, lymphocyte count, and D-dimer influence the result of the infected patient [13];
- Understanding the role of preconditions associated with COVID-19—The main objective is to identify which characteristics are associated with the patient's death. As of July 2020, the main conclusion was that this outcome (death by COVID-19) is associated with male individuals and over the age of 60 years [14].

From these examples, it is possible to observe that there are studies on the implementation of decision support tools in the medical context being developed but also studies regarding the coronavirus and its evolution. However, there are no studies combining both themes.

3. Materials & Methods

The Directorate-General provided the portrayed data for Portuguese Health, and it refers to patients infected with COVID-19. It was collected by medical professionals between 2 March 2020 and 28 February 2021.

3.1. Design Science Research

Since this is a research project and, to understand if it is possible to characterize the clinical typology of patients infected with Coronavirus (as well as the outcome of the disease), two methodologies were followed: Design Science Research (DSR) as a research methodology, and Cross-Industry Standard Process for Data Mining (CRISP-DM). DSR consists of 6 phases: 1. Identifying the problem and motivation; 2. Defining objectives of the solution; 3. Design and development; 4. Demonstration; 5. Evaluation; 6. Communication. These phases provide guidelines for evaluation and interaction in research projects. To put DSR in action, it is necessary to use a practical methodology to help drive the project, so, Cross-Industry Standard Process for Data Mining was chosen [15].

3.2. Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM was the second methodology used, and it is focused on the development of predictive models. The CRISP-DM method provides a global perspective on the life cycle of a data mining project. This cycle, shown in Figure 1—Project Workflow—is divided into six sequential phases. There are dependencies between them; however, it does not have a rigid structure. The current CRISP-DM model stages for data mining projects are Business Understanding, Data Understanding, Data Preparation, Modelling,

Evaluation, and Deployment. The information depicted in this document was achieved after completing the fifth phase—Evaluation [16].

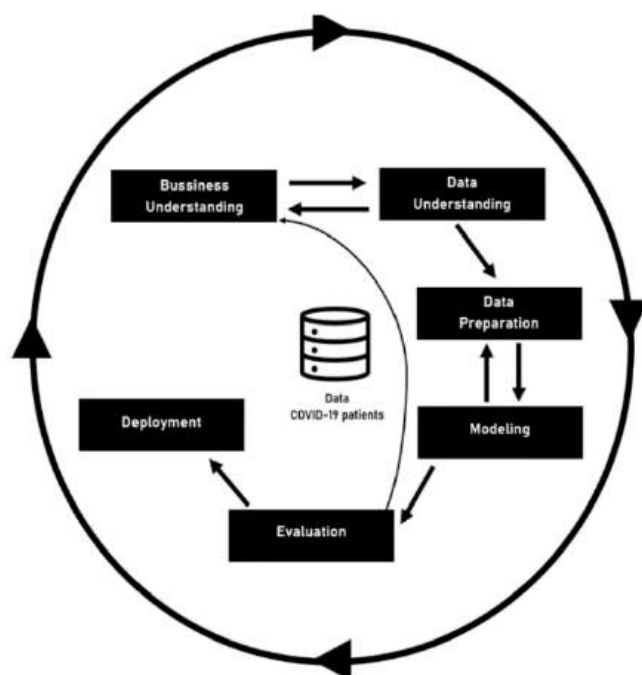


Figure 1. Project Workflow.

To drive this project, it is essential to assure a relation between the research methodology and practical method.

3.3. DSR and CRISP-DM

Since both methodologies are used concurrently, it is possible to point out the relationships between the phases of CRISP-DM and DSR (Design Science Research). The CRISP-DM method comprises the following activities: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [15].

This article portrays the phases of both CRISP-DM and DSR. For example, phases 1 and 2 of the DSR are directly linked to the first activity of CRISP-DM, as it is possible to check in the table. The remaining relationships are also shown in Table 2 [15].

Table 2. Crossover of CRISP-DM and DSR methodologies.

Methodology	Activities	DSR Phases					
		1	2	3	4	5	6
CRISP-DM	Business Understanding	X	X				
	Data Understanding		X	X			
	Data Preparation			X			
	Modeling			X			
	Evaluation				X	X	
	Deployment				X	X	

In other words, in the first phase of the project, there was a necessity to identify the business objective to understand the motivation for the project's emergence. Then, regarding the provided data, the goal was to understand its use and, from there, to define the objectives of the solution to be developed. In a third phase, data is prepared to facilitate its use in future steps. Afterward, the modeling process begins, which in the current project

involves the data mining classification technique. This phase is linked to the previous two. From this point, there may be a need to make changes to obtain the best results. Then, in the Evaluation phase, the results obtained previously were analyzed to understand their usefulness and viability. And finally, after completing the previous stages, it is necessary to apply the obtained resources to make them useful in the environment in which they are required.

3.4. Tools

Python programming language was used for the preparation of the data and consequent analysis of it. Moreover, to be able to apply this same language to the project in question, a set of libraries were used to enable the preparation, analysis, and consequent prediction of the data:

- Panda's library allows the use of the DataFrame object to provide storage and manipulation of the data organized by columns [17];
- Scikit-learn (sklearn) is a machine learning software that provides various algorithms such as classification, regression, and clustering [18];
- TensorFlow is a library used for solving problems with Machine Learning and Deep learning [19].

3.5. Classification Algorithms

A set of classification algorithms was used to achieve this project's aim, namely Logistic Regression, Naive Bayes, Decision Trees, and Deep Learning.

3.5.1. Logistic Regression

Logistic regression is a supervised classification algorithm modeling the data using the sigmoid function. This algorithm is used to predict the probability of a categorical dependent variable [20]. Both the set of resources (input) and the destination variable (output) can only assume discrete values when involved in classification problems. This algorithm builds a model to predict the probability that a given entry belongs to the category numbered as 1. In other words, this model predicts the Probability of $P(Y = 1)$. Therefore, when using the algorithm Logistic Regression, it is necessary to take into account the following assumptions [20]:

- The dependent variable has to be binary;
- Factor 1 of the variable should represent the desired outcome;
- Only the relevant variables should be included in the classification process;
- The independent variables should be independent of each other;
- It requires large sample sizes.

For this model, no hyperparameter was modified to control the learning process; all parameters had their default value.

3.5.2. Naïve Bayes

The Naive Bayes algorithm is a simple probability classifier. Bayesian algorithms predict the class depending on the probability of belonging to that class [21]. It calculates a set of probabilities from the frequency count and the combinations of values in a given data set. This algorithm is based on Bayes' theorem, assuming that all variables are independent. Bayes' theorem follows the following formula [22]:

$$P(A|B) = P((B \setminus A)P(A))/P(B)$$

From this theorem, it is possible to find the probability of event A to happen (what is intended to be predicted—the outcome), given that a particular event B occurred (the comorbidities, symptoms, age, and gender of a given patient). However, this lack of independence is not valid in real contexts because it disregards the correlation between the variables. Hence it is characterized by "Naive" [23].

For this model, no hyperparameter was modified to control the control de learning process; all parameters had their default value.

3.5.3. Decision Trees

Decision Tree is one of the most important and well-known classification algorithms. This algorithm is a nonparametric supervised learning method. The goal is to create a model capable of predicting the value of a target variable by learning simple decision rules inferred from the given data [24]—in other words, it works as a set of “yes” or “no” questions based on specific characteristics to reach the target variable. The base components are nodes and branches, and the next most important steps are splitting, stopping, and pruning, so it is possible to create a decision tree [25].

For this model, one hyperparameter setting was modified, as shown in Table 3, yet the remaining parameters kept their default value. The parameter `max_depth` assumes the value of 20, which establishes the tree’s maximum depth to control the size of the generated tree. This value was achieved after implementing the `GridSearchCV` library, which allows identifying the best values to be applied in the parameters [26,27].

Table 3. Decision Tree hyperparameters settings.

Parameter	Assumed Value
<code>max_depth</code>	20

3.5.4. Deep Learning

Deep Learning (DL) is Machine Learning based on algorithms inspired by the human brain structure and function denominated artificial neural networks (ANNs) [28], also known as feedforward neural networks. With the constant increase of data and processing power, the need to apply both concepts arose. DL consists of a technique associated with neural networks that enable computers to learn through experience from a hierarchy of concepts. This hierarchy allows the computer to learn complex concepts by building them out of simpler ones. Therefore, since the computer can learn from its own experience, there is no longer a need for Human intervention [29]. This method achieves excellent power and flexibility by learning to represent information as a nested hierarchy of concepts. An ANN is composed of 3 components: Input Layer, Hidden Layers, and Output Layers [30]. There are several neurons for input values and others for output values; however, many neurons are interconnected in the hidden layer. So, formally neurons define Deep Learning.

For this model, the changes made to the hyperparameter were the following, as depicted in Table 4:

Table 4. Deep learning hyperparameters settings.

Parameter	Assumed Value
<code>batch_size</code>	64
<code>callback</code>	EarlyStopping
<code>epochs</code>	20

The remaining parameters kept their default value.

4. Case Study

The case description goes through the methodology presented in the CRISP-DM section, as possible to understand in the following points. As previously mentioned, the used data is inserted in the time interval between March 2nd of 2020 and February 28th of 2021. The provided data has 805 141 records, with each record referring to a patient. To provide a better understanding of the used data, attached to the article is a document denominated “COVID-19 Data Analysis”, where it is possible to find relevant general information.

4.1. Business Understanding

The first phase—Business Understanding—focuses on understanding the project's objectives and requirements from a business perspective. It is then possible to design a preliminary data mining project that can achieve the outlined goals. In the project in question, this phase consisted of realizing what type of data would be provided and its use in a data mining project. Therefore, the project intends to develop a platform for clinicians to combat COVID-19, with the primary objective of predicting the evolution of a specific patient's disease—evaluating the impact that each variable has on the disease and predicting the type of discharge. In the study, the particular aspect addressed is trying to predict the outcome or the need for further medical support for clinical patients in Portugal.

4.2. Data Understanding

Data Understanding starts after the initial collection of data to be worked. At this stage, data analysis is carried out to search for possible quality problems and, consequently, obtain a better understanding of them. Due to this type of study, it is also easier to understand if there is any subset that can be obtained considering the available information, thus enriching the subject under investigation. So, it was at this stage that a meticulous analysis of the data at hand was carried out.

The data provided has 805,141 records, collected between the 2nd of March and the 28th of February. Since the patient's age is an important variable in the classification process, all records with no associated age were ignored. Therefore, the number of associated records changed to 739,297 (representing 91.8% of the initial data). Subsequently, a table was created to expose all the relevant information in order to obtain a better perception of the data. However, since that table has large dimensions, it is attached to the article. The table is denominated "COVID-19 Data Analysis".

For the reader to obtain a global perception of the data, Figure 2—Records General Information presents some relevant information to retain. According to the presented figure, it is possible to perceive the gap between records of recovered and infected patients, a crucial detail that can influence the result in the modeling phase; for example, the number of dead patients only represents 1.98% of the cases. The percentage of patients who present comorbidities prior to infection caused by the SARS-CoV-2 virus can also be understood.

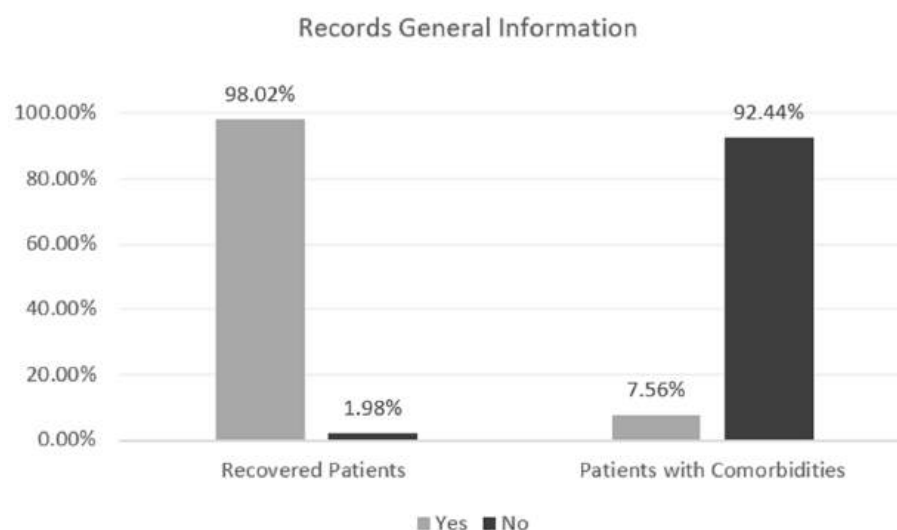


Figure 2. Records General Information.

The comorbidities noted by health professionals are the following: Diabetes, Asthma, Neoplasm, Chronic Lung Disease, Chronic Kidney Disease, Chronic Neurological and Neuromuscular Disease, Chronic Hematological Diseases, Chronic Neurological Deficiency, Liver Pathology, HIV or other Immunodeficiencies, Acute Renal Failure, Cardiac Insufficiency, and Consumption Coagulopathy. The comorbidities distributions are shown in

Figure 3—Comorbidities Distribution and how it is possible to perceive the comorbidities with the largest number of associated records are: Diabetes, Asthma and Neoplasia.

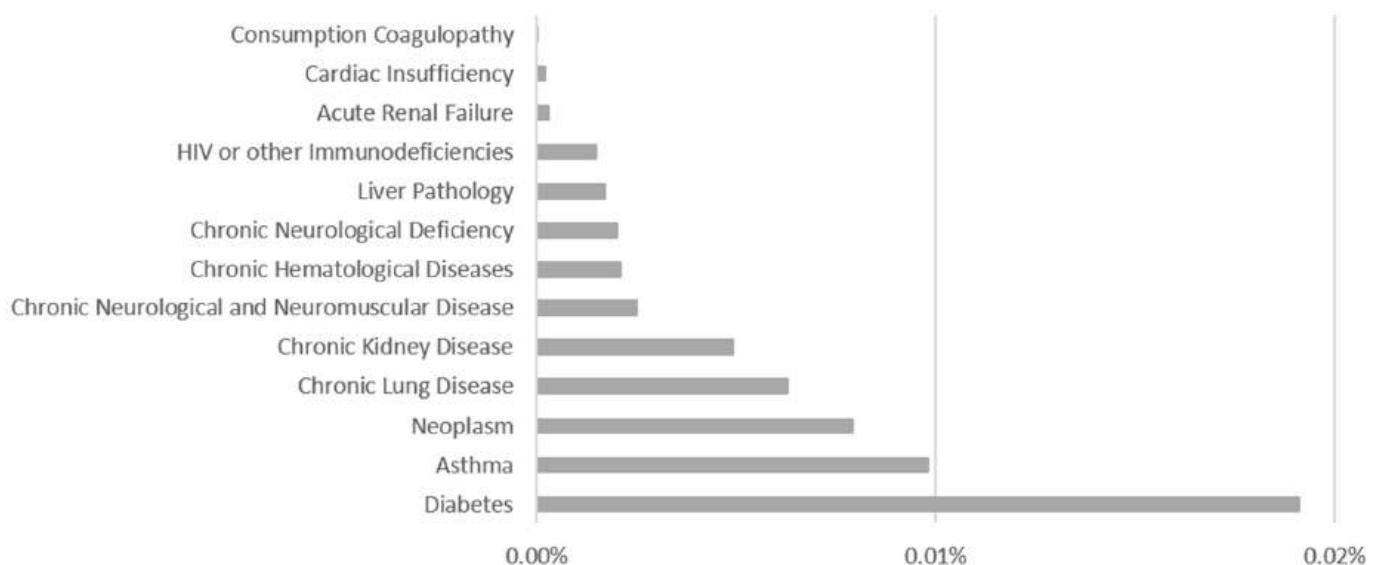


Figure 3. Comorbidities Distribution.

For a better comprehension of the dataset, an analysis of the data is provided in the supplement material section. Table S1 presents a description, the data type, quality issues, percentage of non-nulls, and an example of possible values for each variable.

4.3. Data Preparation

As it is possible to understand by Figure 2—Records General Information, presented in the previous point, the data in use is unbalanced: the discrepancy between patients who died and infected/recovered patients is significantly accentuated, showing a 1:49.4 proportion of the minority class [deaths] to the majority class [infected/recovered], which can result in an Imbalanced Classification. This type of detail needs an important evaluation, given that the minority class instances are easier to be ignored by the learning algorithm due to the high number of instances from the majority class [31]. To counter this problem, Random Oversampling and Undersampling were used.

Oversampling involves the duplication of records from the minority class to correct the imbalanced data. On the other hand, Undersampling consists of a random selection of data from the majority class that involves removing instances of this class until the majority class and the minority class present an equal number of examples for each class [32].

To obtain the best possible result for the classification algorithms, using the Oversampling method, small changes were made to the data. The number of patients infected/recovered is 724,630 (which represents 98% of the selected data) and, on the other hand, the number of patients who died is 14,667 (which represents around 2% of the selected data). Probability sampling techniques were used in order to make sure that all elements of the population have an equal chance of being selected, having been used more specifically Simple Random Sampling [33]. For such, a data sample was created by selecting only the records of patients who died, therefore, it contains the 14,667 records previously mentioned. This data sample was replicated 50 times, so the number of records regarding patients who are infected/recovered or died is balanced, as can be observed in Table 5—Distribution of cases using Oversampling, thus allowing better results. However, this is not the correct method when the need for replication is very high. Furthermore, the Undersampling method was also used in order to compare the results obtained by the two techniques considered. This method requires a random selection of data referring

to Infected/Recovered patients so that both classes involved obtain the same number of records, as seen in Table 6—Distribution of cases using Undersampling.

Table 5. Distribution of cases using Oversampling.

Targets	Replicated	Before		After	
		Number of Cases	Percentage of Cases	Number of Cases	Percentage of Cases
Patients who died	50	14,667	1.98%	733,350	50.3%
Patients infected/recovered	0	724,630	98.02%	724,630	49.7%

Table 6. Distribution of cases using Undersampling.

Targets	Before		After	
	Number of Cases	Percentage of Cases	Number of Cases	Percentage of Cases
Patients who died	14,667	1.98%	14,667	50.0%
Patients infected/recovered	724,630	98.02%	14,667	50.0%

4.4. Modelling

In this phase, several modeling techniques were selected and later applied to the previously treated data, optimized to obtain better results. For this, the modeling includes the following tasks: Selecting modeling techniques, generating a test environment, building the model, and evaluating it [34,35].

To proceed with the application of the algorithms to make predictions on the data, the data had to be divided into two parts: a training set and a testing set. The training set is used by the various classification algorithms so that they can train and adjust the model's constituent parameters. The test set allows evaluating the Accuracy of the final model taken from the training. For the data referring to patients infected by the Coronavirus, the data was divided using Stratified Cross-Validation since it offers more robust and reliable results. This method estimates the Machine Learning model's learning capacity to make predictions on unused data during the training phase, evaluating the classifier performance. Therefore, when a supervised Machine Learning process is carried out, data is separated so that one of these sets can be tested. This procedure has a single parameter denominated "k", and this value refers to the number of sets into which the data set in use was divided. For the project in question, this variable assumed the value of ten, meaning that the data sample in question was divided into ten different groups. The purpose of cross-validation is to test the model's ability to correctly predict new data that was not used in the training phase [36]. The Stratification process reorganizes the data to ensure that each fold has the same proportions of observations and provides a correct representation of the data as a whole [37].

In this case, the data referring to COVID-19's patients will be subject to several classification algorithms to be able to predict the class of data in question. The SVM classification algorithm was also considered; however, it was not possible to obtain results on time due to the amount of data.

The targets under consideration, presented in JSON (JavaScript Object Notation), were the following:

```
{
  "Scenarios":{
    "S1": "All comorbidities, symptoms, age and gender",
    "S2": "All comorbidities, age and gender",
    "S3": "Risk comorbidities, age and gender",
    "S4": "All comorbidities, symptoms and age",
  }
}
```

```

"S5": "All comorbidities and symptoms",
"S6": "All comorbidities",
"S7": "All comorbidities and age",
"S8": "All comorbidities and gender"
},
"Techniques": {
  "T1": "Logistic Regression",
  "T2": "Naive Bayes",
  "T3": "Decision Trees",
  "T4": "Deep Learning"
},
"Sampling Methods": {
  "SM1": "Oversampling",
  "SM2": "Undersampling"
}
}

```

Therefore, 64 models were induced to obtain the published results (8 Scenarios \times 1 Targets \times 4 Techniques \times 2 Sampling Methods).

4.5. Evaluation

In the Evaluation phase, the developed model was evaluated to ensure it allows the achievement of business objectives. In other words, this phase of the project presupposes the implementation of the following stages: Evaluate results, Review the process and Determine the next steps.

So, the previous phase models go through an evaluation process regarding performance and utility. For this, metrics are used to understand which algorithm has the best results for the problem presented.

Therefore, it was necessary to define a threshold to obtain the best possible model. Consequently, from the knowledge of the specialist in the field, the following values were stipulated for the metrics Sensitivity, Accuracy, and Specificity:

- Sensitivity $\geq 90\%$;
- Accuracy $\geq 80\%$;
- Specificity $\geq 80\%$.

The obtained results varied from 0 to 1, and the main objective is to predict the 1. So, to evaluate the performance of the models, the mentioned metrics are used. Metrics related to Sensitivity have priority compared to the others given because it allows for balancing the number of false positives and negatives. However, it is important that there are no unbalanced models, such as predictions with high false positives. So, both Accuracy and Specificity must have relevant values. This metric will also be important for the future implementation of ensembles.

4.6. Deployment

In the Implementation phase, the obtained knowledge has to be presented so that the client—in this case, health professionals—can use it. In other words, it involves making the acquired knowledge useful for decision-making.

The project in question is part of the clinical Intelligence Decision Support System (CIDSS). The CIDSS is an information system focused on the health area, conceptualized to provide support in health professionals' decision-making, in this case, the ones that are on the front line to combat COVID-19. Clinical observations are linked to the domain's knowledge in a specific area of health, and they can influence physician's choices to improve health care services.

Regarding the coronavirus project, a web and mobile application were developed to support the results obtained. Furthermore, to allow its use, the models will be consumed by clinicians as a service. To this end, a third article associated with the project is under

development, where the main outputs from both the clustering and the classification phase will be exposed, explaining more closely why it is a CIDSS project. The deployment process turns the models pervasive and makes it available as a service anywhere and anytime to any user with access privileges. The deployment process and the entire solution will be described in-depth in a further article [38].

5. Results and Discussion

At this stage of the project, the obtained results in the modeling phase are exposed. The classification algorithm's best-obtained result and its respective scenario are shown for the target under study. In this way, it is possible to uncover which is the best model for the considered target.

The prediction models were constructed considering the target under study (Patient Outcome), eight different scenarios (S1 to S8), and applied five different DM techniques (T1 to T5). For such, the models can be identified as an ensemble of a three-dimensional matrix M composed by $s = 8$ scenarios $\times t = 5$ techniques $\times sm = 2$ sampling methods. Each element of M corresponds to a particular model and can be defined as:

$$M_{s,t,sm} = \left\{ \begin{array}{l} s = 1 \dots 8 \\ t = 1 \dots 4 \\ sm = 1, 2 \end{array} \right\}$$

where,

s :	t :	sm :
1 = {All comorbidities, symptoms, age and gender}	1 = Logistic Regression	1 = Oversampling
2 = {All comorbidities, age and gender}	2 = Naive Bayes	2 = Undersampling
3 = {Risk comorbidities, age and gender}	3 = Decision Tree	
4 = {All comorbidities, symptoms and age}	4 = Deep Learning	
5 = {All comorbidities and symptoms}		
6 = {All comorbidities}		
7 = {All comorbidities and age}		
8 = {All comorbidities and gender}		

The table below presents the best results obtained by each scenario ($s = 1$ to 8). For each model, the values of Accuracy, Specificity, and Sensitivity are presented. For example, for model $M_{1,3,1}$ —consisting of scenario 1, Decision Tree technique, and sampling method Oversampling—the Accuracy obtained was 90.99%, which represents the percentage of correctly labeled subjects. Specificity was 86.79%, indicating the percentage of recovered outcome correctly predicted, and Sensitivity was 95.14%, which represents the percentage of death outcome correctly identified.

For example, taking into account the $M_{1,3,1}$ model (composed by scenario 1—Decision Tree technique and sampling method Oversampling), the metrics Accuracy, Specificity, and Sensitivity were noted. The Accuracy obtained was 90.99%, representing the percentage of correctly labeled subjects; Specificity was 86.79%, which indicates the percentage of recovered outcome correctly predicted, and Sensitivity was 95.14%, representing the percentage of death outcome correctly identified.

In this way, it is possible to understand the importance of the Sensitivity metric. It is preferable to obtain a forecast that indicates if the patient is going to die, but that, in reality, ends up recovering, than to predict the opposite, since it can affect the decisions of health professionals. All the results obtained can be consulted in the Supplementary Materials—“Table S2—Classification Results”.

Therefore, the best results obtained for Patient outcome, where bold is the metrics that achieved the threshold, shown in Table 7—Metrics for Patient Outcome, were as follows:

Table 7. Metrics for Patient Outcome.

Model	Accuracy	Specificity	Sensitivity
$M_{1,3,1}$	90.99%	86.79%	95.14%
$M_{2,3,1}$	88.94%	84.43%	93.39%
$M_{3,3,1}$	88.91%	84.60%	93.16%
$M_{4,3,1}$	90.67%	86.08%	95.20%
$M_{5,3,2}$	87.66%	82.19%	93.12%
$M_{6,3,1}$	58.72%	96.80%	21.08%
$M_{7,3,1}$	88.13%	84.80%	91.41%
$M_{8,3,2}$	60.86%	97.00%	24.74%

As previously mentioned, the purpose of the project is to understand each patient's outcome, positive or negative, based on their characteristics. Therefore, after a brief analysis of the results presented, the main outputs gathered are discussed in this section. Firstly, it is important to remember the threshold previously defined to find the best model for the project, which is defined by the following metrics (in order of importance):

- Sensitivity $\geq 90\%$;
- Accuracy $\geq 80\%$;
- Specificity $\geq 80\%$.

All values that managed to reach or exceed the defined threshold were marked in bold. That said, with the exception of models $M_{6,3,1}$ and $M_{8,3,2}$, all the other models can be considered for the prediction of the outcome of a given patient. In general, the models present interesting results for Sensitivity, and the lowest values are obtained by scenario 6 (all comorbidities) and scenario 7 (all comorbidities and age). The models that obtained the best results, both in Sensitivity and in the other metrics, were $M_{1,3,1}$ and $M_{4,3,1}$ —which are associated with scenario 1 (All comorbidities, symptoms, age, and gender) and scenario 4 (All comorbidities, symptoms, and age). All models (except the two with the lowest results) were able to reach the stipulated threshold (since, for example, the Sensitivity value ranges from 91.41% to 95.20%). It is possible to notice that the more detailed the patient's profile was, the higher the probability of the model obtaining better results. Another interesting point to draw from the results gathered is that for all models, the classification algorithm that showed the best metrics was T3—Decision Tree. Regarding the sampling methods, the method that presented the best results was Oversampling, being present in six of the eight models presented.

Therefore, the $M_{4,3,1}$ model is the best since it is the one that presents the best result in the Sensitivity metric (95.20%), and in the other metrics, it presents values higher than those of the defined threshold. This model uses the Decision Tree as a classification and oversampling technique for sampling technique.

6. Conclusions

This paper presents evidence that it is possible to predict the outcome of a specific patient infected with the SARS-CoV-2 virus using the characteristics indicated by the data provided, thus making it possible to assist clinicians at crucial decision-making moments. The article exposes the authors' work from the moment of data collection and analysis to the implementation of the data in the modeling phase to extract knowledge from them.

The results were evaluated in terms of a collection of three metrics and in accordance with the thresholds established; however, priority was given to the Sensitivity metric. Six of the exposed models meet the threshold. This means that for the target under study—Patient's outcome—health professionals will be able to predict the outcome, death, or recovery of a given patient. The model with the best results is $M_{4,3,1}$ since it presents the best result for the metric Sensitivity (95.20%) and the following metrics—Accuracy (90.67%) and Specificity (86.08%), while also showing positive values above the threshold. This model consists of scenario 4 (all comorbidities, symptoms, and age), decision tree technique, and sampling method oversampling.

These results indicate that clinicians use the predictions to understand the most likely outcome of the patients, which allows health professionals to make better decisions on how to act towards an infected patient. In terms of future work, the continuous reception and exploration of new records regarding patients infected by COVID-19 will allow the exploration of new patterns, techniques and broadcast the solution next to the medical community. Simultaneously, it will make possible the continuous improvement of the results obtained by the predictive models. Induced Models are part of the inference layer of the Clinical Decision Support System (CIDSS) developed and can be easily used by clinicians approved by the Institutions that use the platform.

The interested reader should consult the official page of the project (<https://iocovid19.research.iotech.pt>) for more information about future work to be developed, such as model optimization with the most recent data and the conclusion of the CIDSS deployment.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/fi13040102/s1>, Table S1: Dataset analysis, Table S2: Patient's Outcome results.

Author Contributions: Conceptualization, F.P. and J.V.; methodology, C.F.; validation, F.P., J.V. and C.F.; formal analysis, A.T.F.; investigation, A.T.F., C.F., J.V. and F.P.; resources, F.P.; data curation, A.T.F., J.V., C.F.; writing—original draft preparation, A.T.F.; writing—review and editing, F.P.; visualization, A.T.F.; supervision, F.P.; project administration, F.P.; funding acquisition, F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Portugal 2020 program through NORTE-01-02B7-FEDER-048344.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to being the property of the Ministry of Health.

Acknowledgments: We acknowledge the Portuguese Directorate-General of Health for providing us access to the COVID-19 Clinical Dataset.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Ferreira, T.; Fernandes, C.; Vieira, J.; Portela, F. The clinical reality of COVID19 in Portugal—A clustering analysis. Unpublished.
2. DGS. Perguntas Frequentes. Available online: <https://covid19.min-saude.pt/category/perguntas-frequentes/> (accessed on 13 December 2020).
3. DGS. Relatório de Situação. Available online: <https://covid19.min-saude.pt/relatorio-de-situacao/> (accessed on 6 April 2021).
4. Óbitos por Algumas Causas de Morte (%). Available online: [https://www.pordata.pt/Portugal/%C3%93bitos+por+algumas+causas+de+morte+\(percentagem\)-758](https://www.pordata.pt/Portugal/%C3%93bitos+por+algumas+causas+de+morte+(percentagem)-758) (accessed on 29 March 2021).
5. Nascimento, F. Risco de morrer por Covid-19 em Portugal está entre 0,7 e 2 por cento. Available online: <https://www.tsf.pt/portugal/sociedade/risco-de-morrer-por-covid-19-em-portugal-esta-entre-07-e-2-por-cento-13489927.html> (accessed on 4 April 2021).
6. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: Oxford, UK, 2012; pp. 12–18.
7. Nu Phyu, T. Survey of Classification Techniques in Data Mining. *IMECS* **2009**, *1*, 1.
8. Bradley, A.P. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
9. Visa, S.; Ramsay, B.; Ralescu, A.L.; Van Der Knaap, E. Confusion Matrix-based Feature Selection. *MAICS* **2011**, *710*, 120–127.
10. Agarwal, R. The 5 Classification Evaluation Metrics Every Data Scientist Must Know. Available online: <https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226> (accessed on 4 April 2021).
11. Portela, F.; Santos, M.F.; Machado, J.; Abelha, A.; Silva, Á.; Rua, F. Pervasive and Intelligent Decision Support in Intensive Medicine—The Complete Picture. In *Information Technology in Bio- and Medical Informatics*; Springer International Publishing: Cham, Switzerland, 2014; pp. 87–102.
12. Veloso, R.; Portela, F.; Filipe Santos, M.; Silva, Á.; Rua, F.; Abelha, A.; Machado, J. Using Domain Knowledge to Improve Intelligent Decision Support in Intensive Medicine—A Study of Bacteriological Infections. In *Proceedings of the International Conference on Agents and Artificial Intelligence*, Lisbon, Portugal, 10–12 January 2015.
13. Hu, C.; Liu, Z.; Jiang, Y.; Shi, O.; Zhang, X.; Xu, K.; Suo, C.; Wang, Q.; Song, Y.; Yu, K.; et al. Early Prediction of Mortality Risk among Patients with Severe COVID-19, Using Machine Learning. *Int. J. Epidemiol.* **2021**, *49*, 1918–1929. [CrossRef] [PubMed]

14. Nogueira, P.J.; de Araújo Nobre, M.; Costa, A.; Ribeiro, R.M.; Furtado, C.; Bacelar Nicolau, L.; Camarinha, C.; Luís, M.; Abrantes, R.; Vaz Carneiro, A. The Role of Health Preconditions on COVID-19 Deaths in Portugal: Evidence from Surveillance Data of the First 20293 Infection Cases. *J. Clin. Med.* **2020**, *9*, 2368. [CrossRef] [PubMed]
15. Fernandes, G. Pervasive Data Science Applied to the Services Society. Master's Thesis, University of Minho, Guimarães, Portugal, 2019.
16. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 1–13 April 2000.
17. McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, Numpy, and Ipython*, 2nd ed.; O'Reilly Media: Newton, MA, USA, 2017; pp. 3–4.
18. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. Available online: <https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf> (accessed on 20 April 2021).
19. Fandango, A. *Mastering TensorFlow 1.x: Advanced Machine Learning and Deep Learning Concepts Using TensorFlow 1.x and Keras*; Packt Publishing: Birmingham, UK, 2018.
20. Li, S. Building A Logistic Regression in Python, Step by Step. Available online: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> (accessed on 4 April 2021).
21. Vijayarani, S.; Muthulakshmi, M. Comparative analysis of bayes and lazy classification algorithms. *Int. J. Adv. Res. Comput. Commun. Eng.* **2013**, *2*, 3118–3124.
22. Yildirim, S. Naive Bayes Classifier—Explained—Towards Data Science. Available online: <https://towardsdatascience.com/naive-bayes-classifier-explained-50f9723571ed> (accessed on 4 April 2021).
23. Saritas, M.M. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *Int. J. Intell. Syst. Appl. Eng.* **2019**, *7*, 88–91. [CrossRef]
24. Decision Trees—Scikit-Learn 0.24.1 Documentation. Available online: <https://scikit-learn.org/stable/modules/tree.html> (accessed on 4 April 2021).
25. Song, Y.-Y.; Lu, Y. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130–135. [PubMed]
26. Sklearn.Tree.DecisionTreeClassifier—Scikit-Learn 0.24.1 Documentation. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> (accessed on 4 April 2021).
27. Sklearn.Model_Selection.GridSearchCV—Scikit-Learn 0.24.1 Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed on 4 April 2021).
28. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [CrossRef] [PubMed]
29. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: London, UK, 2016.
30. Omid, M.; Mahmoudi, A.; Omid, M.H. Development of Pistachio Sorting System Using Principal Component Analysis (PCA) Assisted Artificial Neural Network (ANN) of Impact Acoustics. *Expert Syst. Appl.* **2010**, *37*, 7205–7212. [CrossRef]
31. Zhu, T.; Lin, Y.; Liu, Y. Synthetic Minority Oversampling Technique for Multiclass Imbalance Problems. *Pattern Recognit.* **2017**, *72*, 327–340. [CrossRef]
32. Lin, W.-C.; Tsai, C.-F.; Hu, Y.-H.; Jhang, J.-S. Clustering-Based Undersampling in Class-Imbalanced Data. *Inf. Sci.* **2017**, *409–410*, 17–26. [CrossRef]
33. Singh, S. Sampling Techniques. Available online: <https://towardsdatascience.com/sampling-techniques-a4e34111d808> (accessed on 5 April 2021).
34. Moro, S.; Laureano, R.; Cortez, P. *ESM'2011. The European Simulation and Modelling Conference*; EUROSIS-ETI: Oostende, Belgium, 2011.
35. Fernandes, C. Smart Cities—Otimização inteligente de parques de estacionamento. Master's Thesis, University of Minho, Guimarães, Portugal, 2020.
36. Cross-Validation: Evaluating Estimator Performance—Scikit-Learn 0.24.1 Documentation. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed on 4 April 2021).
37. Sklearn.Model_Selection.StratifiedKFold—Scikit-Learn 0.24.1 Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html (accessed on 4 April 2021).
38. Ferreira, T.; Fernandes, C.; Vieira, J.; Portela, F. A Pervasive Clinical Intelligent Decision Support System to predict COVID-19 patients' outcome. 2021; Unpublished.

Article

Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities

Albert Weichselbraun ^{1,*} , Philipp Kuntschik ¹ , Vincenzo Francolino ¹, Mirco Saner ² , Urs Dahinden ¹ 
and Vinzenz Wyss ²

¹ Institute for Information Research, University of Applied Sciences of the Grisons, 7000 Chur, Switzerland; philipp.kuntschik@fhgr.ch (P.K.); vincenzo.francolino@fhgr.ch (V.F.); urs.dahinden@fhgr.ch (U.D.)

² IAM Institute of Applied Media Studies, Zurich University of Applied Sciences, 8400 Winterthur, Switzerland; mirco.saner@zhaw.ch (M.S.); vinzenz.wyss@zhaw.ch (V.W.)

* Correspondence: albert.weichselbraun@fhgr.ch; Tel.: +41-81-286-3727

Abstract: Recent developments in the fields of computer science, such as advances in the areas of big data, knowledge extraction, and deep learning, have triggered the application of data-driven research methods to disciplines such as the social sciences and humanities. This article presents a collaborative, interdisciplinary process for adapting data-driven research to research questions within other disciplines, which considers the methodological background required to obtain a significant impact on the target discipline and guides the systematic collection and formalization of domain knowledge, as well as the selection of appropriate data sources and methods for analyzing, visualizing, and interpreting the results. Finally, we present a case study that applies the described process to the domain of communication science by creating approaches that aid domain experts in locating, tracking, analyzing, and, finally, better understanding the dynamics of media criticism. The study clearly demonstrates the potential of the presented method, but also shows that data-driven research approaches require a tighter integration with the methodological framework of the target discipline to really provide a significant impact on the target discipline.

Keywords: Big Data; Web Intelligence; media analytics; social sciences; humanities; linked open data; adaptation process; interdisciplinary research; media criticism

Citation: Weichselbraun, A.; Kuntschik, P.; Francolino, V.; Saner, M.; Dahinden, U.; Wyss, V. Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities. *Future Internet* **2021**, *13*, 59. <https://doi.org/10.3390/fi13030059>

Academic Editor: Carlos Filipe Da Silva Portela

Received: 30 January 2021

Accepted: 21 February 2021

Published: 26 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Recent advances in areas such as Big Data and deep learning have paved the way for the development of Web Intelligence systems that are capable of performing knowledge extraction and data analytics tasks on large and dynamic web and social media corpora in real time. Driven by this success, methods from computer science have expanded to disciplines within the social sciences and humanities, such as business, communication science, economics, healthcare, and even religion. Some researchers have raised concerns about whether the current approach of transferring methods to other fields without considering the target field's theoretical framework, research background, and concepts is really a good strategy for unfolding the full potential of data-driven research approaches [1]. In fact, there are areas within the field of natural language processing that have been doing particularly well, by building upon existing frameworks from other disciplines. In sentiment analysis, for example, many of the most influential researchers draw upon models from psychology and neuroscience [2,3]. The well-known Hourglass of Emotions model and its revisited version [2], for example, blend concepts from psychology, affective neuroscience, and computer science.

Bartlett et al. [4] note that it is telling that computer scientists are rarely called to embrace traditional sociological thought, and they contest the idea that computer scientists should be legitimate interpreters of social phenomena, even if they have been analyzed with data-driven methods. Researchers such as Connolly argue that computer scientists

should receive a more comprehensive training in social sciences [1] to make them better suited for contributing to these fields. Given the wide area of academic disciplines that are considered as social sciences, such a strategy seems challenging and barely actionable in the short term.

This paper, in contrast, proposes a collaborative process that aims at creating a shared understanding between computer scientists and the researchers in the target domain, provides multiple feedback loops to ensure that knowledge extraction and data analytics tasks are well aligned with the underlying theoretical frameworks, and yields results that significantly impact the target domain. The process guides the research design, the collection and formalization of domain knowledge (e.g., as linked open data), data acquisition, data selection, knowledge extraction, and data analytics. It also actively promotes a close collaboration between researchers from different fields to unfold the full potential of their joint research endeavors.

The rest of this paper is structured as follows: Section 2 discusses related research in the fields of Big Data and Web Intelligence. We then provide an overview of the process used for adapting data-driven research methods to other fields. Afterwards, we elaborate on its application to the field of communication science, provide a short discussion of the relevant research framework and background (Section 4), and demonstrate how the process impacts tasks such as collecting domain knowledge (Section 5) and the processes of data acquisition and selection (Section 6), as well as how it affects the choice of knowledge extraction and data analytics techniques (Section 7). Section 8 finally demonstrates the potential of the constructed data analytics platform based on a use case that analyzes the media coverage of the New Year's Eve sexual assaults in Cologne in 2015. The paper closes with the presentation of conclusions in Section 9.

2. Related Work

Big Data and Web Intelligence provide powerful methods for analyzing web and social media content. The potential and capabilities of these data-driven approaches have been successfully demonstrated in many domains, such as political science [5], environmental communication [6,7], financial market analysis [8,9], healthcare [10], and marketing [11].

Ranganath et al. [5] drew upon social movement theories from political science to design a quantitative framework for studying how advocates push their political agendas on Twitter. They used two datasets for analyzing message and propagation strategies, as well as the community structures adopted by these advocates. Chung and Zeng [12] used network and sentiment analysis on Twitter to investigate the discussion on U.S. immigration and border security. The authors uncovered major phases in the Twitter coverage, identified opinion leaders and influential users, and investigated the differences in sentiment, emotion, and network characteristics between these phases.

In the environmental communication domain, Khatua et al. [7] studied the perception of nuclear energy in tweets covering the 2017 Nobel Peace Prize won by the campaign to abolish nuclear weapons and the 2011 Fukushima nuclear disaster. Scharl et al. [6] presented visual tools and analytics to support environmental communication in the Media Watch on Climate Change (<https://www.ecoresearch.net/climate>, accessed on 21 February 2021), the Climate Resilience Toolkit (<https://toolkit.climate.gov/>, accessed on 21 February 2021), and the NOAA Media Watch [13]. All three platforms aggregate and analyze the coverage of environmental topics in different outlets, including news media, Fortune 1000 companies, and social media, such as Twitter, Facebook, Google+, and YouTube. The article discusses (i) the implemented metrics and visualizations for measuring communication success, (ii) monitoring the efficiency and impact of newly published environmental information, programs to engage target groups in interactive events, and the distribution of content through partners and news media, and (iii) tracking communication goals.

However, even in traditional mediums, such as television, where metrics by Nielsen Media Research (<https://www.nielsen.com>, accessed on 21 February 2021) are well-established

standards for audience ratings, Web Intelligence is gaining in importance, since it provides techniques for assessing audience engagement rather than the impact and reach of television programs [14]. Wakamiya et al. [15] and Napoli [14] discussed the advantages of performing complementary analyses of social network activities related to television programs, and Scharl et al. [16] investigated the emotion in online coverage of HBO's *Game of Thrones* in Anglo-American news media, Twitter, Facebook, Google+, and YouTube.

Li et al. [17] drew upon company-specific news articles to study their impact on the movements of stock markets. They concluded that public sentiments voiced in these articles cause fluctuations in the market, although their impact depends on the company as well as the article content. Xing et al. [18] presented an analysis of common mistakes and error patterns within sentiment analysis methods used in the financial domain and provided suggestions on how to counter them. They also provided a comprehensive study of data-driven approaches used for financial forecasting in [9].

Kim et al. [19] investigated the coverage of the Ebola Virus on Twitter and in news media. They created topic and entity networks, computed per-topic sentiment scores, and analyzed the temporal evolution of these networks. Yang et al. [10] developed a recommender system for patients interested in information on diabetes. Their approach was developed on Weibo.com, which is the largest microblogging site in China, and suggested new content based on the users' interests and their attitude towards a topic by considering features extracted from the users' tweets.

The application domain plays an important role in choosing data sources, analytics, and visualizations. Marketing, for instance, often focuses on the discussion of products and product features in online word-of-mouth channels by applying techniques such as opinion mining and conjoint analysis. Xiao et al. [11] extracted consumer preferences from product reviews and used an economic preference measurement model to derive and prioritize customer requirements at a product level based on this information.

A common theme of the work presented above is the expansion of data-driven approaches to other research fields, particularly to social sciences. As outlined in the introduction, such approaches have been highly successful, but have also raised concerns regarding their efficiency and legitimacy in terms of impact on the target domain [1,4]. The presented paper aims at addressing these concerns by proposing a collaborative process that promotes a shared understanding of the research framework and an alignment of hypotheses and goals between the disciplines, as outlined in the next section.

3. Method

Computer scientists that adapt data-driven research methods to other fields often have only a limited understanding of the theoretical framework that guides research within the target domains. Although such knowledge might not be strictly necessary for applying data science to new fields, it seems sensible to suggest that aligning data-driven research with the concepts, research questions, and methodological framework of the target domain will improve the efficiency, effectiveness, and impact of the research outcomes within the target discipline.

This section introduces an iterative process that supports this alignment by promoting a collaborative, interdisciplinary approach that leverages expert knowledge. We have successfully applied this process to a number of interdisciplinary research projects covering domains such as business ethics, communication science, investment, and pharmaceutical drug development.

Figure 1 illustrates the proposed process, which consists of five main tasks that trigger corresponding feedback loops used to align the research design, goals, hypotheses, and data-driven research methods with the target domain and to consequently improve the quality and impact of the created artifacts. The following subsections elaborate on these tasks in greater detail, and are followed by a comprehensive discussion of how this process has been applied to the creation of the Swiss Media Criticism portal in Sections 4–8.

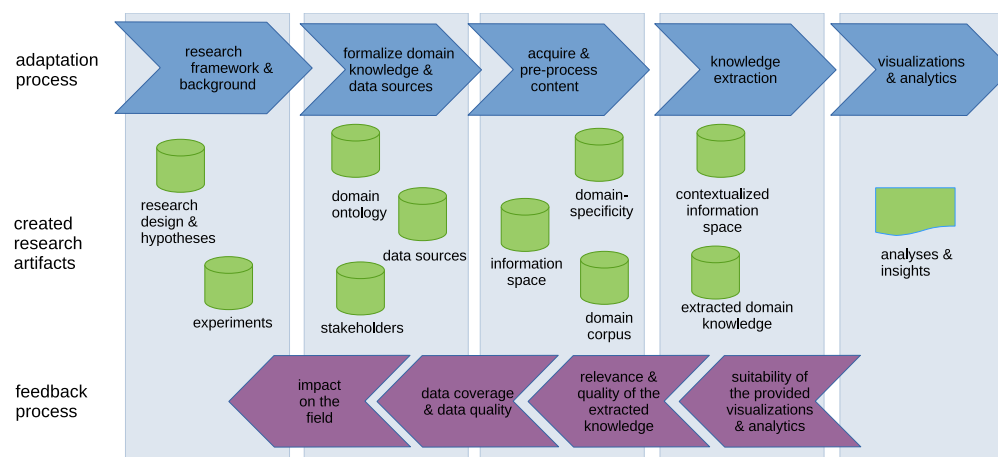


Figure 1. Adapting data-driven research to a new target domain.

3.1. Adaptation to the Research Framework and Background of the Target Domain

The first step aims at creating a joint understanding between computer scientists and researchers from the target domain. It involves introducing the background, research framework, and hypotheses to computer scientists, and communicating the available methods, their potential, and their limitations to the colleagues from the target domain. This stage should also consider how existing concepts from the target domain might support downstream data acquisition, knowledge extraction, and analysis. Finally, the research design, hypotheses, and experiments that help in either confirming or rejecting these hypotheses are created in this step.

An illustrative example of the impact of this first adaptation step is given in Section 4; the definition of media criticism within the communication science literature, which requires a particular stakeholder to voice criticism, had a significant impact on the approach used to detect media-critical content in online news and social media. The domain's research framework has also been instrumental in determining:

1. The required domain knowledge on stakeholders and sources (Section 5), which, in turn, influenced:
2. The sources and source groupings considered by data acquisition components (Section 6) and
3. The types and kinds of entities supported in the named entity-linking process (Section 7.1);
4. Whether a domain-specific affective model or standard sentiment should be used in the knowledge extraction pipeline (Section 7.2);
5. The approach used for data analytics, i.e., investing considerable effort into building the Swiss Media Criticism portal and analytics dashboard, which supports real-time tracking of emerging issues in addition to historical analyses (Section 8).

3.2. Formalization of Domain Knowledge and Selection of Relevant Data Sources

Afterwards, domain experts formalize domain knowledge and collect relevant data sources. The importance of this step cannot be overstated, especially since all later analyses will draw upon data retrieved from sources specified in the formalization step. The domain experts, therefore, need to clearly state which sources are relevant and useful to the analyzed domain and are required for answering the research questions. At the end of this task, they provide:

- Domain knowledge, such as ontological knowledge on entities (e.g., relevant stakeholders, locations, events, etc.) and their relations to each other, which is well suited for supporting data acquisition and knowledge extraction processes. Approaches for computing the domain specificity (Section 6), named entity linking (Section 7.1), and sentiment analysis (Section 7.2) also benefit heavily from domain knowledge.

- Data sources, such as (i) links to relevant web resources (e.g., news media sites), (ii) search terms, and (iii) social media accounts of major stakeholders.

Section 5 outlines the tasks necessary for collecting and formalizing the relevant domain knowledge in greater detail.

3.3. *Acquiring and Preprocessing Textual Data*

The content acquisition process leverages the specified data sources and domain knowledge for content acquisition and filtering. The later step is of particular importance for content sources, such as news media, that cover a broad selection of topics and, consequently, contain both relevant and irrelevant content.

Criteria and concepts from the research framework (e.g., the concept of media criticism) and formalized knowledge (e.g., stakeholders, domain concepts, etc.) from the previous step are instrumental in ensuring that only relevant documents are included in the information space on which analytics will be performed. Depending on the chosen approach, domain experts might provide (i) regular expressions for white- and blacklists (i.e., text patterns that identify relevant and irrelevant content) or (ii) gold-standard documents for training supervised machine learning components and deep learning, or (iii) might ask computer scientists to apply more advanced techniques that combine these approaches, such as ensemble methods.

Experts then browse this corpus in order to evaluate the acquired documents in terms of (i) relevance, (ii) coverage, and (iii) quality (i.e., whether the content is complete and free of noise, such as navigation elements). The feedback on the data quality triggers (i) adaptations of the domain-specificity component, which aims at improving the document relevance and/or coverage (if relevant documents have been filtered), (ii) the inclusion of additional sources to address missing document sources, and (iii) the optimization of the boilerplate removal to improve the document quality.

This iterative feedback process is also a good point to reflect on the necessity of the selected sources. Within the presented use case, for example, we observed that especially web sites of small media outlets tended to violate web standards and caused problems with the boilerplate removal. This observation, combined with the insight that these sites do not contribute a relevant amount of media-critical content and, therefore, have no real impact on the validation of the research hypotheses, led to the decision to remove them from the list of data sources, yielding a better overall content quality.

3.4. *Knowledge Extraction*

Once the data acquisition and preprocessing pipeline has been established, knowledge extraction processes (Section 7) draw upon methods such as (i) named entity linking to identify persons, organizations, and locations relevant to the use case, (ii) sentiment analysis to determine the polarity (i.e., positive versus negative coverage) of the retrieved documents, topics, and stakeholders, and (iii) keyword analysis to extract topics and concepts covered in these documents. The contextual information obtained from the knowledge extraction pipeline yields annotations that form the contextualized information space [20], which is then used for further analyses. Domain experts provide feedback on the annotation quality to aid improvements of the underlying methods, as well as the modeling of the relevant stakeholders.

3.5. *Visualizations and Analytics*

In the last steps, computer scientists draw upon data analytics and visualizations to obtain the results required for verifying or rejecting research hypotheses, to conduct experiments, and to generate insights that are relevant to the target domain. The outcome of this step is not limited to one-time analyses, but may also comprise the creation of expert systems, such as Web Intelligence dashboards, that equip researchers with powerful tools for continuously monitoring relevant web and social media coverage to gain additional insights into stakeholders, new trends, and factors that drive these developments. Finally,

as the case study in Section 8 demonstrates, domain experts are indispensable for interpreting insights generated by data-driven methods and for relating them to the target domain's theoretical framework.

4. Use Case—Analyzing Media Criticism

The following sections demonstrate the application of the introduced process to research in the field of communication science that focuses on media criticism.

As suggested in Section 3, we start by outlining the importance of the topic and the relevant research background within the target discipline. Afterwards, we elaborate on how this background is used to guide the collection of domain knowledge (Section 5), acquire relevant content (Section 6), adapt knowledge extraction methods to the given use case (Section 7), and, finally, create an expert system that is tailored towards performing analyses that help in answering research questions within the domain's theoretical framework.

4.1. Research Background

Mass media play a pivotal role for democratic societies. However, a number of recent national and international debates (e.g., on the media coverage of the 2020 US election and the COVID-19 pandemic) have shown that the performance of mass media is highly contested and trust in journalism is on the decline. A study that was published by Gallup in September 2020 (<https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx>, accessed on 21 February 2021) shows that only 9% of U.S. citizens have a great deal of confidence in media reporting, while more than 60% of the respondents described their confidence in news media as “not very much” or “none”. The performance of the media is also highly contested in Germany, where the term *Lügenpresse* (lying press) was elected as the worst German word of 2014 (https://www.unwortdesjahres.net/fileadmin/unwort/download/pressemitteilung_unwort2014.pdf, accessed on 21 February 2021) and is still used by some ideological groups.

Measuring and understanding the issues, dynamics, and impact of media criticism is a challenging task that can tremendously benefit from systematic studies that cover media criticism from major stakeholders, such as (i) mass media, (ii) media-critical agents, and (iii) social media across different outlets and media. An analysis of the daily output produced by these stakeholders makes it clear that it is no longer feasible to manually collect and analyze these document streams by hand and that a collaborative research design that integrates methods from computer sciences is required.

4.2. Research Framework

Journalism possesses substantial definatory power due to its selection of fragments of reality and the resulting staging of it. However, journalistic descriptions of the reality are at least partly subjective and depend on a number of long-term factors, such as the journalist's socialization, know-how, and self-conception, but also on short-term circumstances, such as the available time for production, events occurring shortly before or after, and the accessible resources (experts, pictures, etc.).

Moreover, citizens depend on trustworthy journalism, which can be achieved by periodic information about current processes in the sector, journalistic work routines, or dominant pressures [21]. Substantial media criticism also empowers the public to overcome its role as an exclusive consumer and enables it to acquire the role of a media-literate agent and citizen who shoulders responsibility for the media system's status quo and quality [22]. Therefore, society cannot go without a continuous, public, and critical debate about journalistic performances [23]. We understand media criticism as optimistic-constructive or negative, public, and reflexive observation, description, and evaluation of media process routines, concerning all relevant participants, referring to accepted rules and standards [21,24]. Crucial is an explicit assessment of a relevant media issue concerning products, agents, actions, or procedures [21,25].

Up to now, academic research lacks a systematic inventory and quantitative empirical basis of editorial-based media criticism, as well as of agents and institutions that practice public media criticism, aside from [26,27]. What is the yearly output of media-critical stakeholders? On which issues do they focus? What resonance do existing agents encounter in mass media? Moreover, little is known from a scientific point of view about the issues, dynamics, and impact of media criticism debates. What seems obvious is that circumstances for editorial-based media critics are deteriorating due to ongoing concentration processes within the media system [28]. Furthermore, on the basis of content analyses, Wyss, Schanne, and Stoffel [27] hint at the often episodic character of media criticism and deplore the absence of explicit evaluations and statements that focus on structural deficits. It remains uncertain whether other agents are able to fill this gap with their own qualitative and systematic coverage. Although the number of critical agents and institutions has increased since digital channels have spread, at least some of them are highly dynamic and show variable lifespans, such as media blogs [29–31]. The societal relevance of media criticism along with recent developments, such as the declining trust in journalistic stakeholders, the structural changes in the media industry, and the neglect of media criticism by communication research [32,33], emphasizes the importance of a systematic scientific analysis of this topic.

4.3. Conclusions

The discussions of research questions, research framework, and the available data analytics capabilities yielded the following insights that directly influenced the project's research design: The dynamics of media criticism are of particular interest to communication science. We, therefore, decided to develop an expert system that acquires, identifies, and monitors media-critical coverage in real time, enabling historical analyses as well as active tracking of current issues. Due to the importance of research questions that focus on the output and impact of different media-critical actors, we have defined three groups from which media-critical coverage will be collected (Section 5.3). The data acquisition component will also draw upon the discipline's definition of media criticism for determining whether an article is relevant for the analysis (Section 6). Finally, the knowledge extraction components (Section 7) will perform named entity linking to identify major stakeholders, phrase detection to automatically obtain associations that provide clues on important topics and the framing behavior of agents, and sentiment analysis to gather insights on whether issues are perceived as positive or negative. Based on these design decisions, the created expert system will allow an efficient and effective analysis of relevant issues, stakeholders, their output, and the impact of their criticism.

5. Collecting Domain Knowledge and Data Sources

5.1. Domain Knowledge on Stakeholders

The research framework emphasizes the importance of gathering qualitative and quantitative insights into (i) the output of media-critical stakeholders and (ii) the issues, dynamics, and impacts of media-critical debates. Consequently, domain experts compiled stakeholder lists that contain media entities (persons and organizations) of all four Swiss language regions, as well as key organizations and persons of foreign countries. Information on international stakeholders was provided because Swiss German media-critical coverage can potentially also refer to entities outside of Switzerland. In addition to active organizations, historical entities were part of the list, as well as existing agents that have not produced any publication output in the last few years. The domain experts also provided background information on entities, such as abbreviations, addresses, and key people. In total, these efforts yielded the object lists outlined in Table 1.

Table 1. Categories of the media-critical entity lists and the corresponding numbers of entities.

Entity Category	Entities
Swiss Stakeholders	1209
Media-critical agents	65
Mass media (print, radio, TV, online)	524
Publishing houses and print offices	85
Publishers, CEOs, and editors in chief	344
Media events and media awards	21
Syndicates and associations	11
Media scholars and (commercial and non-profit) research organizations	50
Foundations and media schools	22
Media politicians and federal councillors	13
Media lawyers	26
Media journalists	48
Foreign stakeholders	39
Selected organizations and persons from foreign countries	39
Miscellaneous	462
Information programs of the Swiss Broadcasting Corporation (SRG)	64
General media-critical keywords (media-relevant terms)	398

In addition, the experts assembled a general keyword list of about four hundred concepts that might indicate media criticism, which contains terms such as journalism, mass media, newspaper, broadcaster, editorial, circulation, tabloid press, practical training, audience rating, or gatekeeper. These terms were manually extracted from a set of media-critical articles that were used by the domain-specificity components (Section 6) and from literature covering communication and media studies. This keyword list was used by the mentioned domain-specificity components as another indicator for media criticism.

5.2. Formalizing Domain Knowledge

The domain experts provided the collected domain knowledge on media-critical stakeholders in a tabular form, specifying information such as the stakeholder's name, possible abbreviations, Twitter accounts, homepages, organization, and type. These tables were converted into the Resource Description Framework (RDF) linked data format, which is more easily interpretable by automated processes and serves as input for the named entity linking component (Section 7.1).

Maali et al. [34] introduced the Publishing Pipeline for Linked Government Data, which utilizes Open Refine (<https://openrefine.org/>, accessed on 21 February 2021) together with an RDF extension (<https://github.com/stkenny/grefine-rdf-extension>, accessed on 21 February 2021) to convert tabular data into RDF. We simplified this pipeline to contain only the steps “Data clean-up” and “Transformation into RDF”, and subsequently adopted it for the targeted lightweight RDF graph used further along in the process.

The main advantages of the outlined procedure are (i) its simplicity and (ii) that the configuration used for the conversion pipeline can be exported, edited, and reused. As such, it is only necessary to create this pipeline once, since it is possible to just reapply it on an updated dataset.

5.3. Selecting Relevant Data Sources

Gaining a comprehensive picture of publicly performed online media criticism requires analyzing the publication output of opinion-leading media-critical agents (press releases, news features, blogs, project reports) and their response rate in relevant mass media and specialized publications focusing on media. In this context, “publicly performed” means mass media online distribution, as well as publications that are accessible on organizational web sites, available for everyone, and, hence, usable cross-systemically [21]. Based

on these insights, the domain experts organized sources of media-critical content into three source categories, as outlined in Table 2: mass media, professional public, and social media.

Table 2. Sources per source type for media-critical content.

Source Type	Sources
Mass media	185
Professional public	170
Social Media	
Media organizations	180
Journalists and media stakeholders	740
Total	1275

1. A total of 185 mass media sources were identified by the experts, comprising web pages of radio stations, TV stations, and printed media with an edition of at least 15,000 copies, as well as online only media. In addition to editions, crucial factors concerning printed media were timeliness, universality, and periodicity. This category also contained well-established TV and radio programs (i.e., programs that have been broadcasted for at least five years).
2. The professional public category gathered articles from 100 Swiss German media-critical agents participating in the public discourse related to media criticism and the corresponding press releases. It included a heterogeneous range of organizations that are either part of the media system (intra-media agents) or belong to another societal system (extra-media agents) [27]. The resulting list contains 170 agents with approximately 100 harvestable URLs.
3. Based on the domain experts' assessment, our research considered two different types of social media accounts: The first one collected tweets from mass media sources, and the second one collected tweets from media- and journalism-related persons. For inclusion in the Twitter sample, profiles needed to fulfill the following three criteria: (i) at least 100 followers, (ii) a relation to Swiss media criticism, and (iii) the majority of the tweets must be written in German. The system monitored 180 Twitter accounts of mass media and 740 accounts of Swiss journalists and media-related persons.

A minor drop-out on the web sources could be determined due to the lack of relevant or up-to-date content, or because content was secured by a paywall or offered exclusively as a podcast, ePaper, or another non-trivial data format. In the case of mass media, program preview sites and audio-only content were removed as well. In the case of paywalls, a subscription for important outlets was organized.

6. Data Acquisition and Preprocessing

Based on the information provided in the previous step, our content acquisition pipeline drew upon web crawlers and the Twitter web API to retrieve potentially relevant content. Since between 50 and 60% of a typical web page consists of noise, such as navigation menus, links to related documents, advertisements, and copyright notes, the content acquisition processes also deploy boilerplate removal [35] to identify and remove noise elements as well as overview pages (i.e., summarization pages and entry points).

Afterwards, a domain-specificity classifier ensures that only relevant documents are included in the corpus (or information space) by filtering irrelevant documents. The information space used for analyzing media criticism, therefore, will contain mostly media-critical documents rather than arbitrary media coverage. The following sections describe the evolution of the domain-specificity component within the Swiss Media Criticism project.

6.1. Keyword-Based Approach

The first version of the content acquisition pipeline drew upon black- and whitelist items to determine the domain specificity of documents. Keyword-based queries are very

common in social sciences and have the advantage of being well accepted within this discipline. Nevertheless, they provide low performance in terms of recall, since media criticism can be voiced in manifold ways and settings, ranging from sport events like the ski accident of Michael Schuhmacher in 2013, to tragic disasters in aviation like the Germanwings accident in 2015, to political statements like the Böhmermann affair in 2016, and to public health and policy issues, such as the reporting on the COVID-19 crisis. Knowledge-aware text classification systems address this problem by considering both the document's vocabulary and background knowledge, such as the presence of media-critical entities and terms in the classification process.

6.2. Knowledge-Aware Text Classification

The next iteration of the domain-specificity component drew upon the formalized domain knowledge to calculate the probabilistic affiliation of a new and unknown text with a given set of categories. In the presented project, two categories—media criticism and not media criticism—were used, corresponding to relevant and irrelevant content.

To gather the needed domain-specificity information, domain experts (i.e., communication scientists) collected a gold-standard corpus of both media-critical and non-media-critical documents. Special attention was paid to finding document pairs dealing with the same topic, where one document contained media criticism while the other one did not, as this helped to minimize the risk of topic-specific bias. In addition, the domain experts aimed to include a wide range of topics, such as sports, direct democracy, affirmative actions, disabilities, the Holocaust, offshore leaks, the Pope, Islam, climate change, the energy transition, and airplane accidents, in the gold standard. All selected documents were published in a Swiss medium after 2010. The media-critical documents in the sample only considered texts for which media criticism was clearly identified as the main topic and was not just peripherally mentioned. The final gold standard comprised 503 media-critical documents and 643 corresponding non-media-critical counterparts, which were used for training the classifier.

This iteration used the Naïve Bayes classification algorithm due to its simplicity and explainability, which allowed the determination of the reasons for a correct (or incorrect) classification result. Being able to observe and correct the process if necessary helped in building the domain expert's trust in the classifier, and this was therefore identified as a crucial step in creating the system.

The overall probability of an unknown text was derived from the probabilities of the contained terms: First, the document was converted into a “bag of words” representation that also considered n-grams and skip-grams. A stopword filter and a frequency-based filter, which remove terms that have been used less than three times in all collected documents, were applied to speed up the computation time. The prior probability for each element of this bag of words was received from the knowledge base calculated previously. The overall probability of an article being media critical given its included terms $P(M|t)$ was calculated with the cross-product of its terms' prior probability $P(M)$ and the likelihood $P(t|M)$ that they were contained divided by its evidence $P(t)$.

$$P(M|t) = \frac{\text{priorprobability} \times \text{likelihood}}{\text{evidence}} = \frac{P(M)P(t|M)}{P(t)} \quad (1)$$

In this scenario, prior probability describes the general probability of an unknown text containing media criticism according to the training data, while likelihood refers to the probability of an unknown text's term being related to a media-critical text. Meanwhile, evidence means the probability of an unknown text's terms being contained in any of the trained categories (in this case media criticism M and not media criticism $\neg M$).

$$\text{evidence} = P(t) = P(M)P(t|M) + P(\neg M)P(t|\neg M) \quad (2)$$

We further introduced ensemble heuristics, where the classification functions as a baseline, and further knowledge sources, such as (i) source whitelists, (ii) black- and whitelists, and (iii) named entity annotation and text patterns, function as regulators. Documents that contain no or only a few keywords received a penalty on the estimated probability of containing media criticism. The ensemble methods applied in this iteration achieved results with a recall of over 71%, which is already considerably better than the keyword-based approach.

6.3. Deep Learning for Text Classification

The final iteration of the domain-specificity component drew upon a transformer architecture to perform the classification process. As outlined in Figure 2, the classifier used a Bidirectional Encoder Representations from Transformers (BERT) language model [36] with twelve transformer layers and a maximum sequence length of 512 tokens to create a contextualized representation of the input document, in which each token was represented by a 768-dimensional vector. Text classification tasks usually only consider the BERT symbol for classification output ([CLS]), which starts each document and provides a 768-dimensional vector representation of its content. Since combining multiple hidden layers has been shown to improve accuracy [36], we pooled the output of the last four hidden layers and fed it through two linear layers, which then provided the final classification result.

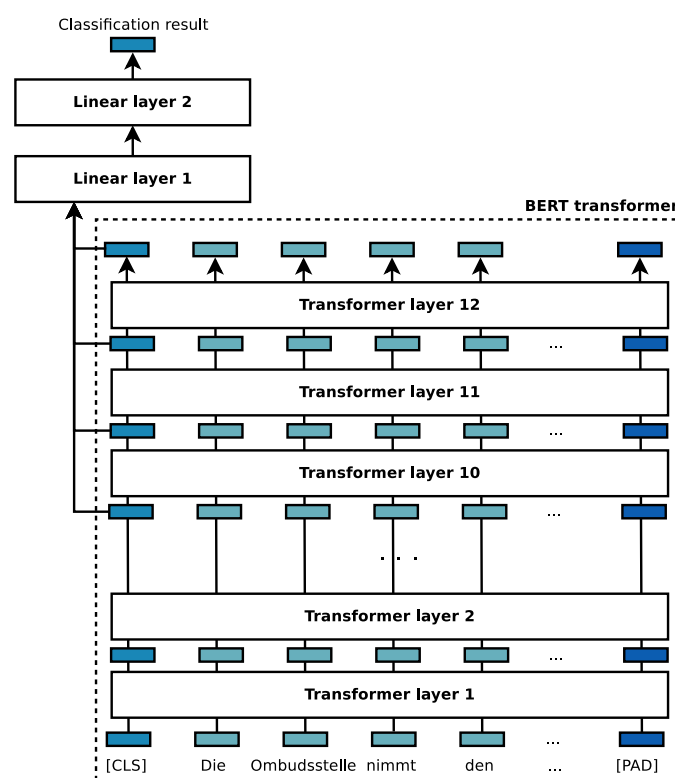


Figure 2. Architecture of the deep learning classifier.

We used the Hugging Face transformer library (<https://huggingface.co>, accessed on 21 February 2021) in conjunction with PyTorch (<https://pytorch.org>, accessed on 21 February 2021) to implement the classifier and evaluated it using the following two pre-trained transformer models:

- German BERT base (<https://huggingface.co/bert-base-german-cased>, accessed on 21 February 2021): A case-sensitive BERT transformer that has been trained on over 10 GB of textual data comprising the German Wikipedia dump (6 GB), the OpenLegalData dump (2.4 GB), and 3.6 GB of news articles.

- Multilingual BERT (<https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>, accessed on 21 February 2021): A case-insensitive, multilingual BERT model that has been fine-tuned for sentiment analysis on product reviews in English, Dutch, German, French, Spanish, and Italian.

Experiments that aimed at optimizing the model's hyperparameters yielded the following settings: a batch size of four, a dropout value of 0.3 to prevent overfitting, an Adam optimizer with a learning rate of $\alpha = 3 \times 10^{-5}$, a binary cross-entropy (BCE) loss function, and a training limit of 35 epochs with early stopping if the validation loss does not improve.

Both models were fine-tuned on the domain-specificity classification task using the gold standard created by our domain experts. In our experiments, the German BERT Base model achieved an F1 score of 92.5% with both a precision and a recall of 92.5%. The multilingual BERT model even topped this performance with an F1 score of 95.8% (recall: 95.6%, precision: 95.8%). This considerable boost in performance came at the cost of a lower explainability of the provided classification results.

7. Knowledge Extraction

The selection of the knowledge extraction techniques was guided by the research framework discussed in Section 4, which requires (i) the analysis and tracking of major stakeholders, (ii) the identification of dominant issues, and (iii) the classification of media criticism as either optimistic–constructive or negative. Weichselbraun et al. [37] discussed the use of domain-specific affective models, which support capturing emotions that go beyond standard sentiment and emotion models. An assessment that compared the optimistic–constructive and negative dimensions from communication science literature with standard sentiment polarity (dimensions: positive and negative) concluded that, for the purpose of the joint research project, the use of sentiment polarity is an efficient (availability of high-quality sentiment lexicons) and effective (sufficiently high correlation between both metrics) strategy.

Consequently, we deployed named entity linking for identifying stakeholders (Section 7.1), the phrase extraction method described by Weichselbraun et al. [38] for tracking associations and dominant issues, and sentiment analysis (Section 7.2) to automatically determine whether the feedback was positive or negative.

7.1. Named Entity Linking

Named entity linking identifies mentions of named entities, such as persons or organizations that are important stakeholders in the public discourse on media criticism, as well as locations, and links them to structured knowledge sources, such as the DBpedia, GeoNames, and custom linked open data repositories. Therefore, it paves the way for analytics that assign sentiments to entities, identify trends, and reveal relations between these entities. Transforming the domain knowledge on media criticism assembled by the domain experts to a linked data format (Section 5.2) enables us to leverage these data for named entity linking.

The webLyzard platform used in this project draws upon Recognize [39], a named entity linking component that queries linked open data sources to obtain textual, contextual, and structural information on entities, which is then used to identify entity mentions in text documents. Figure 3 outlines this process. Recognize uses analyzers, i.e., graph mining components, that retrieve the relevant information from the available linked data sources.

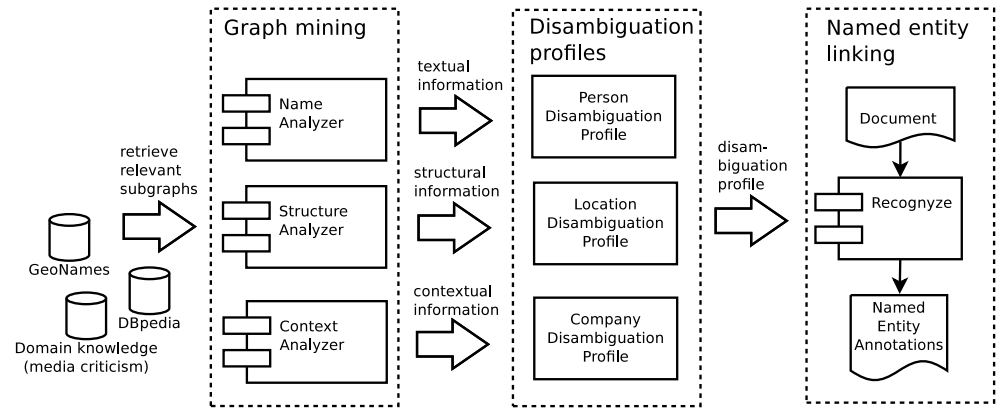


Figure 3. Named entity linking with Recognize.

Name analyzers yield textual information that comprises different named variants and abbreviations used to refer to an entity (e.g., BBC, British Broadcasting Corporation, etc.). Context analyzers mine contextual information, such as keywords obtained from the organization’s description, products and services offered by an organization, its web address, and mail and email addresses. Structural information is obtained by structure analyzers, which reveal relations between the extracted entities—for instance, that John Reith was the founder of the BBC or that Future Media is an operational division of the BBC.

The obtained information was then used to create disambiguation profiles that identified potential named entity mentions in textual content, thus helping to disambiguate these mentions and ground them to the correct entity in the linked open data repository.

7.2. Sentiment Analysis

Sentiment analysis computes the polarity (positive versus negative) of targets, such as documents, sentences, topics, and named entities. Therefore, it is useful to determine how targets are perceived by the public or to identify conflicting targets, i.e., targets with a high standard deviation of the sentiment value, which indicates that the coverage of these targets is framed differently in the analyzed outlets.

The Swiss Media Criticism portal uses a context-aware sentiment analysis approach to determine the text sentiment, which considers the text’s context prior to evaluating its sentiment. Therefore, it combines a static sentiment lexicon that contains sentiment terms t_i that either indicate a positive or negative sentiment (e.g., good and excellent versus bad and horrible) with a contextualized sentiment lexicon, which determines the sentiment value for ambiguous sentiment terms ($t_i \in T_{ambig}$) based on the text’s context C_i , expressed as a set of non-sentiment terms within the text. The term expensive, for instance, is considered negative in conjunction with context terms such as overpriced, while the context terms high value and quality might indicate a positive usage of this term.

Aggregating the sentiment value of all terms t_i within a sentence with context C_i yields the total text’s sentiment (Equation (3)).

$$s_{\text{sentence}} = \sum_{i=1}^n \mathcal{N}(t_i) \cdot s''(t_i, C_i) \quad (3)$$

The sentiment lexicon $s''(t_i, C_i)$ yields the contextualized sentiment value for ambiguous sentiment terms t_i and falls back to the static sentiment lexicon $s(t_i)$ for unambiguous sentiment terms or if no context terms are available.

$$s''(t_i, C_i) = \begin{cases} s'(t_i, C_i) & \text{if } t_i \in T_{ambig} \text{ and } C_i \neq \emptyset \\ s(t_i) & \text{otherwise} \end{cases} \quad (4)$$

A function $\mathcal{N}(t_i)$ considers negations by inverting the term sentiment if t_i has been negated.

$$\mathcal{N}(t_i) = \begin{cases} -1 & \text{if } t_i \text{ is negated} \\ +1 & \text{otherwise} \end{cases} \quad (5)$$

The system also propagates the sentence sentiment to topics, sources, and entities, i.e., it allows assessment of whether a certain entity occurs frequently in a positive or negative context or how a certain new media site frames a particular topic.

8. Visualizations and Analytics—The Swiss Media Criticism Portal

The expert system that was created, the “Swiss Media Criticism portal and analytics dashboard”, was developed within the Radar Media Criticism Switzerland project, funded by the Swiss National Science Foundation, and was built upon the webLyzard platform (<https://www.webylizard.com>, accessed on 21 February 2021), which provides components for scalable knowledge acquisition and extraction [40], advanced analytics [16], and visualizations [6]. The expert system has been actively used by communication scientists from the project consortium, and efforts towards extending this system to further countries, domains, and research groups are planned.

Figure 4 illustrates the Swiss Media Criticism portal and analytics dashboard, which enables users to search, refine, analyze, and interpret media-critical coverage from Swiss online sources.

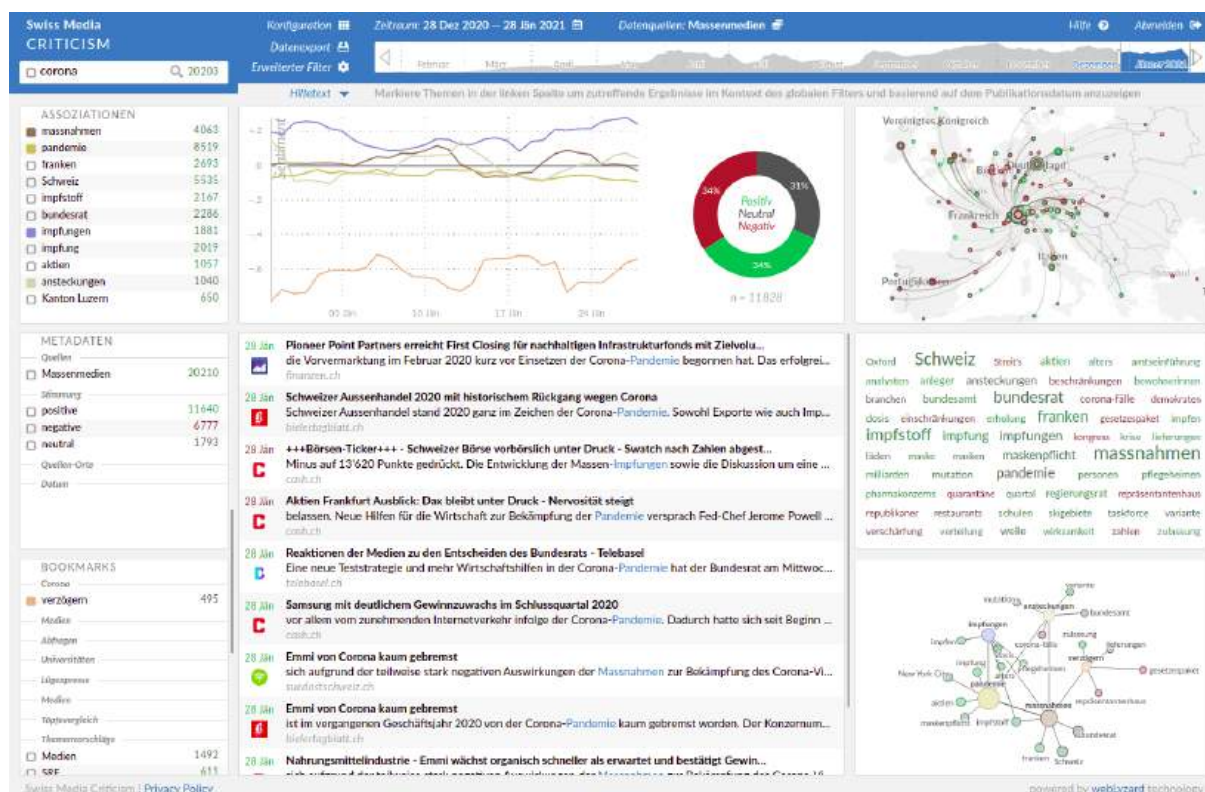


Figure 4. Screenshot of the Swiss Media Criticism portal featuring a query for the search term “corona”.

The system aggregates media-critical content and yields keyphrase statistics (left bottom), relevant documents (center), real-time analytics, such as the geographic distribution of the media coverage (right top), a tag cloud highlighting terms associated with the current query (right middle), and a keyword graph outlining how the associated terms are connected to each other (right bottom). The case study in the following section outlines how the computed associations and the corresponding visualizations provide clues con-

cerning the framing behavior of the agents and support the development of hypotheses that describe the effects that the coverage has among the target audience.

The created expert system provides powerful analytics for addressing the major issues and questions raised within the research framework:

- The platform provides real-time coverage of current (e.g., Figure 4) and past (e.g., Section 8.1) issues, thus enabling experts to analyze the dynamics of both current and past issues and to participate in the discourse with data-driven insights.
- Stakeholder groups (i.e., mass media, professional public, and social media) are organized in different samples, allowing domain experts to analyze and contrast the output and impact of these groups.
- Named entity linking enables analyses that focus on the stakeholder groups that have been defined by the domain experts (Section 5).
- Phrase extraction automatically identifies terms and concepts that are associated with stakeholders, locations, and queries, supporting domain experts in uncovering important topics, dominant issues, and their framing.
- Sentiment analysis provides insights into the perception of stakeholders and issues.
- Drill-down analyses ensure that aggregated results and trends presented in the portal are valid and help in understanding the underlying reasons for the observed effects.

The following case study demonstrates how these analytics and visualizations are instrumental in understanding the media-critical discourse and its dynamics by supporting analyses that (i) aid in answering fundamental questions on the temporal course and lifespan of issues, (ii) highlight important agents participating in the discourse, and (iii) identify the sentiment (positive versus negative) of its coverage.

8.1. Case Study: New Year's Eve Sexual Assaults in Cologne

During the night of New Year's Eve in 2015/2016, numerous women were robbed and sexually molested at the Cologne main station. Generally, the suspects were described as African- or Arabic-looking. A public debate emerged over immigrants, sexism, and cultural values. In this context, the role of mass media coverage of delicate topics was critically discussed, and the term *Lügenpresse* (lying press) was suddenly socially acceptable again.

Searching for media-critical coverage of these events using the keywords *Silvesternacht* (night of New Year's Eve), *Köln* (Cologne), and *Medien* (media) during the period from 31 December 2015 to 31 March 2016 yielded 72 documents in the mass media and professional public categories, as well as 227 Tweets from media-critical stakeholders.

Figure 5 compares the media-critical coverage in mass media and in Tweets authored by media-critical stakeholders.

For mass media, the most relevant concepts were *Köln* (Cologne, 36 mentions), *Täter* (offenders, 36 mentions), and *Übergriffe* (assaults, 16 mentions), indicating that the discourse was focused on these particular events.

The discussion of media-critical stakeholders, in contrast, focuses on the implications of the events for Switzerland, as indicated by the keywords *Frauen* (women, 17 mentions), *Schweiz* (Switzerland, 12 mentions), and *Durchsetzungsinitiative* (an upcoming people's vote regarding the deportation of criminal foreign citizens, nine mentions). It is also interesting to note that the term "refugeeswelcome", which was present only in the Twitter tag cloud, became a hashtag for some of these discussions.

The geographic distribution of the locations mentioned in the articles indicates that the coverage focused on Cologne, Stuttgart, and Sweden, where similar events were reported, as well as in Poland due to an article that covered the drastic means by which Poland's government tried to strengthen its influence on the media, mentioning the Polish media coverage of the events in Cologne.

Figure 6 reveals the topic's life cycle, which seems to be typical for mass media coverage, with a heavy increase in coverage when the issue first became apparent three days after New Year's Day. The mostly episodic mass media coverage, which has been criticized

by media scholars, seemed to continue as far as media-critical coverage is concerned. Media critics seem to operate similarly to their colleagues.

alt angst begriffe berlin besten bitte christofmoser
 debatte deutschland dsi durchsetzungsinitiative fakten
 faznet frage frauen gewalt guter halt junge kritik
 kultur köln lesenswert leser linken mbinswanger medien
 merkel monibol männer nacht nickluethi nzz phwampfler
 polizei problem radio reaktion recht refugeeswelcome richter
 schellmisch schweiz schweizer silvesternacht simongemperi
 srf srfnews svp tagesanzeiger text thema tipps täter
 verhalten wasser watsonnews welt worte zahl zeitonline
 zumindest zuwanderung zürcher zürich

andrej anständig anzeigen ausschaffungsinitiative ausweisung
 automatismus christlichen durchsetzungsinitiative express faz
 flüchtlings frauen gastbeitrag gewaltenteilung grundhaltung
 hauptbahnhof individuen innenpolitische korrektheit kriminelle
 köln lügenpresse maizières massenhaften mindestlöhne
 muslimische nordafrika normalität parteiintern pegida polen
 polizei polnische pressemitteilung prinzipien publizist rassismus
 rechtsstaat reflex schröder selbstzensur sexuelle sexuellen
 silvester soziologe stans tatsachen täter tätern unliebsame
 unterstellungen verfassungsgericht verhältnismässigkeit vorfälle
 vorfällen vorkommnisse waage werte willkommenskultur zeichen
 zeilen zuwanderer zuzug übergriffe übergriffen



Figure 5. Tag clouds of the media-critical coverage of the New Year's Eve sexual assaults in Cologne on Twitter (top), in news media (bottom), and in locations mentioned in the news media coverage (right).

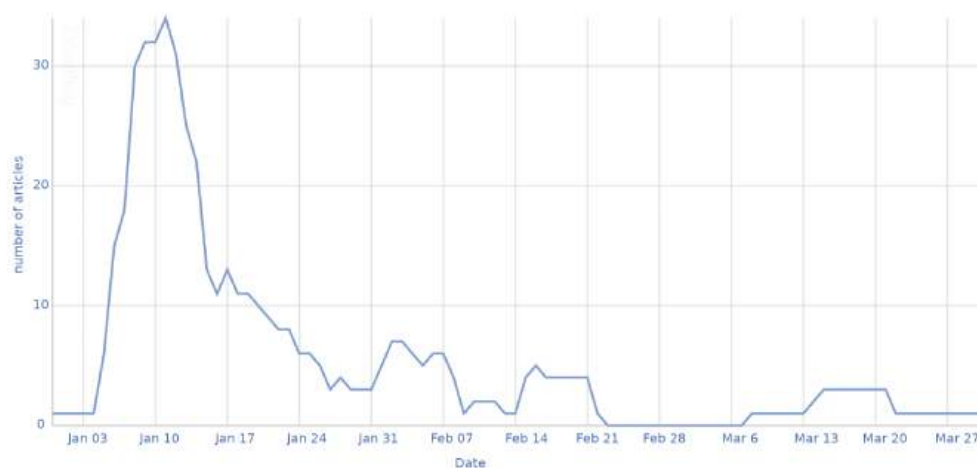


Figure 6. Frequency graph illustrating the topic's life cycle in mass media in the first quarter of 2016.

After a fortnightly high during which the issue was dominant, the media lost interest, as hardly any new information became available or there was a lack of similar follow-up events. The steep decline was followed by a steady loss of importance, which ended in disappearance. When considering that almost three hundred monitored agents only yielded 72 hits over three months, it became obvious that this event did not attract a lot of media-critical coverage in Switzerland. Hence, the hypothesis of media-critical coverage not being widespread is supported. Swiss journalists held themselves back from criticizing their professional colleagues in Germany, but also omitted a discussion of what should be done differently by Swiss media if a comparable event would happen in Switzerland.

At the same time, the system revealed which sources participated in a public debate and how strongly each source had been involved (Figure 7). From a media studies perspective, this feature is relevant concerning the distinction of mass media with institutionalized

forms of media criticism and editorial teams without similar structures or for recognizing differences between tabloid press and quality media. Institutionalization means that there is at least a personal specialization or an organizational beat in the topic field of media journalism. According to structuration theory, institutionalization of media criticism should lead to more regular and more qualitative coverage [41]. Moreover, the coverage's geographic distribution becomes visible; therefore, it is also more discernible if one regional press cooperation reports more often than another.

The deepening analysis in Table 3 points out that only 22 of the 185 mass media titles covered the event from a media-critical point of view. This corresponds to an average of 0.31 contributed articles per title in the category of the mass media. The reaction of the professional media in this case was even lower. Only 8 of the 100 professional public media titles covered the issue. This corresponds to an average of 0.13 contributed articles per title in the category of the professional public. This result shows that mass media criticism cannot be fully substituted with criticism by professional public agents. Moreover, the analysis allows a comparison between tabloid papers and quality press. In the sample, we found 24 articles coming from five quality press titles, but only five articles in five tabloid press titles. We could argue that substantial media criticism is not something that tabloid press is predestined for due to its meta-level nature and its complex context with ethics or law. Therefore, media criticism is a topic that tabloids seem to avoid, fearing they could perform poorly.



Figure 7. Cropped screenshot of the Swiss Media Criticism portal demonstrating the most frequent sources of the selected articles. Included are the source name (*Quelle*), number of documents (*Anzahl*), reach (*Reichweite*), influence (*Einfluss*), and sentiment.

Table 3. Average number of articles per medium.

Category	Articles (Agents)	Average per Medium
<i>Top Categories</i>		
Mass media	59 (185)	0.31
Professional public	13 (100)	0.13
<i>Sub-Categories</i>		
Quality media	24 (5)	4.80
Tabloid press	5 (5)	1.00
Institutionalized media criticism	17 (5)	3.40

The data show that forms of institutionalized media criticism structures within the newsroom do indeed lead to more coverage. In the category of the media with institutionalized media criticism, we found an average of 3.4 contributed articles per title, whereas editorial teams without such specialization published far fewer articles.

Although the analysis shows certain tendencies, it is clear that one case study is not enough to determine whether institutionalization is meaningful or not. Just by skimming the articles, it becomes evident that the expression “lying press” is hardly existent in Swiss mass media coverage. The Swiss journalists do not seem to be willing to support the polemical use of this inadequate term. In addition, mass media only perform “embedded media criticism”, which means that critical content is not the main topic of an article, but is peripherally mentioned in a few sentences.

Figure 8 shows the sentiment for the press coverage of the New Year’s Eve sexual assaults in Cologne. The figure indicates the framing behavior of the involved agents—more precisely, whether an issue is mainly framed with positive, neutral, or negative terms. This allows, for example, the creation of hypotheses regarding the effects of this coverage on the audience.



Figure 8. Sentiment analysis of the topic’s perception in the first quarter of 2016.

In the case of Cologne, once more, we recognized a typical pattern known from general media coverage: There is often a first outrage wave among journalists when such an issue emerges. A highly emotionalized coverage arises, leading to a very negative issue sentiment. Indeed, most of the media-critical Cologne coverage was clearly negatively framed. After a few weeks, the sentiment becomes more neutral, as journalists gain distance from an event and contextualize an incident. Later, the sentiment even becomes positive, as voices often arise that demand that society learns and improves and that things have to be handled differently the next time in order to avoid a similar event. After this conclusion phase, the sentiment stays neutral as media coverage vanishes. If there is a new aspect to be released to the public or a new event occurs that can be associated with the previous one, the neutral phase ends and is replaced by more emotional coverage again.

8.2. Validation of the Results

Validation of automated methods plays a key role in ensuring a high data quality and in obtaining reliable results and insights. Within the Swiss Media Criticism project, we implemented multiple validation routines that have been tailored to meet the specific needs of the domain experts’ use cases, some being fully automated tests, and others requiring heavy human interaction.

For instance, the recall of the data acquisition pipeline (i.e., whether all available media-critical coverage can be found in the expert system) was of particular importance. Therefore, we complemented standard automatized tests that computed the precision, recall, and F1 measure of the domain-specificity components with manual checks, in which domain experts scanned newspapers for relevant articles and verified that the articles were also actually available in the portal.

To evaluate information extraction tasks, such as named entity linking and sentiment analysis, we introduced group exercises into lectures and tasked students with evaluating data quality in seminars. These efforts were complemented by evaluations performed by computer scientists using tools such as Orbis [42] that aid validations of computer-generated annotations by visually comparing them with gold-standard annotations.

The feedback received from the validation steps was collected in an issue management system, which allowed all stakeholders to track the improvements in data quality and reliability, and has been proven to strengthen participation and involvement.

8.3. Discussion of the Case Study

For newly evolving discourses, the Swiss Media Criticism portal provides an automated quantitative overview for crucial parameters and answers to basic research questions in almost real time. What is the temporal course and lifespan of an issue? Which agents participate to which extent in an ongoing discourse? How is an issue framed and which sentiment does it encounter during coverage? In addition, the portal instantly delivers a sample of articles that can easily be used and modified for more in-depth, manual content analysis, as search results are exportable to various data formats.

Several benefits can be highlighted from the application of data-driven research methods to the domain of media criticism: (i) availability of a complete inventory, such as extensive document repositories, (ii) volatile sources of information (such as Twitter and comments in forums) can be captured, and (iii) the efficiency of the content analysis is much higher than with conventional methods.

In addition, the system provides powerful analytics and visualization to gain a better understanding of the target domain, of spatial and temporal effects, and of the framing of topics due to the computation of associations and the sentiment. The Swiss Media Criticism portal also supports exporting documents and visualizations, enabling the application of further linguistic and statistical methods to the document corpus.

9. Conclusions

Recent literature has raised serious concerns about the effectiveness of computer scientists that apply methods from their field to other disciplines without a proper understanding of the necessary research background within the target domain [1,4]. This article addresses these concerns by suggesting a strong collaborative adaptation process that has been developed within a number of research projects and that aids interdisciplinary research groups in building a common understanding of (i) the research framework within the target discipline and (ii) the benefits that data-driven research methods could yield. The process guides researchers in:

- Aligning their research design and hypotheses with the target discipline's research framework and background to ensure that the envisioned research has a real impact on that discipline;
- Leveraging theories and concepts from the target domain in the design of indicators and metrics;
- Considering the target domain's research framework in the development of the entire data acquisition, processing, and analytics process and integrating domain expert input (e.g., on formalized domain knowledge and data sources) into their development;
- Organizing the collaboration between the groups by defining interfaces and feedback loops.

We then applied this approach to the field of communication science and discussed its impact on the design of the whole data-driven research process. Our experiences with the domain-specificity component, which was improved based on multiple feedback loops, demonstrate that iterative feedback and refinements—for example, of system components—are crucial for ensuring success. The presented process also guides the evolution of the system in a controlled and structured way, enabling computer scientists and domain experts to systematically monitor the impact of each feedback loop and to determine when a sufficient quality has been reached.

A case study that drew upon the developed system demonstrated how a close alignment between the target discipline and computer scientists helps in creating research designs that yield insights of high relevance to the target domain. The Swiss Media Criticism portal, for example, provided sophisticated analytics that helped communication

scientists in identifying, tracking, and understanding public media criticism debates. An analysis that would take weeks with conventional means can be performed in close to real time, enabling the research team to provide continuous reports on media-critical events and to investigate temporal effects. The alignment of data-driven research methods with the research framework from communication science also ensures that the obtained results adhere to the target domain's scientific standards and yield relevant contributions to this field. Computer scientists also benefit from this process, since the interdisciplinary approach provides them with insights into the target domain's research frameworks and with innovative views on the strengths and weaknesses of their technology, since each scientific discipline poses its own typical questions for its objects of investigation, and technology cannot answer all questions ad hoc.

Author Contributions: Conceptualization, A.W., P.K., M.S., U.D. and V.W.; Data curation, P.K., V.F. and M.S.; Funding acquisition, A.W., U.D. and V.W.; Methodology, A.W., U.D. and V.W.; Project administration, U.D. and V.W.; Resources, V.F. and M.S.; Software, A.W. and P.K.; Supervision, A.W., U.D. and V.W.; Visualization, P.K. and V.F.; Writing—original draft, A.W., P.K., V.F., M.S., U.D. and V.W.; Writing—review & editing, A.W., P.K., V.F., M.S., U.D. and V.W. All authors have read and agreed to the published version of the manuscript.

Funding: The work presented in this paper was conducted as part of the “Radar Media Criticism Switzerland” project funded by the Swiss National Science Foundation under the project number: 150327.

Data Availability Statement: Publicly available datasets have been created and utilized within this study. These data can be found at <https://github.com/media-criticism/swiss-dataset> (accessed on 21 February 2021). The gold standard dataset used for the training and evaluation of the domain-specificity component contains documents that are copyrighted by third parties and, therefore, cannot be legally redistributed in Switzerland.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Connolly, R. Why computing belongs within the social sciences. *Commun. ACM* **2020**, *63*, 54–59. [CrossRef]
- Susanto, Y.; Livingstone, A.; Ng, B.C.; Cambria, E. The Hourglass model revisited. *IEEE Intell. Syst.* **2020**, *35*, 96–102. [CrossRef]
- Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
- Bartlett, A.; Lewis, J.; Reyes-Galindo, L.; Stephens, N. The locus of legitimate interpretation in Big Data sciences: Lessons for computational social science from -omic biology and high-energy physics. *Big Data Soc.* **2018**, *5*, 2053951718768831.
- Ranganath, S.; Hu, X.; Tang, J.; Liu, H. Understanding and Identifying Advocates for Political Campaigns on Social Media. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16), San Francisco, CA, USA, 22–25 February 2016; ACM: New York, NY, USA, 2016; pp. 43–52. [CrossRef]
- Scharl, A.; Herring, D.; Rafelsberger, W.; Hubmann-Haidvogel, A.; Kamolov, R.; Fischl, D.; Föls, M.; Weichselbraun, A. Semantic Systems and Visual Tools to Support Environmental Communication. *IEEE Syst. J.* **2017**, *11*, 762–771. [CrossRef]
- Khatua, A.; Cambria, E.; Ho, S.S.; Na, J.C. Deciphering Public Opinion of Nuclear Energy on Twitter. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. ISSN 2161-4407. [CrossRef]
- Cavalcante, R.C.; Brasileiro, R.C.; Souza, V.L.F.; Nobrega, J.P.; Oliveira, A.L.I. Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Syst. Appl.* **2016**, *55*, 194–211. [CrossRef]
- Xing, F.Z.; Cambria, E.; Welsch, R.E. Natural language based financial forecasting: A survey. *Artif. Intell. Rev.* **2018**, *50*, 49–73. [CrossRef]
- Yang, D.; Huang, C.; Wang, M. A social recommender system by combining social network and sentiment similarity: A case study of healthcare. *J. Inf. Sci.* **2017**, *43*, 635–648. [CrossRef]
- Xiao, S.; Wei, C.P.; Dong, M. Crowd intelligence: Analyzing online product reviews for preference measurement. *Inf. Manag.* **2016**, *53*, 169–182. [CrossRef]
- Chung, W.; Zeng, D. Social-media-based public policy informatics: Sentiment and network analyses of U.S. Immigration and border security. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 1588–1606. [CrossRef]
- Scharl, A.; Herring, D.D. Extracting Knowledge from the Web and Social Media for Progress Monitoring in Public Outreach and Science Communication. In Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia'13), Salvador, Brazil, 5–8 November 2013; ACM: New York, NY, USA, 2013; pp. 121–124. [CrossRef]
- Napoli, P.M. Social TV Engagement Metrics: An Exploratory Comparative Analysis of Competing (Aspiring) Market Information Regimes. *SSRN Electron. J.* **2013**. [CrossRef]

15. Wakamiya, S.; Lee, R.; Sumiya, K. Towards Better TV Viewing Rates: Exploiting Crowd's Media Life Logs over Twitter for TV Rating. In Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication (ICUIMC'11), Seoul, Korea, 21–23 February 2011 ; ACM: New York, NY, USA, 2011; pp. 39:1–39:10. [CrossRef]
16. Scharl, A.; Hubmann-Haidvogel, A.; Jones, A.; Fischl, D.; Kamolov, R.; Weichselbraun, A.; Rafelsberger, W. Analyzing the Public Discourse on Works of Fiction—Automatic Emotion Detection in Online Media Coverage about HBO's Game of Thrones. *Inf. Process. Manag.* **2016**, *52*, 129–138. [CrossRef] [PubMed]
17. Li, Q.; Wang, T.; Li, P.; Liu, L.; Gong, Q.; Chen, Y. The effect of news and public mood on stock movements. *Inf. Sci.* **2014**, *278*, 826–840. [CrossRef]
18. Xing, F.; Malandri, L.; Zhang, Y.; Cambria, E. Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; International Committee on Computational Linguistics: Barcelona, Spain (Online), 2020; pp. 978–987.
19. Kim, E.H.J.; Jeong, Y.K.; Kim, Y.; Kang, K.Y.; Song, M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *J. Inf. Sci.* **2016**, *42*, 763–781. [CrossRef]
20. Hubmann-Haidvogel, A.; Scharl, A.; Weichselbraun, A. Multiple Coordinated Views for Searching and Navigating Web Content Repositories. *Inf. Sci.* **2009**, *179*, 1813–1821. [CrossRef]
21. Malik, M. *Journalismusjournalismus. Funktion, Strukturen und Strategien der journalistischen Berichterstattung*; Springer: Berlin/Heidelberg, Germany, 2004.
22. Wyss, V.; Keel, G. Media Governance and Media Quality Management: Theoretical Concepts and an Empirical Example from Switzerland. In *Press Freedom and Pluralism in Europe: Concepts and Conditions*; Intellect: Bristol, TN, USA; Chicago, IL, USA, 2009; pp. 115–128.
23. Sutter, T. *Medienanalyse und Medienkritik. Forschungsfelder einer Konstruktivistischen Soziologie der Medien*; VS Verlag: Wiesbaden, Germany, 2010.
24. Schmidt, S.J. Zur Grundlegung einer Medienkritik. In *Neue Kritik der Medienkritik. Werkanalyse, Nutzerservice, Sales Promotion oder Kulturkritik*; Herbert von Halem Verlag: Köln, Germany, 2005; pp. 21–40.
25. Scodari, C.; Thorpe, J. *Media Criticism. Journeys in Interpretation*; Kendall Hunt Publishing: Dubuque, IA, USA, 1993.
26. Meier, C.; Weichert, S. *Basiswissen für die Medienpraxis. Journalismus Bibliothek 8*; 2012. Available online: https://www.halem-verlag.de/wp-content/uploads/2012/09/9783869620237_inhalt.pdf (accessed on 21 February 2021).
27. Wyss, V.; Schanne, M.; Stoffel, A. Medienkritik in der Schweiz—Eine Bestandsaufnahme. In *Qualität der Medien. Schweiz-Suisse-Svizzera. Jahrbuch 2012*; Schwabe: Basel, Switzerland, 2012; pp. 361–376.
28. Puppis, M.; Schönhagen, P.; Fürst, S.; Hofstetter, B.; Meissner, M. Arbeitsbedingungen und Berichterstattungsfreiheit in Journalistischen Organisationen. Available online: <https://www.bakom.admin.ch/dam/bakom/de/dokumente/2014/12/journalistenbefragungimpressum.pdf.download.pdf/journalistenbefragungimpressum.pdf> (accessed on 21 February 2021).
29. Eberwein, T. *Raus aus der Selbstbeobachtungsfalle! Zum medienkritischen Potenzial der Blogosphäre*; Springer: Berlin/Heidelberg, Germany, 2008.
30. Eberwein, T. *Typen und Funktionen von Medienblogs*; Springer: Berlin/Heidelberg, Germany, 2008.
31. Eberwein, T. Von "Holzhausen" nach "Blogville"—Und zurück. Medienbeobachtung in Tagespresse und Weblogs. In *Journalismus und Öffentlichkeit. Eine Profession und ihr gesellschaftlicher Auftrag*; Festschrift für Horst Pöttker; VS Verlag: Wiesbaden, Germany, 2010; pp. 143–165.
32. Kleiner, M.S. *Einleitung; Grundlagentexte zur sozialwissenschaftlichen Medienkritik*; VS Verlag: Wiesbaden, Germany, 2010; pp. 13–85.
33. Russ-Mohl, S.; Fengler, S. *Medien auf der Bühne der Medien. Zur Zukunft von Medienjournalismus und Medien-PR*; Dahlem University Press: Berlin, Germany, 2000.
34. Maali, F.; Cyganiak, R.; Peristeras, V. A Publishing Pipeline for Linked Government Data. In Proceedings of the 9th Extended Semantic Web Conference, Heraklion, Greece, 27–31 May 2012.
35. Lang, H.P.; Wohlgenannt, G.; Weichselbraun, A. TextSweeper—A System for Content Extraction and Overview Page Detection. In Proceedings of the International Conference on Information Resources Management (Conf-IRM), Vienna, Austria, 21–23 May 2012.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
37. Weichselbraun, A.; Steixner, J.; Brasoveanu, A.M.P.; Scharl, A.; Göbel, M.; Nixon, L.J.B. Automatic Expansion of Domain-Specific Affective Models for Web Intelligence Applications. *Cogn. Comput.* **2021**, doi:10.1007/s12559-021-09839-4.
38. Weichselbraun, A.; Scharl, A.; Gindl, S. Extracting Opini Targets from Environmental Web Coverage and Social Media Streams. In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS-49), Koloa, HI, USA, 5–8 January 2016; IEEE Computer Society Press: Los Alamitos, CA, USA, 2016.
39. Weichselbraun, A.; Streiff, D.; Scharl, A. Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *Int. J. Artif. Intell. Tools* **2015**, *24*, 1540008. [CrossRef]
40. Scharl, A.; Weichselbraun, A.; Göbel, M.; Rafelsberger, W.; Kamolov, R. Scalable Knowledge Extraction and Visualization for Web Intelligence. In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS-49), Koloa, HI, USA, 5–8 January 2016; IEEE Computer Society Press: Los Alamitos, CA, USA, 2016.

41. Wyss, V. Journalismus als duale Struktur. Grundlagen einer strukturationstheoretischen Journalismustheorie. In *Theorien des Journalismus*; Ein diskursives Handbuch; VS Verlag: Wiesbaden, Germany, 2004; pp. 305–320.
42. Odoni, F.; Kuntschik, P.; Brasoveanu, A.M.; Rizzo, G.; Weichselbraun, A. On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance. In Proceedings of the 14th International Conference on Semantic Systems (SEMANTICS 2018), Vienna, Austria, 10–13 September 2018; Elsevier: Vienna, Austria, 2018.

Article

Dashboard COMPRIME_COMPRI_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal

Nuno Marques da Costa , Nelson Mileu * and André Alves

Institute of Geography and Spatial Planning, Universidade de Lisboa, 1600-276 Lisbon, Portugal; nunocosta@campus.ul.pt (N.M.d.C.); andrejoelalves@campus.ul.pt (A.A.)

* Correspondence: nmileu@campus.ul.pt; Tel.: +351-21-044-30-00

Abstract: Due to its novelty, the recent pandemic of the coronavirus disease (COVID-19), which is associated with the spread of the new severe acute respiratory syndrome coronavirus (SARS-CoV-2), triggered the public's interest in accessing information, demonstrating the importance of obtaining and analyzing credible and updated information from an epidemiological surveillance context. For this purpose, health authorities, international organizations, and university institutions have published online various graphic and cartographic representations of the evolution of the pandemic with daily updates that allow the almost real-time monitoring of the evolutionary behavior of the spread, lethality, and territorial distribution of the disease. The purpose of this article is to describe the technical solution and the main results associated with the publication of the COMPRIME_COMPRI_MOv dashboard for the dissemination of information and multi-scale knowledge of COVID-19. Under two rapidly implementing research projects for innovative solutions to respond to the COVID-19 pandemic, promoted in Portugal by the FCT (Foundation for Science and Technology), a website was created. That website brings together a diverse set of variables and indicators in a dynamic and interactive way that reflects the evolutionary behavior of the pandemic from a multi-scale perspective, in Portugal, constituting itself as a system for monitoring the evolution of the pandemic. In the current situation, this type of exploratory solutions proves to be crucial to guarantee everyone's access to information while simultaneously emerging as an epidemiological surveillance tool that is capable of assisting decision-making by public authorities with competence in defining control policies and fight the spread of the new coronavirus.

Citation: Marques da Costa, N.; Mileu, N.; Alves, A. Dashboard COMPRIME_COMPRI_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal. *Future Internet* **2021**, *13*, 45. <https://doi.org/10.3390/fi13020045>

Academic Editors: Carlos Filipe Da Silva Portela and Andrew Crooks

Received: 17 January 2021

Accepted: 9 February 2021

Published: 12 February 2021

Keywords: dashboard; WebGIS; data analytics; COVID-19; SARS-CoV-2

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In December 2019, several cases of pneumonia of unknown origin appeared in China. Earlier that year, the World Health Organization (WHO) had published a list of ten major global threats, including pandemics associated with respiratory diseases [1]. Later, in January 2020, a new severe acute respiratory syndrome coronavirus (SARS-CoV-2) was identified and recognized as responsible for the spread of the COVID-19 disease. The first records of infection by SARS-CoV-2, initially circumscribed in the Chinese territory, quickly spread, with a progressive increase in the number of infected by the new coronavirus and its spread to dozens of countries.

Due to the rapid progression of the disease, the increase in the number of deaths associated with it, and its widespread expansion, the WHO declared COVID-19 as a pandemic on 11 March 2020. By this date, more than 100,000 people were infected in 114 countries, and the number of deaths was already 4000 [2]. Nine months later, the number of confirmed cases in the world has exceeded 60 million, and the number of deaths is over 1.5 million [3], attesting to the high transmissibility of the disease.

Already in a pandemic situation, the WHO requested several countries with cases of infection to maintain an open policy of communication and dissemination of information about the situation in their countries. This request resulted from the fact that during the last SARS epidemic (severe acute respiratory syndrome), an emerging disease caused from infection by the SARS coronavirus, in 2003, the lack of transparent and updated disclosure of the number of cases made it difficult to control and manage the epidemic, thus evidencing that “transparency is the best policy” [4]. Following this request, the WHO itself set an example by providing a dashboard that gathers a set of data on the evolution of the pandemic situation in the world [3].

In this framework, health authorities in the various affected countries have communicated the evolution of the pandemic situation by making data available to the public. At this juncture, also the public interest for access to information has been high, as shown by the study REACT COVID—Survey on Food and Physical Activity in the Context of Social Contention [5], where it was found that almost 80% of respondents in Portugal seek information on health care and more than 94% have accessed information on COVID-19. However, about 56% state that they have difficulty understanding it. This last finding is an indication of the need to develop simpler representations capable of transmitting information on the pandemic situation to the population with different levels of health literacy. The use of exploratory data methods, such as dashboards applied to monitoring the coronavirus pandemic, has been developed by several authors, for their versatility, speed of analysis, and ease of understanding by decision makers and the general public. For Boulos and Geraghty, dashboards became an essential source of information during the COVID-19 outbreak, contributing to the protection and reduction of its harmful effects [6].

The main objective of this article is to present the dashboard COMPRIME_COMPRI_MOv technical solution and data analytics results for monitoring the evolution of the COVID-19 disease situation in Portugal. The dashboard was published online in order to be a means of dissemination of multi-scale information and knowledge but also a support for analysis of the main time trends and territorial patterns of SARS-CoV-2 contagion in Portugal, and it allows access to multiple indicators and interactive exploration of graphics and dynamic maps. The research path was based on the information collected from the project's stakeholders, namely the dashboard purpose, dashboard users, functional functionalities, design features, and how it could support the decision process.

The development of the dashboard was associated with two research projects, both of them rapid implementation projects for innovative solutions in response to the pandemic of COVID 19 under the exceptional funding line “RESEARCH 4 COVID-19” of the FCT (Portuguese Foundation for Science and Technology) (https://www.fct.pt/apoios/research4covid19/edicao_1/index.phtml.pt) (accessed on 9 February 2021), which aims to support research and development projects and initiatives that meet the needs of the National Health Service:

1. COMPRIME—Conhecer Mais PaRa Intervir MELhor (Get to Know More for Intervention)—has as its main objective to identify the propagation dynamics of SARS-CoV-2, in its relations with the demographic and socioeconomic profiles of the territories, at the municipality scale, identifying the determining factors of this propagation;
2. COMPRI_MOv—Conhecer Mais PaRa Intervir melhor no contexto da Mobilidade (Get to Know More for Intervention in the context of mobility)—aims to characterize the mobility of populations given the intensity, motivation, and geographical pattern of the flows and, associating these dynamics with epidemiological data, assess the risk of propagation associated with mobility. The project intends to propose a monitoring system to support the decision and present the basis of a model for the simulation of propagation based on mobility.

Thus, within these projects, we developed the dashboard, which is structured in four components corresponding to the scales of analysis: international, national, regional, and municipal. The dashboard is published online at the following address: <https://www.comprime-compri-mov.com/dashboards.html> (accessed on 9 February 2021),

with an English version at the following address: <https://www.comprime-compri-mov.com/dashboardsenglish.html> (accessed on 9 February 2021). The information used in the dashboard is merely an evolutionary follow-up based on official information from the Directorate General of Health [7] in the form of daily epidemiological reports, based on which several metrics were calculated. Thus, the COMPRIE_COMPRI_MOv dashboard does not include any output or conclusions obtained in the scope of the projects or information of predictive character, being its function to represent the evolution of the epidemiological situation based on official information.

This article is organized in six parts. The first corresponds to this introduction, which is followed by a brief contextualization of the importance and contributions of using dashboards for the study of COVID-19. The third part concerns the methodologies used, dividing itself between the background, architecture, and the data. The following shows the technical solution designed and briefly describes the main processes and trends of propagation of COVID-19 in Portugal, making it possible to withdraw from the dashboard. In the fifth part, the results achieved with the dashboard are discussed, and in the last part, the conclusions are presented.

2. Use of Dashboards in the Context of the Pandemic

The use of spatial analysis in tracking and understanding the spread of infectious diseases using cartography is an old process. One of the most widespread examples of analysis in the literature is the map of physician John Snow, who in the nineteenth century identified the origin of an outbreak of cholera through the spatial relationships between the occurrences of deaths by disease and the location of water wells in the city of London [8]. With the development of Geographic Information Systems (GIS), the possibilities of analyzing, visualizing, and detecting disease patterns have increased significantly, as proved by the growing number of publications [9]. Using examples at different scales, Zhou et al. analyzed the contribution of GIS and big data in combating the pandemic, with the visualization of information constituting one of the challenges in the response to the pandemic [10]. In addition, Franch-Pardo et al. present several studies using spatial analysis and other GIS techniques to study the geographical dimension of COVID-19, some of which the authors add in the category of web-based mapping as they are cartographic representations of the pandemic situation available online [11]. In the context of a pandemic, the forms of representation of the disease situation have multiplied, with dashboards being the predominant solution. As Sarfo and Karuppannan point out, making data available on interactive dashboards in real time, or near real time, has become a useful tool by which many countries present specific information on COVID-19 [12]. In this sense, Boulos and Geraghty present several examples of WebGIS and dashboards to track the spread of the new coronavirus with the aim of discussing additional ways these tools contribute to combating outbreaks of infectious diseases and epidemics [6]. This type of solution is crucial by facilitating transparent access to information to the entire population in a simple way, ensuring the monitoring of the situation evolution from an epidemiological surveillance point of view, not only regarding the number of new infections but also other parameters and at various scales, also allowing the health authorities a rigorous monitoring. WHO and Johns Hopkins University were pioneers in representing the pandemic situation in various graphic and cartographic ways, with the Center for Systems Science and Engineering (CSSE) dashboard at Johns Hopkins University [13] being used as an official source of international information. Other examples could be given, and there are even published studies by researchers such as Fernandez-Lozano and Cedron [14] that have developed an interactive and dynamic dashboard for monitoring COVID-19 to support the epidemiological study of the disease in Spain, or Barone et al. [15], who propose a set of ways to explore and analyze epidemiological data of COVID-19 from a spatio-temporal perspective with explorative and non-inferential metrics. Without neglecting the importance that these solutions have for epidemiological monitoring and surveillance, it is crucial to address others that add a predictive component to the evolution and that naturally have a greater

weight in helping make decisions regarding the control of contagions. In this sense, Florez and Singh [16] combine international monitoring of the evolution of the number of cases and deaths associated with COVID-19 with the prediction of the evolution of these indicators using quadratic equations. Despite the simple projection using purely mathematical models, ignoring other variables and geographical differentiations of the phenomenon diffusion with importance for a more rigorous estimation, this is still a positive contribution toward the analysis and evaluation of risk of future infection.

In the European context, the European Centre for Disease Control and Prevention (ECDC) has developed its own solution for monitoring developments [17] in the European Union and the European Economic Area, by countries and regions, while maintaining the international situation monitoring.

In Portugal, the national health authority official online service is the DGS (Direção-Geral da Saúde – Directorate-General of Health) interactive platform [18]. This platform reproduces the official reports published daily [7] at national, regional, and municipal level (weekly), but no other metrics and indicators are available with the official information. The National Institute of Statistics (INE) has also designed a solution of the same kind, with special focus on municipalities and where it makes available, for each analysis unit, statistical variables, and indicators from its database in an attempt to complement the reading of the epidemiological situation [19].

The “COVID-19 Insight” platform [20] is one of the most complete proposals for multi-scale and multi-thematic epidemiological monitoring which, in addition to including a municipal risk index, allows monitoring and forecasting the impacts on the economy and changes in mobility, as well as epidemiological estimation models of various relevant indicators.

Finally, and with the same type of architecture as the DGS and INE solutions, the Portuguese Association of Geographers (APG) partially replicates the most relevant national information by monitoring in greater detail the evolution of cases per municipality, providing the absolute variation in the number of new cases and the accumulated per inhabitant [21].

3. Materials and Methods

3.1. Background

The transformation of large volumes of data into information using dashboards is a practice in several domains. From the typical dashboards associated with management [22], through its adoption in the context of smart cities [23,24] or in health [25], several examples can be found in the literature. Although there are several dashboard definitions [23,24], they all share the fact that it constitutes a communication mechanism to support decision making. In this article, the definition of Yigitbasioglu and Velcu was adopted, where a “dashboard can be regarded as a data-driven decision support system, which provides information in the particular format to the decision maker” [22]. Another important aspect in the construction of a dashboard, is related to the type of use. In this respect, Stephen Few’s Information Dashboard Design [26] establishes a useful taxonomy, proposing three categories: strategic, operational, and analytical. Taking into account the objectives associated with the research project, it was considered that the COMPRIE_COMPRI_MOv dashboard fit into the operational typology [27], providing descriptive measurements using indicators based on original data and other related data, in order to provide multi-temporal and multi-scale information.

The research methodology was adapted from Yigitbasioglu and Velcu [22], organizing in the following conceptual relationships: purpose, users, design features, and decision-making (Figure 1). The implementation of concepts and relationships between entities was based on discussion and analysis among the stakeholders involved in the project. Firstly, the entities involved established the main objective of the dashboard to monitor cases of infection and to calculate indicators using auxiliary information to allow understanding/explaining the spread of the virus in the territory. The first step was the assembly of

appropriate time series data and geographic information. Based on the number of cases and auxiliary demographic information, the indicators that make up the dashboard are calculated. The calculation of indicators (e.g., confirmed cases per 10,000 inhabitants by municipality, risk classification) allows establishing the relationship between the pandemic situation and the socio-economic context.

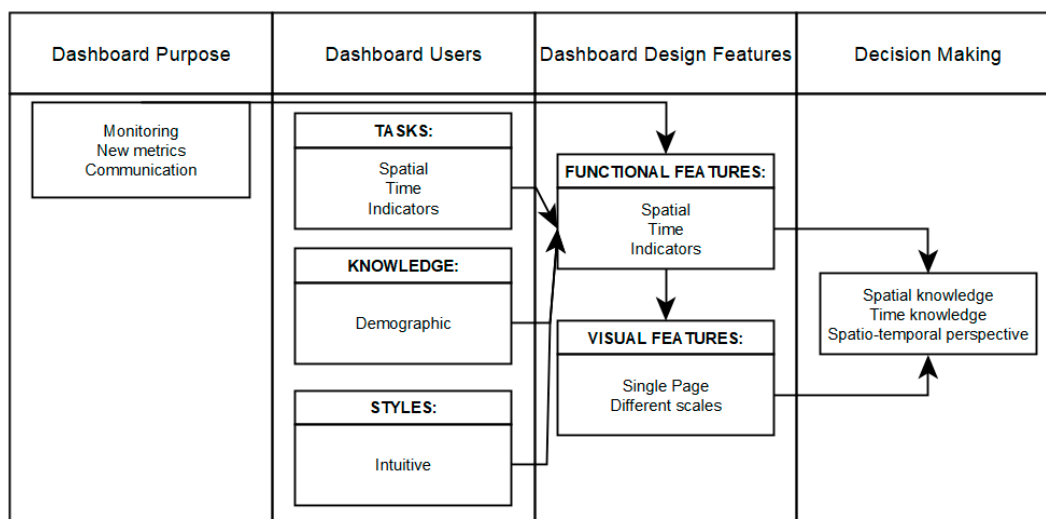


Figure 1. Dashboard research paths.

The functional features identified by the project's stakeholders were the visualization of information through maps, the use of simple and dynamic graphics, and the visualization of the big statistical numbers associated with the number of cases. The principles of visualizing data identified by the project stakeholders were the visualization on a one-page style and drilldown style containing all the maps and graphs (Figure 2). The screen is organized by geographic scales, and each level includes elements such as maps, graphs, and indicators. Based on the requirements defined by the stakeholders and the purpose of the dashboard, the integration of the map and other elements assumed the map as part of the Graphic User Interface (GUI) [23]. The interactive spatio-temporal exploration, and spatio-temporal information decoding constitutes the support for decision-making.

In order to provide an interactive exploration of spatiotemporal data, the time sliders were developed to make a web maps depicting time series information using proportional symbols [28] and choropleth maps.

3.2. Architecture

To integrate data and the logic from various systems and derive new functionalities, such as multi-temporal and multi-scale management, a combined approach based on Esri's ArcGIS Online technology was chosen (Figure 3). All existing graphical elements, except external incorporations and cartography, were developed from ArcGIS Online dashboard functionalities.

WebGIS cartography was developed using the Leaflet map library and jQuery function library, in a solution that combines HTML (HyperText Markup Language), CSS (Cascading Style Sheets) and JS (JavaScript), for allowing greater development freedom, interactivity and query speed. The cartography was developed combining the Leaflet map library and some of its plugins such as Leaflet.migrationLayer (for viewing imported case flows) with the jQuery library. Both the proportional circle maps and the choropleths present in the dashboard allow the user to set the date of the data series to be represented by a time slider from jQuery.

Due to the external cartographic development, the cartography was published online on a hosting platform and then incorporated into the dashboard.

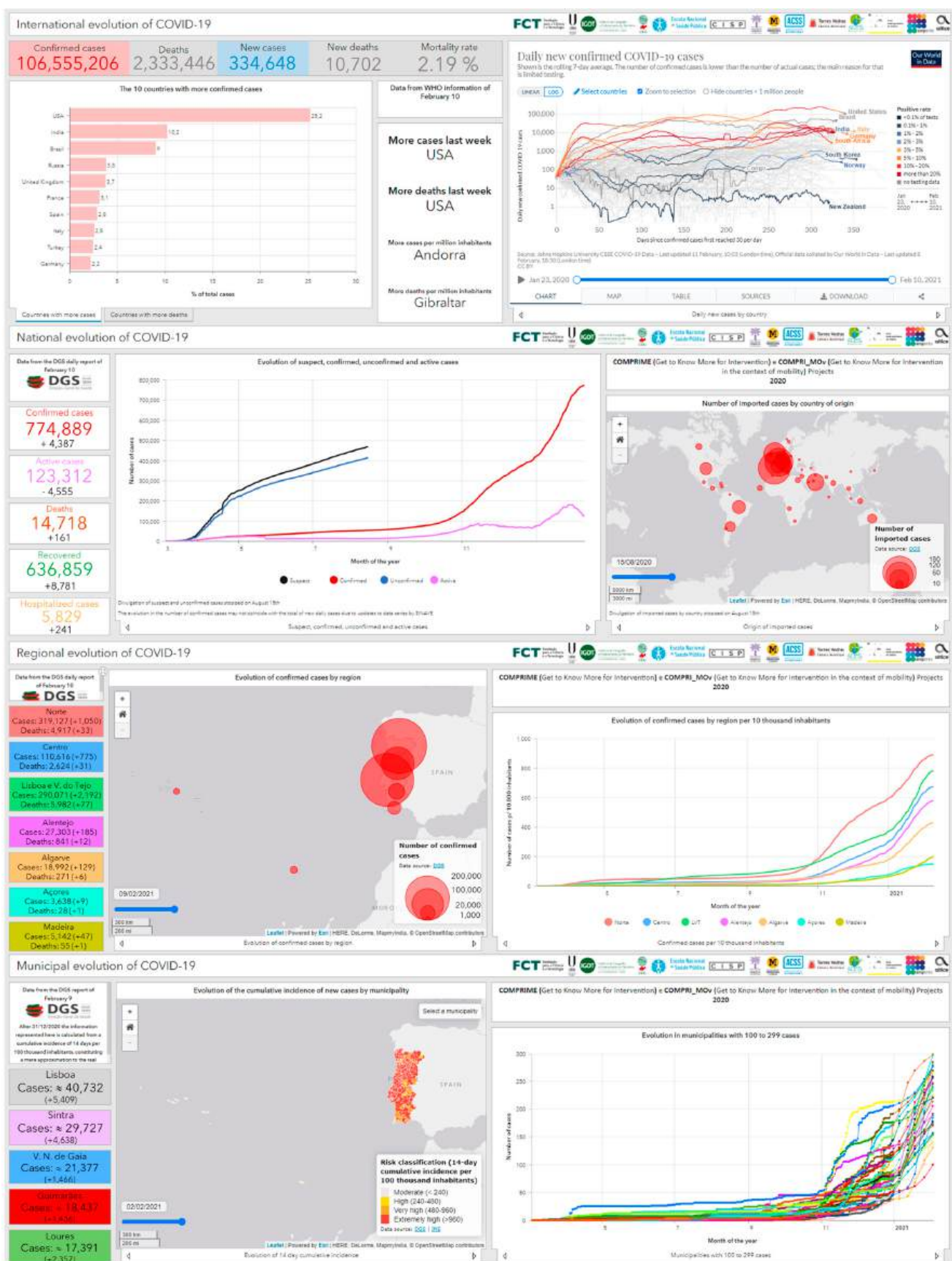


Figure 2. Dashboard design.

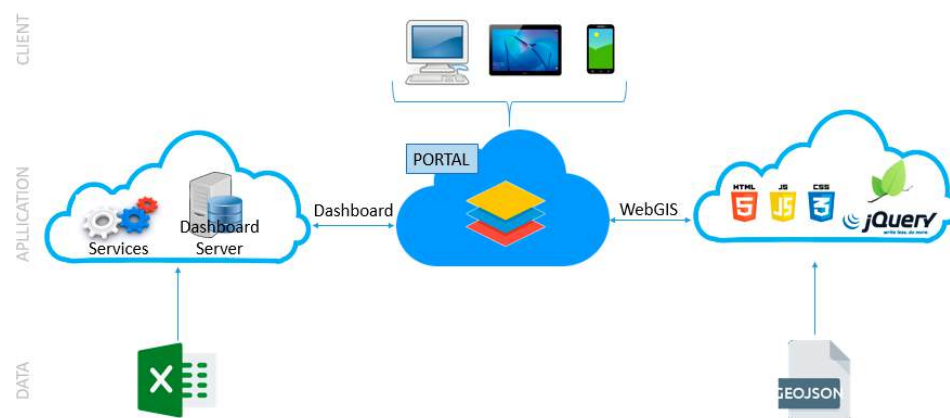


Figure 3. Dashboard architecture.

3.3. Data

The data for the national territory come from the daily reports [7] and dashboard of DGS [18]. Despite the work developed by DGS for the daily preparation of epidemiological newsletters and their availability to the public, the information is made available exclusively in Portable Document Format and dashboard, i.e., it is not provided in a format that could be easily manipulated and submitted to analysis and modeling tasks or can be integrated into a GIS, which is a limitation to the use of the data. Therefore, the use of the information for analysis tasks requires the manual collection of the information, subject of course to errors. It was also necessary to edit and correct the series when they represent reductions in the number of accumulated cases due to recounts. This edition was imperative for a proper representation of the graphic and cartographic information. Various changes that have occurred in the information made available in the daily reports should also be noted. That information no longer includes imported cases, which limits the knowledge of the geographical origin of new cases associated with international contacts. It should also be noted that historical corrections in the series made by DGS sometimes occur due to delays in reporting by laboratories or due to incorrect allocation of cases to territorial units, but the correction process is not made available by DGS. The main consequence is that the accumulated numbers of some variables do not match with the sum of daily values. Whenever this happens, data are updated as much as possible from the available information. In addition, the information on a municipal scale is no longer disclosed with a fixed periodicity, and the disclosure of accumulated cases per municipality has been discontinued. Now, it provides at 14 days the cumulative incidence per 100,000 inhabitants. These changes also constitute constraints to the maintenance and update of the information present in the dashboards.

The data on the international scale are based on information from the WHO dashboard [3] and the portal “Coronavirus Pandemic (COVID-19)”, which compiles a set of data from the ECDC [29]. Due to the use of these two sources for monitoring the situation at an international level, it is sometimes possible that there is not a complete matching of the quantities recorded by both.

The information regarding dashboard charts and graphs has been structured in Excel files, due to its ease of import and daily update in ArcGIS Online. Regarding WebGIS cartography, the information is structured in GeoJSON format that contains all the associated data series, topology, and coordinates.

4. Results

This dashboard ensures the monitoring of the evolution of COVID-19 at various scales, in a close articulation between the geographic and temporal dimensions. The interactivity and dynamism of the various graphic elements (Figure 4a, b) is a guarantee of ease of consultation and access to all available information. The production of dynamic WebGIS cartography with time control ensures the representation of variables in a spatio-

temporal perspective (Figure 4c–f), proving to be of great contribution to the interpretation of processes and trends of evolution of the phenomenon and its propagation in space and over time.

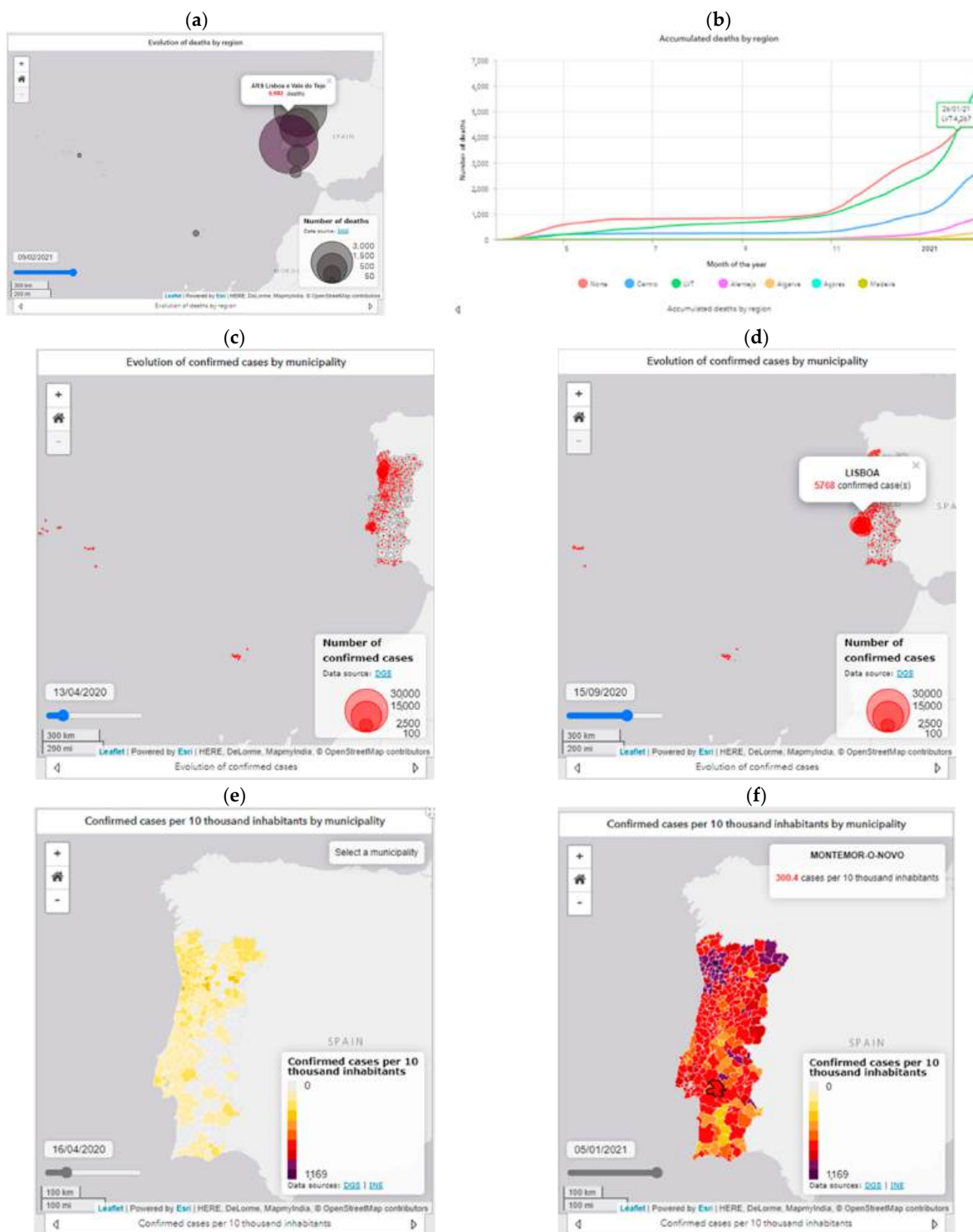


Figure 4. Interactivity of the dashboard elements. Source: <https://www.comprime-compri-mov.com> (accessed on 9 February 2021).

Since the publication of the dashboard in October 2020 until the month of November 2020, there were about 600 visits and 1191 page views, corresponding to a daily average of 10 visits.

The dashboard is structured in four components, each representing a different scale of analysis (international, national, regional, and municipal).

4.1. International Scale

In monitoring the situation on an international scale (Figure 5), the dashboard is divided into two parts:

1. The left half that results from the WHO data collection [3] in which the main figures (confirmed cases, deaths, new cases, new deaths, and mortality rate) and the countries that register a rapid increase in them in absolute terms and by their population are highlighted. The proportions of cases and deaths in the world context are also represented for the 10 most affected countries (Figure 5a,c).
2. The right half where six external elements are incorporated: daily variation of new confirmed cases per country (Figure 5b), new cases per million inhabitants, new deaths per million inhabitants (Figure 5d), evolution of the total number of cases and deaths in the world, evolution of vaccination doses administered and, finally, the WHO dashboard.

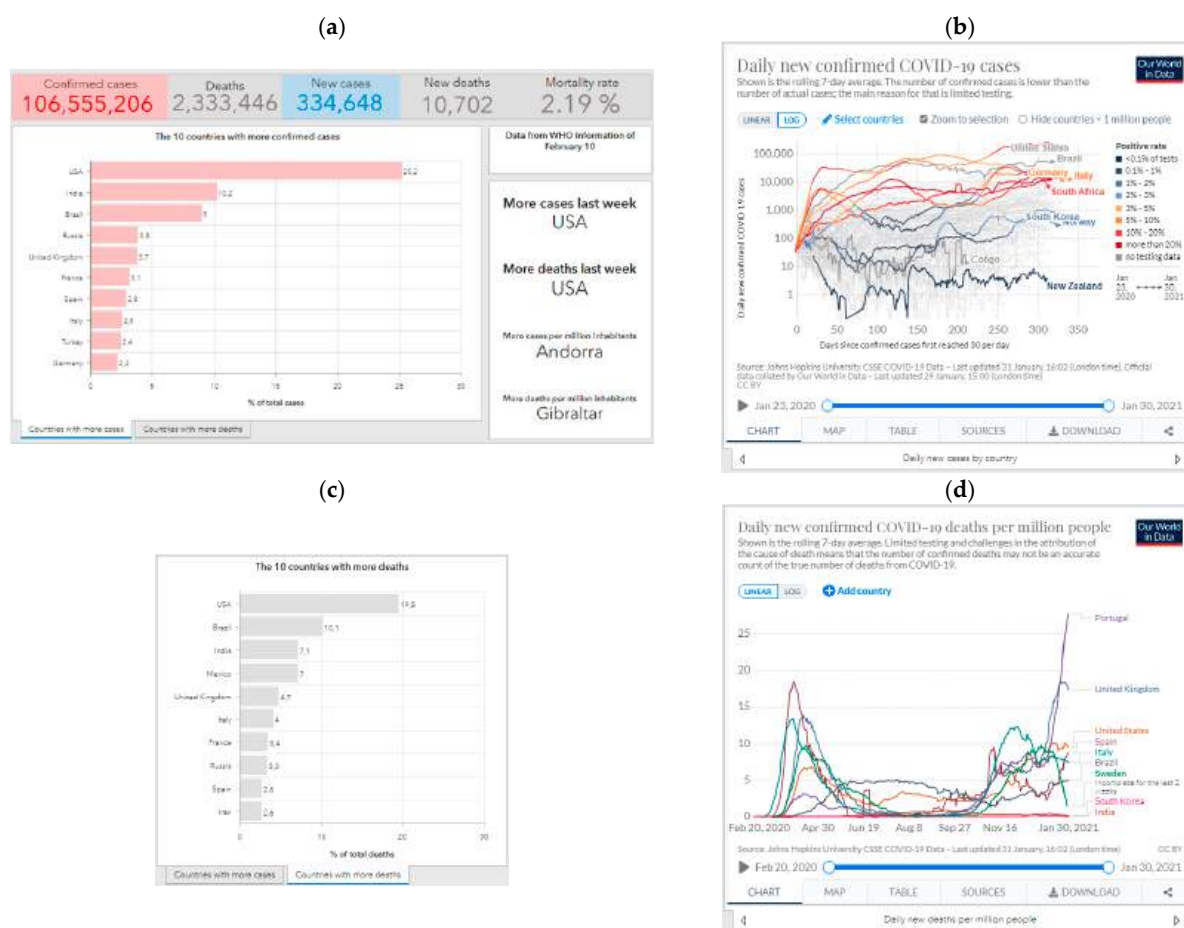


Figure 5. Example of dashboard elements for the international scale. Source: <https://www.comprime-compri-mov.com> (accessed on 9 February 2021).

In the context of the international component, the option of incorporating the external elements mentioned, as opposed to their collection and subsequent editing and representation, results from the slowness of constant updating of the situation for all countries with incidence of the disease. Moreover, the incorporation of these elements allows the user to access a vast amount of information published in these sources, contributing to the access to more information on the phenomenon.

4.2. National Level

As for the national context, particular importance is given to the large numbers (confirmed cases, active, deaths, recovered, and hospitalized) using 11 line graphs to represent the evolution of the main variables made available in the DGS daily bulletins. Metrics calculated from the information in the DGS reports were also included, such as the average of new infections in the last 14 days and the confirmed and active cases per 10,000 inhabitants. It also includes two interactive maps (one of flows and one of proportional circles), representing the origin of imported cases (with temporal variation), which is accompanied by two horizontal bar graphs expressing the number and relative percentage of the 10 countries of origin of the most cases imported into Portugal, and also two vertical bar graphs stacked at 100% with the distribution of confirmed cases and deaths by age.

The metrics and indicators available at the national scale (Figure 6) are indicative of the evolution of the main variables, such as the number of confirmed, active, recovered, and death cases.

The evolution of confirmed cases (Figure 6a) depends on the number of new cases per day (Figure 6c) which, in turn, depends on the testing that is performed. In the first months of known spread of the new coronavirus in Portugal (March and April), high values of infection were reached with the number of accumulated cases exceeding 20,000. Later, and in the context of general confinement (March to May), a slowdown is observed, but the numbers increase again in the beginning of September, with the return to schools and face-to-face work after vacations, reaching in the months of October and November values of contagion much higher than those recorded at the beginning of the pandemic. The average of new 14-day infections in November was seven times higher than the maximum registered in the first wave (5587 against 800). The variation from September shows that the first wave (March, April, and May) was very circumscribed in time and of low magnitude, with consecutive maximums of new infections in October and November, with the number of confirmed cases doubled in less than 1 month (101,860 cases at 18 October 2020 and 204,664 at 12 November 2020, reaching 300,462 at 30 November 2020). In the end of December, this second wave showed signals of being under control; however, after Christmas and New Year's Eve festivities, the number of new cases increased as never seen before, being 15 times higher than the first wave and more the double the second. January was the worst month, representing about 30% of the accumulated cases since the beginning of the pandemic in Portugal. The evolution of the disease's transmissibility index (Figure 6e) reveals a complex trajectory that does not always coincide with the evolution of new cases, but at the time of higher incidence of new cases, it was higher than 1. The number of recovered patients has always evolved favorably with an exponential increase in recent months, as has the number of cases. The number of hospitalizations has also been increasing, as have intensive care unit admissions, putting great pressure on health care and revealing the severity and magnitude of this new phase. The fatality of the disease as a proportion of those infected was especially high in the first phase, and it decreased during and after the period of national confinement in a clear trend of control and maturation. However, the current number of daily deaths is much higher than in the first and second waves, successively reaching record mortality rates. January was also the worst month relative to COVID-19 deaths with almost 6000 deaths. It took nine months of the pandemic to reach the same number. Fatal cases occur mainly in the older age groups, and there are no substantial differences in proportions between males and females. The same is not true for confirmed cases in which, in a very short initial period, men were the most affected, and since the end of March, women have been the most affected. As far as testing is concerned, the daily variation varies substantially, yet the general trend is one of continuous increase, reaching the maximum number of tests ever performed. With the positive rate increasing, it reveals that the increase in the number of cases in recent months is not exclusively the result of more testing. On this scale, a map is also available with the number of imported cases per country of origin (Figure 6b), a map representing the flows per country of origin (Figure 6d), and a graph with the top 10 countries of origin (Figure 6f).

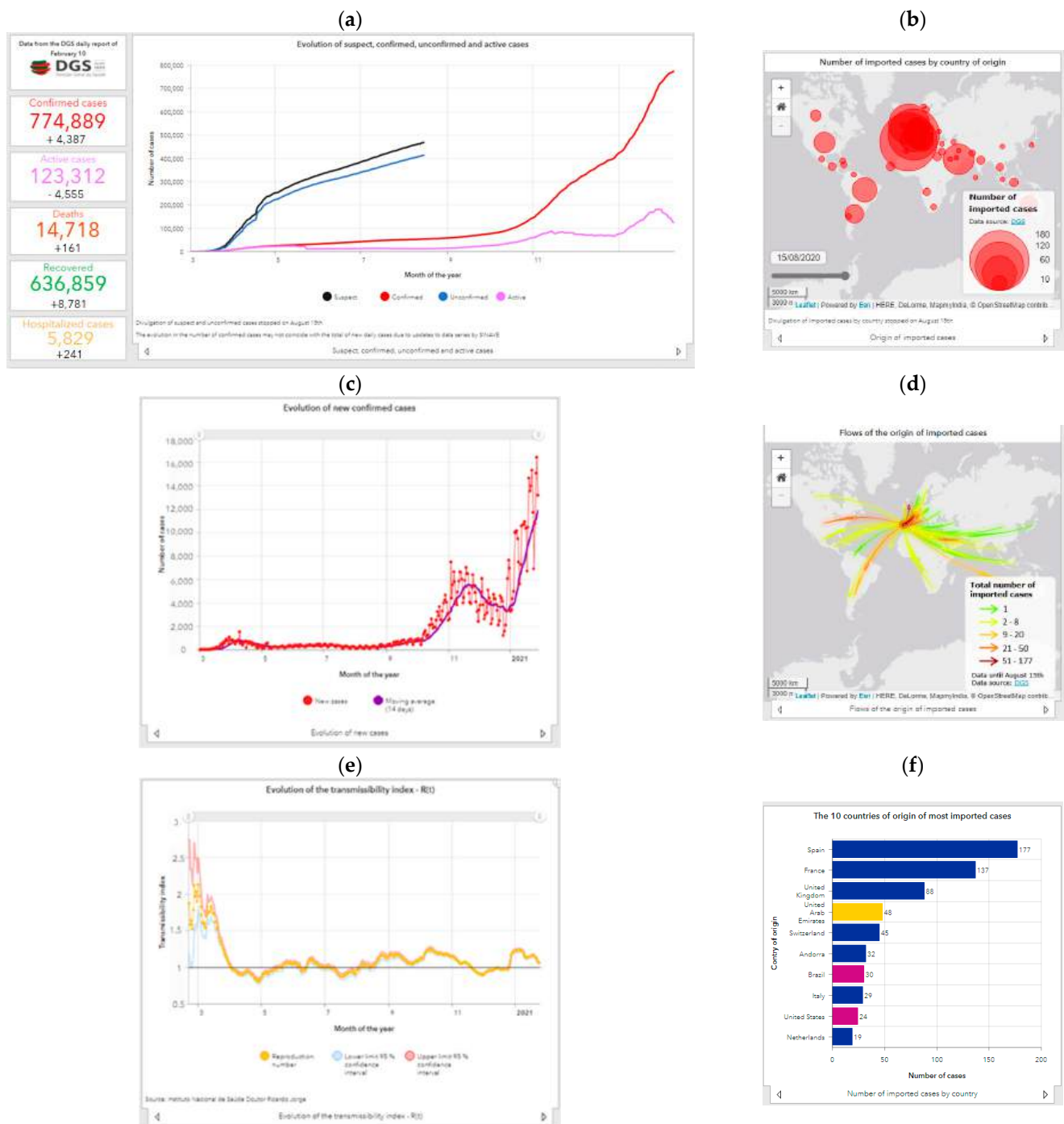


Figure 6. Example of dashboard elements for the national scale. Source: <https://www.comprime-compri-mov.com> (accessed on 9 February 2021).

4.3. Regional Level

At the regional scale, the cases and deaths are differentiated by a region on the interactive maps, which allows changing the date represented (Figure 7a,c,e). Seventeen graphs were elaborated to show the evolutionary behavior of the virus propagation at the regional scale, highlighting the accumulated curves individualized by Regional Health Administration (RHA), the number of new cases per region, as well as the average evolution (Figure 7b,d,f). In addition to the number of cases, graphs by the RHA referring to deaths are presented with the representation of their accumulated number (information also mapped) and the mortality rate.

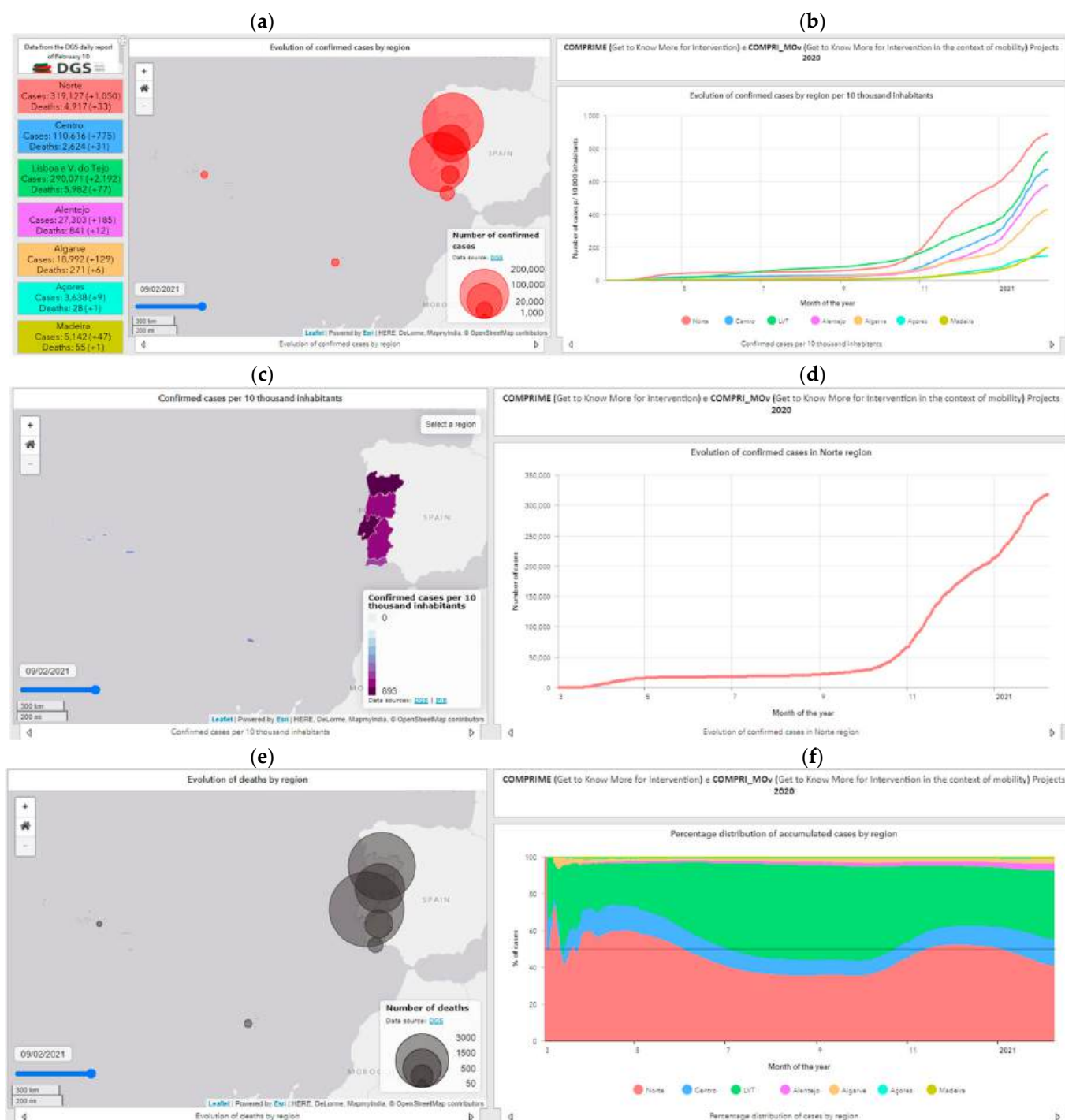


Figure 7. Example of dashboard elements for the regional scale. Source: <https://www.comprime-compri-mov.com>. (accessed on 9 February 2021).

The spatial spread of the disease started in the Norte RHA and Lisboa e Vale do Tejo (LVT) RHA, since these administrative units include the metropolitan areas of Lisbon and Porto. The phenomenon quickly spread to the Centro RHA, and only at a later stage was there a higher incidence in Alentejo and Algarve. Subsequently, the LVT overcame the Norte region in daily and accumulated cases, becoming the main focus of contagion in the national territory, while the remaining regions showed a situation of control and low contagion. In a more recent period (October), the Norte region again reached daily incidence values higher than any other unit in the country, in a clear repetition of the initial trend of the pandemic, currently registering more confirmed cases than all the other RHAs

together. With the beginning of the third wave, the LVT RHA far surpassed the incidence registered in Norte. The index of transmissibility of the disease presents several “peaks” due to the occurrence of large outbreaks and localized outbreaks of infection by COVID 19, especially in the Alentejo RHA.

The Madeira and Azores RHA are characterized by its own evolution of greater stability, which is not strange its insularity, with the number of new contagions very residual and practically zero in the period of confinement. From July onwards, there was a more expressive increase, although it was quite controlled, maintaining these two regions as the two with the lowest number of cases per 10,000 inhabitants. After January, they seem to show the same trajectory of the rest of country trajectory with record incidences.

Regarding deaths, the mortality rate has been decreasing considerably in all RHA, remaining below 5% in all territorial units, with the Norte LVT accumulating the highest number of fatal cases, closely followed by the Norte RHA.

4.4. Municipal Level

Given the municipal context, the dashboard (Figure 8) highlights the municipalities with the highest number of confirmed cases. In parallel, the municipalities are represented in a WebGIS through six variables with the option of temporal change: risk classification based on the cumulative incidence of new cases at 14 days per 100,000 inhabitants (Figure 8a); number of confirmed cases (Figure 8c); number of cases per classes (Figure 8e); confirmed cases per 10,000 inhabitants (Figure 8g); infection rate per 10,000 inhabitants (Figure 8i).

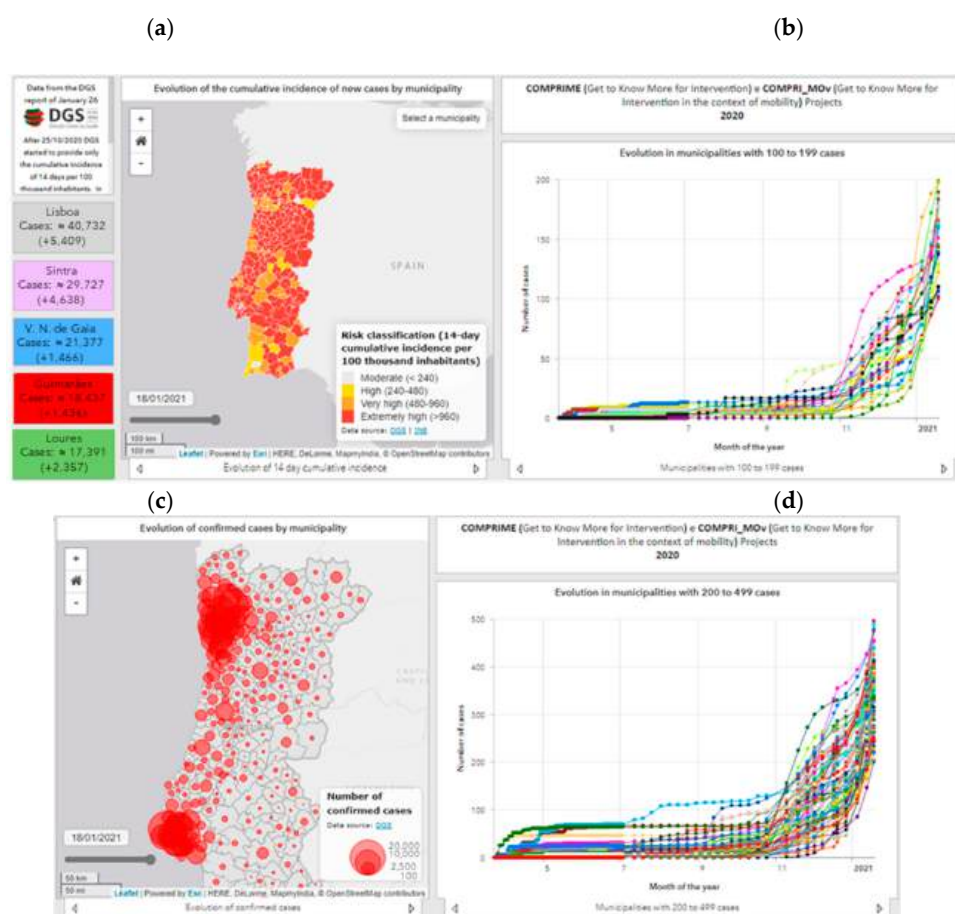


Figure 8. Cont.

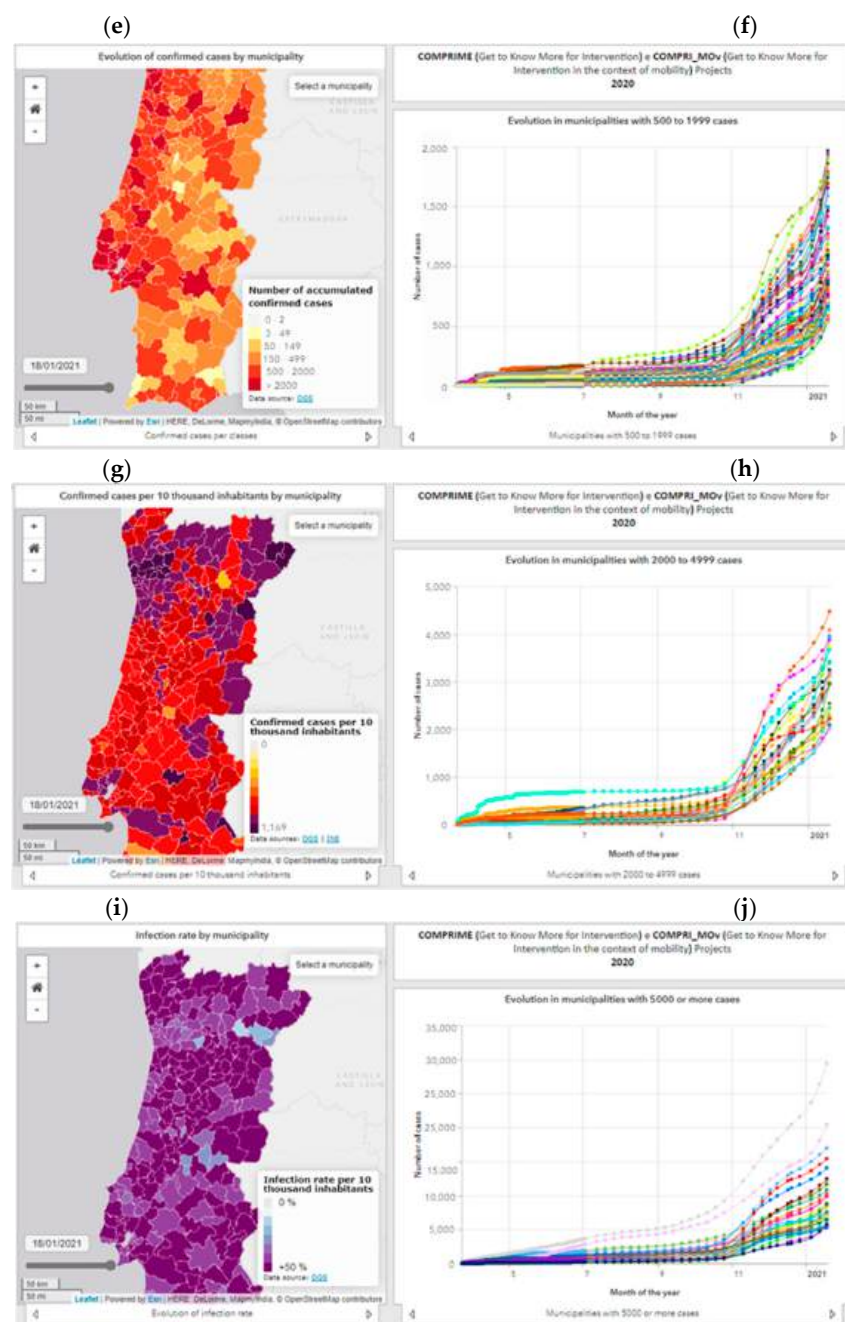


Figure 8. Example of dashboard elements for the municipal scale. Source: <https://www.comprime-compri-mov.com> (accessed on 9 February 2021).

The representation of these variables on a municipal scale, associated with a timeline, allows identifying different geographical patterns, as well as spatio-temporal trends in an interactive way. The cartography is accompanied by five graphs of the evolution of the number of confirmed cases, allowing interpreting the different timings of contagion and the epidemic phase of the municipalities: municipalities with 100 to 199 cases (Figure 8b); municipalities with 200 to 499 cases (Figure 8d); municipalities with 500 to 1999 cases (Figure 8f); (iv) municipalities with 2000 to 4999 cases (Figure 8h); and municipalities with 5000 or more cases (Figure 8j).

5. Discussion

The high number of accesses to the dashboard allows us to state that the use of exploratory data methods is an approach that facilitates the understanding of the main epidemiological patterns and trends.

The cartography available on the dashboard allows a temporal analysis of the evolution of the propagation of new cases, noting that in the first moment (March to early April), the evolution of the number of infections was higher in municipalities with higher population density and associated with concentrations of employment, as is the case of municipalities in metropolitan areas [30]. The spatial spread of the phenomenon in mainland Portugal occurred from metropolitan areas to non-metropolitan coastal municipalities, particularly in the northern and central regions first, followed by the spread from the coast to the interior from the main cities and urban axes of the coast [31].

Later, the phenomenon reaches territories with lower population density and an older population, with outbreaks also occurring, mainly in nursing homes, leading to a significant increase in the number of outbreaks of infection dispersed throughout the municipalities of the interior in the Norte and Centro regions and, later, the municipalities of the Oeste e Médio Tejo [30].

Another component responsible for the dispersed and random number of new cases results from the visit of emigrants to Portugal, which caused some contagions in the regions Norte and Centro [31].

While in the first wave, the phenomenon expanded practically to the entire national territory, both through hierarchical diffusion and by contagion, the second phase started in September, which demonstrates not only the further geographical expansion of the phenomenon, with virtually all municipalities registering cases of infection by COVID-19, but also a substantial increase in the values of confirmed cases in the most urban municipalities and their nearest functional dependencies, proving not only the growth of territorial diffusion but also the consolidation of spatial dispersion of the disease.

Thus, we can state that the diffusion of COVID-19 in Portugal is strongly related to the hierarchy of the urban network, spreading the infection from the main urban centers to those closest to them, and with the concentration of employment and industrial production. Industrial companies with a high number of workers have been the scene of large outbreaks of infection, often with high geographical extent. Social facilities such as nursing homes and senior residences have been associated with large localized outbreaks of contagion in a dispersed and localized pattern and explain the high number of cases in the older territorial units [32].

Dashboards can improve transparency and accountability; however, there are various risks and challenges related [33]. The greatest limitations to the development of the dashboard are related to data access. For example, information on a municipal scale is no longer disclosed at a fixed periodicity, and the disclosure of accumulated cases per municipality has been discontinued. These changes are constraints to the maintenance of the data series of the various indicators, requiring various calculations in order to guarantee the continuity of the series. Moreover, in the municipal case, due to these calculations, only approximate values represent the actual information, with no public access to the real number of cases per municipality. Another aspect refers to the fact that the data are not made available in a format that is easily manipulated (human-readable and machine-readable), that can be submitted to analysis and modeling tasks, or that can be integrated in a GIS, which is a limitation to the data use.

The main benefit for citizenship after implementing the dashboard is the involvement of citizens and the possibility of knowing the consequences of public policies and behaviors that influence the evolution of the pandemic. In this way, citizens will have the opportunity to know the spatio-temporal evolution and discuss the results. Future problems in the use and exploration of the dashboard are the lack of resources, maintenance, and updating, leading citizens to lose trust and to abandon the use of the dashboard.

The evolution of the dashboard is expected to integrate predictive models supported by mobility data, and the limitations associated with real-time data availability will be overcome through the provision of webservices by the national health authority.

6. Conclusions

Geographic information systems in general and web technologies in particular are indispensable tools in the provision and sharing of pandemic information in real time to understand the processes of contagion and support decision-making. The dashboard elaborated and presented in this article allows monitoring “almost in real time” the evolution of the COVID-19 disease at different scales and according to different perspectives, making available in a dynamic and interactive way the main indicators and variables that synthesize the pandemic situation from an exploratory and not inferential point of view.

The possibility of combining the cartographic representation of various indicators with their temporal differentiation allows a clear understanding of the dynamics and processes of virus spread throughout the national territory at regional and municipal scales. This is the main advantage of the solution developed compared to similar ones, and it is possible to consult information regarding several indicators for all the dates when official data are available.

Although the existing knowledge about the spread of the disease is not yet sufficient to control it, the tools for surveillance and epidemiological control, such as dashboards, continue to have an added importance in monitoring infections from a spatio-temporal perspective. The availability of this type of platform, which is an aggregator of official information providing it in an accessible, interactive, and transparent way, contributes to amplifying the dissemination of knowledge, providing important insights on the spread in space and time of the disease, supporting the population and other entities with detailed information so they can make as informed decisions as possible.

Author Contributions: Project administration, N.M.d.C.; funding acquisition, N.M.d.C.; supervision, N.M.d.C.; formal analysis N.M.d.C.; writing—review and editing, N.M.d.C.; Conceptualization, N.M.; methodology, N.M.; supervision, N.M.; writing—review and editing, N.M.; writing—original draft preparation, A.A.; software, A.A.; data curation, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by national funds from FCT-Foundation for Science and Technology (RESEARCH 4 COVID-19): Project COMPRIME (Get to Know More for Intervention)-ID: 596685735 and Project COMPRI_MOv (Get to Know More for Intervention in the context of mobility)-ID: 613765655.

Data Availability Statement: The original data used in the dashboard study is available in the following URL: <https://covid19.min-saude.pt/relatorio-de-situacao/>. The processed data used in the dashboard is available in the following URL: <https://www.comprime-compri-mov.com>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO—World Health Organization. Ten Threats to Global Health in 2019. 2019. Available online: <https://www.who.int/news-room/feature-stories/ten-threats-to-global-health-in-2019> (accessed on 19 September 2020).
2. WHO—World Health Organization. WHO Director-General’s Opening Remarks at the Media Briefing on COVID-19—11 March 2020. 2020. Available online: <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020> (accessed on 10 October 2020).
3. WHO—World Health Organization. Coronavirus Disease (COVID-19) Outbreak Situation, 2020. WHO, 2020. Available online: <https://covid19.who.int/> (accessed on 6 December 2020).
4. WHO—World Health Organization. *SARS: How a Global Epidemic Was Stopped*; World Health Organization Western Pacific Region: Geneva, Switzerland, 2006.
5. DGS—Direção-Geral da Saúde. REACT-COVID: Inquérito sobre Alimentação e Atividade Física em Contexto de Contenção Social. 2020. Available online: <https://nutrimento.pt/noticias/react-covid/> (accessed on 9 February 2021).

6. Boulos, K.; Geraghty, E. Geographical tracking and mapping of coronavirus disease COVID 19/severe acute respiratory syndrome coronavirus 2 (SARS CoV 2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *Int. J. Health Geogr.* **2020**, *19*, 8. [CrossRef]
7. DGS—Direção-Geral da Saúde. Relatório de Situação. Lisboa: Ministério da Saúde—Direção Geral da Saúde. 2020. Available online: <https://covid19.min-saude.pt/relatorio-de-situacao/> (accessed on 9 February 2021).
8. Cliff, A.; Haggett, P. *Atlas of disease Distributions: Analytic Approaches to Epidemiological Data*; Blackwell Publishers: Oxford, UK, 1993.
9. Lyseen, A.K.; Nøhr, C.; Sørensen, E.M.; Gudes, O.; Geraghty, E.M.; Shaw, N.T.; Bivona-Tellez, C. A review and framework for categorizing current research and development in health related geographical information systems (GIS) studies. *Yearb Med. Inform.* **2014**, *23*, 110–124. [CrossRef]
10. Zhou, C.; Su, F.; Pei, T.; Zhang, A.; Du, Y.; Luo, B.; Cao, Z.; Wang, J.; Yuan, W.; Zhu, Y.; et al. COVID-19: Challenges to GIS with Big Data. *Geogr. Sustain.* **2020**, *1*. [CrossRef]
11. Franch-Pardo, I.; Napoletano, B.M.; Rosete-Verges, F.; Billa, L. Spatial analysis and GIS in the study of COVID-19. A review. *Sci. Total Environ.* **2020**, *739*, 140033. [CrossRef] [PubMed]
12. Sarfo, A.; Karuppannan, S. Application of Geospatial Technologies in the COVID-19 Fight of Ghana. *Trans. Indian Natl. Acad. Eng.* **2020**, *5*. [CrossRef]
13. Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534. [CrossRef]
14. Fernandez-Lozano, C.; Cedron, F. Shiny Dashboard for Monitoring the COVID-19 Pandemic in Spain. *Proceedings* **2020**, *54*, 23. [CrossRef]
15. Barone, S.; Chakhunashvili, A.; Comelli, A. Building a statistical surveillance dashboard for COVID-19 infection worldwide. *Qual. Eng.* **2020**, *32*, 754–763. [CrossRef]
16. Florez, H.; Singh, S. Online dashboard and data analysis approach for assessing COVID-19 case and death data. *F1000Research* **2020**, *9*, 1–13. [CrossRef] [PubMed]
17. ECDC—European Centre for Disease Prevention and Control (2020). Situation Updates on COVID-19. Available online: <https://www.ecdc.europa.eu/en/covid-19/situation-updates> (accessed on 11 October 2020).
18. DGS—Direção-Geral da Saúde & Esri. Coronavírus (COVID-19): Prevenção Através de Dashboards. 2020. Available online: <https://www.esri-portugal.pt/pt-pt/landing-pages/covid19> (accessed on 9 February 2021).
19. INE—Instituto Nacional de Estatística. Dashboard COVID-19: COVID-19 | Contexto e Impacto. 2020. Available online: <https://ine-pt.maps.arcgis.com/apps/opstdashboard/index.html#/7af78fbbdd9456397317f822dac503d> (accessed on 9 February 2021).
20. COTEC e NOVA IMS—Information Management School da Universidade Nova de Lisboa. COVID 19 Insights. 2020. Available online: <https://insights.cotec.pt/> (accessed on 9 February 2021).
21. APG—Associação Portuguesa de Geógrafos. Acompanhamento da Pandemia COVID-19 pela APG. Available online: <http://www.apgeo.pt/acompanhamento-da-pandemia-covid-19-pela-apg> (accessed on 9 February 2021).
22. Yigitbasiglu, O.; Velcu, O. A review of dashboards in performance management: Implications for design and research. *Int. J. Account. Inf. Syst.* **2012**, *13*, 41–59. [CrossRef]
23. Jing, C.; Du, M.; Li, S.; Liu, S. Geospatial Dashboards for Monitoring Smart City Performance. *Sustainability* **2019**, *11*, 5648. [CrossRef]
24. Kourtit, K.; Nijkamp, P. Big data dashboards as smart decision support tools for i-cities—An experiment on Stockholm. *Land Use Policy* **2018**, *71*, 24–35. [CrossRef]
25. Simms, R.; Ping, A.; Yelland, A.; Beringer, A.; Fox, R.; Draycott, T. Development of maternity dashboards across a UK health region; current practice, continuing problems. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2013**, *170*, 119–124. [CrossRef] [PubMed]
26. Few, S. *Information Dashboard Design: The Effective Visual Communication of Data*; O'Reilly Media: Boston, MA, USA, 2006.
27. Pappas, L.; Whitman, L. Riding the technology wave: Effective dashboard data visualization. In *Symposium on Human Interface*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 249–258.
28. Donohue, R.; Sack, C.; Roth, R. Time Series Proportional Symbol Maps with Leaflet and jQuery. *Cartogr. Perspect.* **2013**, 43–66. [CrossRef]
29. Roser, M.; Ritchie, H.; Ortiz-Ospina, E.; Hasell, J. Coronavirus Pandemic (COVID-19). 2020. Available online: <https://ourworldindata.org/coronavirus> (accessed on 9 February 2021).
30. Marques da Costa, E.; Marques da Costa, N. A Pandemia de COVID-19 em Portugal Continental—uma análise geográfica da evolução verificada nos meses de março e abril. *Hygeia Rev. Bras. Geogr. Médica Saúde* **2020**, *72*, 72–79. [CrossRef]
31. Marques da Costa, E.; Marques da Costa, N. O processo pandémico da Covid-19 em Portugal Continental. Análise geográfica dos primeiros 100 dias. *Finisterra* **2020**, *115*, 11–18. [CrossRef]
32. Sá Marques, T.; Santos, H.; Honório, F.; Ferreira, M.; Ribeiro, D.; Barbosa, M. O Mosaico Territorial do Risco ao Contágio e à Mortalidade por COVID-19 em Portugal Continental. *Finisterra* **2020**, *115*, 19–26. [CrossRef]
33. Matheus, R.; Janssen, M.; Maheshwari, D. Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Gov. Inf. Q.* **2020**, *37*, 101284. [CrossRef]

Article

Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks

Aleksandr Romanov ¹, Anna Kurtukova ^{1,*}, Alexander Shelupanov ¹, Anastasia Fedotova ¹ and Valery Goncharov ²

¹ Department of Security, Tomsk State University of Control Systems and Radioelectronics, 634050 Tomsk, Russia; alexx.romanov@gmail.com (A.R.); saa@tusur.ru (A.S.); afedotowaa@icloud.com (A.F.)

² Department of Automation and Robotics, The National Research Tomsk Polytechnic University, 634050 Tomsk, Russia; gvi@tpu.ru

* Correspondence: av.kurtukova@gmail.com

Abstract: The article explores approaches to determining the author of a natural language text and the advantages and disadvantages of these approaches. The importance of the considered problem is due to the active digitalization of society and reassignment of most parts of the life activities online. Text authorship methods are particularly useful for information security and forensics. For example, such methods can be used to identify authors of suicide notes, and other texts are subjected to forensic examinations. Another area of application is plagiarism detection. Plagiarism detection is a relevant issue both for the field of intellectual property protection in the digital space and for the educational process. The article describes identifying the author of the Russian-language text using support vector machine (SVM) and deep neural network architectures (long short-term memory (LSTM), convolutional neural networks (CNN) with attention, Transformer). The results show that all the considered algorithms are suitable for solving the authorship identification problem, but SVM shows the best accuracy. The average accuracy of SVM reaches 96%. This is due to thoroughly chosen parameters and feature space, which includes statistical and semantic features (including those extracted as a result of an aspect analysis). Deep neural networks are inferior to SVM in accuracy and reach only 93%. The study also includes an evaluation of the impact of attacks on the method on models' accuracy. Experiments show that the SVM-based methods are unstable to deliberate text anonymization. In comparison, the loss in accuracy of deep neural networks does not exceed 20%. Transformer architecture is the most effective for anonymized texts and allows 81% accuracy to be achieved.

Keywords: authorship; text mining; machine learning; attribution; neural networks; deep learning; forensic intelligence

Citation: Romanov, A.; Kurtukova, A.; Shelupanov, A.; Fedotova, A.; Goncharov, V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* **2021**, *13*, 3. <https://doi.org/10.3390/fi13010003>

Received: 10 December 2020

Accepted: 23 December 2020

Published: 25 December 2020

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

It is now known that it is possible to determine the individual characteristics of the author on the basis of the writing style, since each text has a specific linguistic personality [1].

The topic of attribution overlaps with information security [2–5]. With the constant increase in volume of transmitted and received documents, there are many opportunities for the illegal use of personal data. An example is a type of fraud in which an attacker sends an employee of an organization an email on behalf of a manager asking them to perform a specific action (e.g., to divulge confidential information of the organization or to transfer funds). In addition, quite often there are situations related to hacking the victim's social media accounts and sending messages on the victim's behalf. One solution to this kind of problem is to compare the writing style of the suspicious texts with others for which it is certain that they were written by the person. As a result of the comparison, it is possible to determine the author. Establishing general differences in the documents based on the writing style is most relevant if there are no other data that would allow the author to be identified.

One type of violation in cyberspace is a copyright infringement and related rights of a text, which can be expressed, for example, by claiming a text by another author for material gain or attempting to pass off the authorship of a created text as the authorship of another person. The effectiveness of intellectual property protection in the digital space is determined by an ability to resist such violations and threats of their occurrence. Authorship identification methods allow determining such infringements and establishing the identity of the text creator.

Interest in the topic is also due to a growth in the volume of text data, the evolution of technology, and social networks. Thus, automatic identification of authorship is a growing area of research, which is also important in the fields of forensic science and marketing.

In this article, we solve the problem of identifying the author of a Russian-language text using a support vector machine and deep neural networks. Literary texts written by Russian-speaking writers were used as input data. The article includes an overview of related works, the statement of the text authorship problem, a detailed description of approaches to solving the authorship identification problem, an impact evaluation of attacks on the developed approaches, and a discussion of the results obtained.

2. Related Works

An excellent overview of articles up to 2010 is presented in [1]. However, since then, methods based on deep neural networks (NN) have become more and more popular, replacing classical methods of machine learning. For example, the topic of author identification is considered annually at the PAN conference [6]. As part of the conference, researchers were offered two datasets of different sizes, containing texts by well-known authors.

The authors of [7] emphasize that they have proposed an approach that takes into account only the topic-independent features of a writing style. Guided by this idea, the authors chose several features such as the frequency of punctuation marks, highlighting the last word in a sentence, consideration of all existing categories of functional words, abbreviations and contractions, verb tenses, and adverbs of time and place. An ensemble of classifiers was used in the work. Each of them accepts or rejects the supposed authorship. Research is distinguished by the application of an approach that, in general, is aimed at recognizing a person based on his behavior. Here, Equal Error Rate (EER) has been applied as the thresholding mechanism. Essentially, the EER corresponds to the point on the curve where the false acceptance rate is equal to the false rejection rate. The results are 80% and 78% accuracy for the large and small datasets, respectively. The results of the approach allowed the authors to take third place among all the submitted works.

In [8], stylometric features were extracted for each pair of documents. The absolute difference between the feature vectors was used as input data for the classifier. Logistic regression was used for a small dataset, and a NN was used for a large one. These models achieved 86% and 90% accuracy for small and large datasets, respectively. As a result, the authors of the study took second place.

The work that achieved the best result in the competition [9] presents the combination of NN with statistical modeling. Research is aimed at studying pseudo metrics that represent a variable-length text in the form of a fixed-size feature vector. To estimate the Bayesian factor in the studied metric space, a probability layer was added. The ADHOMINEM system [10] was designed to transmit the association of selected tokens into a two-level bi-directional long short-term memory (LSTM) network with an attention mechanism. Using additional attention levels made it possible to visualize words and sentences that were marked by the system as “very significant”. It was also found that using the sliding window method instead of dividing a text into sentences significantly improves results. The proposed method showed excellent overall performance, surpassing all other systems in the PAN 2020 competition on both datasets. The accuracy was 94% for the large dataset and 90% for the small one.

The authors of [11] took into account the syntactic structure of a sentence when determining the author of a text, highlighting two components of the self-supervised

network: lexical and syntactic sub-network, which took a sequence of words and their corresponding structural labels as input data. The lexical sub-network was used to code a sequence of words in a sentence, while the syntactic sub-network was used to code selected labels, e.g., parts of speech. The proposed model was trained on the publicly available LAMBADA dataset, which contains 2662 texts of 16 different genres in English. The consideration of the syntactic structure made it possible to eliminate the need for semantic analysis. The resulting accuracy was 92.4%.

The work in [12] provides an overview of the methods for establishing authorship with the possibility of their subsequent application in the field of forensic research on social networks. According to the authors, in forensic sciences, there is a significant need for new attribution algorithms that can take context into account when processing multimodal data. Such algorithms should overcome the problem of a lack of information about all candidate authors during training. Functional words have been chosen as a feature, as they are quite likely to appear even in small samples and can therefore be particularly effective for analyzing social networks. Combinations of different sets of n -grams at symbol and word level with n -grams at the part-of-speech level were investigated. An accuracy of 70% was obtained for 50 authors.

The main idea of the study [13] is to modify the approach to establishing authorship by combining it with pre-trained language models. The corpus of texts consisted of essays by 21 undergraduate students written in five formats (essay, email, blog post, interview, and correspondence). The method is based on a recurrent neural network (RNN) operating at the symbol level and a multiheaded classifier. In cross-thematic authorship determination, the results were 67–91%, depending on the subject, and in cross-genre, 77–89%, depending on the genre.

The essence of [14] is to research document vectors based on n -grams. Experiments were conducted on a cross-thematic corpus containing some articles from 1999 to 2009 published in the English newspaper *The Guardian*. Articles by 13 authors were collected and grouped into five topics. To avoid overlapping, those articles for which content included more than one category were discarded. The results show that the method is superior to linear models based on n -gram symbols. To train the Doc2vec model, the authors used a third-party library called GENSIM 3. The best results were achieved on texts of large sizes. Accuracy for different categories ranged from 90.48 to 96.77%.

In [15], an ensemble approach that combines predictions made by three independent classifiers is presented. The method based on variable-length n -gram models and polynomial logistic regression and used to select the highest likelihood prediction among the three models. Two evaluation experiments were conducted: using the PAN-CLEF 2018 test dataset (93% accuracy) and a new corpus of lyrics in English and Portuguese (52% accuracy). The results demonstrate that the proposed approach is effective for fiction texts but not for lyrics.

The research conducted in [16] used the support vector machine (SVM). Parameters for defining the writing style were highlighted at different levels of the text. The authors demonstrated that more complex parameters are capable of extracting the stylometric elements presented in the texts. However, they are most efficiently used in combination with simpler and more understandable n -grams. In this case, they improve the result. The dataset included 20 samples in four different languages (English, French, Italian, and Spanish). Thus, five samples from 500 to 1000 words in each language were used. The challenge was to assign each document in the set of unknown documents to a candidate author from the problem set. The results were 77.7% for Italian, 73% for Spanish, 68.4% for French, and 55.6% for English.

Authorship identification methods are used not only for literary texts but also to determine plagiarism in scientific works. For example, [17] presents a system for resolving the ambiguity of authorship of articles in English using Russian-language data sources. Such a solution can improve the search results for articles by a specific author and the calculation of the citation index. The link.springer.com database was used as the initial

repository of publications, and the eLIBRARY.ru scientific electronic library was used to obtain reliable information about authors and their articles. To assess the quality of the comparison, experiments were carried out on the data of employees of the A.P. Yershov Institute of Informatic Systems. The sample included 25 employees, whose publications are contained in the link.springer.com system. To calculate the similarity rate of natural language texts, they were presented as vectors in multidimensional space. To construct a vector representation of texts, a bag-of-words algorithm was used with the term frequency-inverse document frequency (TF-IDF) measure. Stop-words were preliminarily removed from the texts, and stemming of words was carried out. Experiments were also provided on the vectorization of natural language texts using the word2vec. The average percentage of the number of publications of authors recognized by the system was 79%, while the number of publications that did not belong to the author but were assigned to his group was close to zero. The approaches used in the system are applicable for disambiguating authorship of publications from various bibliographic databases. The implemented system showed a result of 92%.

There were only a few works that achieved a high level of author identification in Arabic texts. In [18], the Technique for Order Preferences by Similarity to Ideal Solution (TOPSIS) was used to select the basic classifier of the ensemble. More than 300 stylometric parameters were extracted as attribution features. The AdaBoost and Bagging methods were applied to the dataset in Arabic. Texts were taken from six sources. Corpora included both short and long texts by three hundred authors writing in various genres and styles. The final accuracy was 83%.

A new area of research is attribution, which uses not only human-written texts but also texts obtained using generation [19]. Several recently proposed language models have demonstrated an amazing ability to generate texts that are difficult to distinguish from those written by humans. In [20], a study of the problem of authorship attribution is proposed in two versions: determining the authorship of two alternative human-machine texts and determining the method that generated the text. One human-written text and eight machine-generated texts (CTRL, GPT, GPT2, GROVER, XLM, XL-NET, PPLM, FAIR) were used. Most generators still produce texts that significantly differ from texts written by humans, which makes it easier to solve the problem. However, the texts generated by GPT2, GROVER, and FAIR are of significantly better quality than the rest, which often confuses classifiers. For these tasks, convolutional neural networks (CNN) were used, since the CNN architecture is better suited to reflect the characteristics of each author. In addition, the authors improved the implementation of the CNN using n -gram words and part-of-speech (PoS) tags. The result in the “human-machine” category ranges from 81% to 97%, depending on the generator, and, for determining the generation method, 98%.

The author of [21] presented the software product StylometRy, which allows the identification of the author of a disputed text. Texts were presented in the form of a bag-of-words model. Naive Bayesian classifier, k -nearest method, and logistic regression were chosen as classifiers, and pronouns were used as linguistic features. The models were checked in L. Tolstoy, M. Gorky, and A. Chekhov texts. The minimum text volume for analysis was 5500 words. The accuracy of the model for texts over 150,000 characters was in the range of 60–100% (average 87%).

The scientific work [22] describes the features of four styles of the Russian language—scientific, official, literary, and journalistic. The parameters selected for texts analysis were: the ratio of the number of verbs, nouns, adjectives, pronouns, particles, and interjections to the number of words in the text, the number of “noun + noun” constructions, the number of “verb + noun” constructions, the average word length, and the average sentence length. Decision trees were used for classification. The accuracy of the analysis of 65 texts of each style was 88%. The highest accuracy was achieved when classifying official and literary texts, and the lowest was achieved for journalistic texts.

The authors of [23] present the analysis and application of various NNs architectures (RNN, LSTM, CNN, bi-directional LSTM). The study was conducted based on three datasets

in Russian (Habrahabr blog—30 authors, average text length 2000 words; vk.com—50 and 100 authors, average text length 100 words; Echo.msk.ru—50 and 100 authors, average text length 2000 words). The best results were achieved by CNN (87% for Habrahabr blog, 59% and 53% for 50 and 100 authors with vk.com, respectively). Character's trigrams performed significantly better for short texts from social networks, while for longer texts, both trigram and tetragram representations achieved almost the same accuracy (84% for trigrams, 87% for tetragram representations).

The object of research study [24] is journalistic articles from Russian pre-revolutionary magazines. The information system Statistical Methods of Literary Texts Analysis (SMALT) has been developed to calculate various linguistic and statistical features (distribution of parts of speech, average word and sentence length, vocabulary diversity index). Decision trees were used to determine the authorship. The resulting accuracy was 56%.

The problem of authorship attribution of short texts obtained from Twitter was considered in scientific work [25]. Authors proposed a method of learning text representations using a joint implementation of words and character n -grams as input to the NNs. Authors used an additional feature set with 10 elements: text length, number of usernames, topics, emoticons, URLs, numeric expressions, time expressions, date expressions, polarity level, and subjectivity level. Two series of comparative experiments were provided to test using CNN and LSTM. The method achieved an accuracy of 83.6% on the corpus containing 50 authors.

The authors of [26] applied integrated syntactic graphs (ISGs) to the task of automatic authorship attribution. ISGs allow for combining different levels of language description into a single structure. Textual patterns were extracted based on features obtained from the shortest path walks over integrated syntactic graphs. The analysis was provided on lexical, morphological, syntactic, and semantic levels. Stanford dependency parser and WordNet taxonomy were applied in order to obtain the parse trees of the sentences. The feature vectors extracted from the ISGs can be used for building syntactic n -grams by introducing them into machine learning methods or as representative vectors of a document collection. Authors showed that these patterns, used as features, allow determining the author of a text with a precision of 68% for the C10 corpus and also performed experiments for the PAN'13 corpus, obtaining a precision of 83.3%.

An approach based on joint implementation of words, n -grams, and the latent Dirichlet allocation (LDA) was presented in [27]. The LDA-based approach allows the processing of sparse data and volumetric texts, giving a more accurate representation. The described approach is an unsupervised computational methodology that is able to take into account the heterogeneity of the dataset, a variety of text styles, and also the specificity of the Urdu language. The considered approach was tested on 6000 texts written by 15 authors in Urdu. The improved sqrt-cosine similarity was used as a classifier. As a result, an accuracy of 92.89% was achieved.

The idea of encoding the syntax parse tree of a sentence into a learnable distributed representation is proposed in [28]. An embedding vector is created for each word in the sentence, encoding the corresponding path in the syntax tree for the word. The one-to-one correspondence between syntax-embedding vectors and words (hence their embedding vectors) in a sentence makes it easy to integrate obtained representation into the word-level Natural Language Processing (NLP) model. The demonstrated approach has been tested using CNN. The model consists of five types of layers: syntax-level feature embedding, content-level feature embedding, convolution, max pooling, and softmax. The accuracy obtained on the datasets was 88.2%, 81%, 96.16%, 64.1%, and 56.73% on five benchmarking datasets (CCAT10, CCAT50, IMDB62, Blogs10, and Blogs50, respectively).

The authors of [29] combined widely known features of texts (verbs tenses frequency, verbs frequency in a sentence, verbs usage frequency, commas frequency in a sentence, sentence length frequency, words usage frequency, words length frequency, characters n -gram frequency) and genetic algorithm to find the optimal weight distribution. The genetic algorithm is configured with a mutation probability of 0.2 using a Gaussian convolution on the values with a standard deviation of 0.3 and evolved over 1000 generations. The method

was tested on the Gutenberg Dataset, consisting of 3036 texts written by 142 authors. The method is implemented using Stanford CoreNLP, stemming, PoS tagging, and genetic algorithm. The obtained accuracy was 86.8%.

There is no generally accepted opinion regarding the set of text features that provides the best result. In most works, text features such as bigrams and trigrams of symbols and words, functional words, the most frequent words in the language, the distribution of words in parts of speech, punctuation marks, and the distribution of word length and sentence length have proven to be effective. It is incorrect to judge the accuracy of the methods applied to the Russian language based on the results of research in the English language or any other languages because of the specific structure of each language. The choice of approach depends on the text language, the authorship identification method, and the accuracy of the available analysis methods. Particularly, the peculiarity of the Russian language in comparison with English, for which most of the results are presented, is its flexibility and, consequently, more complex word formation and a high degree of morphological and syntactic homonymy, which makes it difficult to use some features useful for the English language. The problems of genre, sample representativeness, and dataset size also limit the implementation of some approaches.

Investigations aimed at finding a method with high separating ability with a large number of possible authors are not always useful when solving real-life tasks. It is necessary to continue further research aimed at finding new methods or improving/combining existing methods of identifying the author, as well as conducting experiments aimed at finding features that allow accurately dividing the styles of authors of Russian-language texts. By using these features, it will be possible to work with small samples.

3. Problem Statement

We define the identification of the text author as the process of determining the author based on a set of general and specific features of the text that formed the author's style.

The problem of identifying the author of the text with a limited set of alternatives is formulated as follows. There are the set of texts $T = \{t_1, \dots, t_k\}$ and the set of authors $A = \{a_1, \dots, a_l\}$. For a certain subset of texts $T' = \{t_1, \dots, t_m\} \subseteq T$, the authors are known; i.e., there are the set of text–author pairs $D = \{(t_i, a_j)\}_{i=1}^m$. It is necessary to determine which author from set A is the true author of the remaining texts (anonymous or disputed) $T'' = \{t_{m+1}, \dots, t_k\} \subseteq T$.

In this statement, the author's identification problem can be considered as a multi-label classification task. In this case, set A is the set of predefined classes and their labels, set D is the set of training samples, and objects to be classified are included in the set T'' . The goal is to develop a classifier that solves the problem—finding the objective function $F : T \times A \rightarrow [-1, 1]$, which assigns some text from the set T to its true author. The function value is described as the degree to which the object belongs to the class, where 1 corresponds to the completely positive solution, while -1 , on the contrary, is a negative one.

4. Methods for Determining the Author of a Natural Language Text

Early research [1] was aimed at evaluating the accuracy and the speed of classifiers based on machine learning algorithms. Then, the best results in all parameters were demonstrated by the SVM classifier. However, over the past 10 years, many solutions based on deep NNs appeared in the field of NLP: RNN and CNN for multi-label text categorization, category text generation, and learning word dependencies, and hybrid networks for aspect-based sentiment analysis. These solutions significantly exceed the effectiveness of traditional algorithms. As of 2020, LSTM, CNN with self-attention, and Transformer [30,31] are the models that successfully solve related text analysis problems. Thus, the purpose of the study was to compare SVM with modern classification methods based on deep NN. The enumerated models, their mathematical apparatuses, as well as the techniques of their application to the task of authorship attribution are described below.

4.1. Support Vector Machine

The SVM classifier is similar to the classical perceptron. Application of its kernel transformations allows training radial basis function network and perceptron with a sigmoidal activation function, the weights of which are determined by solving a quadratic programming problem with linear constraints, while training a standard NN implies solving the problem of non-convex minimization without restrictions. In addition, SVM allows working directly with a high-dimensional vector space without preliminary analysis and also without manually selecting the number of neurons in the hidden layer.

The main difference between SVM and deep-learning models is that SVM is unable to find unobvious informative features in text that have not been pre-processed. Therefore, it is necessary to first extract such features from the text.

Let us denote the set of letters of the alphabet, numbers, and separators $\mathbf{A} = \{a_1, a_2, \dots, a_{|A|}\}$, the set of possible morphemes $\mathbf{M} = \{m_1, m_2, \dots, m_{|M|}\}$, the language dictionary $\mathbf{W} = \{w_1, w_2, \dots, w_{|W|}\}$, the set of phrases $\mathbf{C} = \{c_1, c_2, \dots, c_{|C|}\}$, the set of sentences $\mathbf{S} = \{s_1, s_2, \dots, s_{|S|}\}$, and the set of paragraphs $\mathbf{P} = \{p_1, p_2, \dots, p_{|P|}\}$. Then, the text T can be represented as sequences of elements as follows:

$$T = \{a_j^i\}_{i=1}^{N_a} = \{m_j^i\}_{i=1}^{N_m} = \{w_j^i\}_{i=1}^{N_w} = \{c_j^i\}_{i=1}^{N_c} = \{s_j^i\}_{i=1}^{N_s} = \{p_j^i\}_{i=1}^{N_p}, \quad (1)$$

where $a_j^i \in \mathbf{A}$, $m_j^i \in \mathbf{M}$, $w_j^i \in \mathbf{W}$, $c_j^i \in \mathbf{C}$, $s_j^i \in \mathbf{S}$, $p_j^i \in \mathbf{P}$; $N_a, N_m, N_p, N_w, N_c, N_s$ —the number of characters, morphemes, words, phrases, sentences, paragraphs in the text.

Thus, the SVM feature space can be described as vectors of features that reflect the properties of text elements: $\{a'_1, \dots, a'_k\}$ for symbols, $\{m'_1, \dots, m'_l\}$ for morphemes, $\{w'_1, \dots, w'_n\}$ for words, $\{c'_1, \dots, c'_r\}$ for phrases, $\{s'_1, \dots, s'_t\}$ for sentences, and $\{p'_1, \dots, p'_u\}$ for paragraphs.

In the study, when classifying with SVM, informative features are used as an unordered collection as inputs of the SVM. The frequencies of single text's elements are used as follows:

$$t_k = \begin{cases} 1 & \Leftrightarrow w_j^i \in \mathbf{W} \\ 0 & \Leftrightarrow w_j^i \notin \mathbf{W} \end{cases}, j = \overline{1, n_i}, k = \overline{1, |\mathbf{W}|}, \quad (2)$$

In addition, the texts elements sequences of some length (n -grams) or a limited number of them from the dictionary are used as follows:

$$f(a_i, \dots, a_{i+n-1}) = \frac{C(a_i, \dots, a_{i+n-1})}{L}, \quad (3)$$

$$P(a_i | a_{i-n+1} \dots a_{i-1}) = \frac{C(a_{i-n+1}, \dots, a_i)}{C(a_{i-n+1}, \dots, a_{i-1})}, \quad (4)$$

where L —total number of counted n -grams; k —threshold value; $f()$ —relative frequency of the element in the text; a —the symbol; $P()$ —the probability of the element appearing in the text; n —the length of the n -gram.

It should be noted that for texts of small volumes, it is supposed to use frequencies smoothed by the methods of Laplace (5), Good-Turing (6), and Katz (7), which makes it possible to estimate the probabilities of non-occurring events:

$$P_{ADD}(a_i, \dots, a_{i+n-1}) = \frac{1 + C(a_i, \dots, a_{i+n-1})}{\mathbf{W} + \sum_i C(a_i, \dots, a_{i+n-1})}, \quad (5)$$

where P_{ADD} —estimates of Laplace; \mathbf{W} —the language dictionary; $C()$ —the number of occurrences of the element in the text.

$$P_{GT}^* = \frac{C^*}{N}, P_{GT}^* = \frac{N_1}{N} C^* = (C + 1) \frac{N_{C+1}}{N_C}, \quad (6)$$

where P_{GT} —estimates of Laplace; N —the total number of the considered elements of the text; N_C —the number of text elements encountered exactly C times; C^* —discounted Good Turing estimate.

$$P_{KATZ}(a_i|a_{i-n+1}, \dots, a_{i-1}) = \begin{cases} P^*(a_i|a_{i-n+1}, \dots, a_{i-1}), & \text{if } C(a_{i-n+1}, \dots, a_i) > k \\ \alpha(a_{i-n+1}, \dots, a_{i-1}) P_{KATZ}(a_i|a_{i-n+2}, \dots, a_{i-1}), & \text{if } 1 \leq C(a_{i-n+1}, \dots, a_i) \leq k. \end{cases} \quad (7)$$

where t_k —the fact of the existence of the j -th word of the i -th text in the dictionary \mathbf{W} ; P_{KATZ} —estimates of Katz; $\alpha()$ —weight coefficient.

In the process of authorship attribution of natural language text using classical machine learning methods, not only standard feature sets can be used; features obtained as a result of solving related tasks such as determining the author's gender and age, the level of the author's education, the sentiment of the text, etc. can also be used. However, as a part of this study, aspect-oriented analysis was also used for informative features extraction. Such a type of analysis involves understanding the meaning of a text by identifying aspect terms or categories. Thus, it becomes possible to extract keywords and opinions related to aspects.

There are two well-known approaches to implementing aspect analysis: statistical and linguistic. The statistical approach is performed as an extraction of aspects, determination of the threshold value for them, and selection such aspects, the values of which are indicated above the given threshold. The linguistic approach takes into account the syntactic structure of the sentence and searches for aspects by patterns.

We decided to use a combination of these methods. Aspects chosen were nouns and noun phrases (statistical approach), and the syntactic structure of the sentence was determined based on the dependencies between words (linguistic approach).

Multi-layered NN, consisting of fully connected layers, was implemented to extract aspects. The following training parameters were used:

- Optimization algorithm—adaptive moment estimation (Adam);
- Regularization procedure—dropout (0.3);
- Loss function—Binary cross-entropy;
- Hidden layers activation function—ReLU;
- Function of activation of the output layer—Sigmoid.

The principle of operation of SVM is to construct a hyperplane in the space of high-dimensional features in such a way that the gap between the support vectors (the extreme points of the two classes) is maximized. The mapping of the original data onto space with the linear separating surface is performed using a kernel transformation:

$$(\Phi(x), \Phi(x')) = k(x, x'), \quad (8)$$

where $(\Phi(x), \Phi(x'))$ is the inner product between the sample being recognized and the training samples, and k is some mapping of the original space onto the space with the inner product (the space of dimension sufficient for linear separability).

Then the function performing the classification looks like this:

$$f(x) = \left\{ \sum_{i=1}^l \alpha_i y_i k(x_i, x) \right\} + b, \quad (9)$$

where α is the optimal coefficient, k is the kernel function, y is the label of class, b is the parameter that ensures the fulfillment of the second Karush-Kuhn-Tucker condition for all input samples corresponding to Lagrange multipliers that are not on the boundaries.

The optimal coefficient α is determined by maximizing the objective function:

$$W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j k(x_i, x_j), \quad (10)$$

where the maximization condition:

$$\sum_{i=1}^l \alpha_i y_i = 0, \quad (11)$$

in the positive quadrant $0 \leq \alpha_i \leq C, i = \overline{1, l}$.

The regularization parameter C determines the ratio between the number of errors in the training set and the size of the gap.

4.2. Deep Neural Networks

A distinctive feature of deep NNs is their ability to analyze a text sequence and extract informative features by itself. In some studies, texts should be accepted by the model unchanged [1]. However, in solving the problem of determining the author of a natural language text, preliminary preparation is an important stage.

The purpose of preprocessing is to cleaning the dataset from noise and redundant information. Within the framework of the study, the following actions were taken to clean up the texts:

- Converting text to lowercase;
- Removing stop-words;
- Removing special characters;
- Removing digits;
- White space formatting.

The data obtained from the results of preprocessing must be converted into a vector-understandable NN. For this purpose, it was decided to use word embeddings—a text representation, where words having a similar meaning are defined by vectors close to each other in hyperspace. The received word representations are fed to the inputs of the deep NN.

4.2.1. Long Short-Term Memory

LSTM is a successful modification of the classical RNN, which avoids the problem of vanishing or exploding gradients. This is due to the fact that the semantic weights of the LSTM model are the same for all time steps during error backpropagation. Therefore, the signal becomes too weak (exponentially decreases) or too strong (exponentially increases). This is the problem that LSTM solves.

The LSTM model contains the following elements:

- Forget Gate “f”—an NN with sigmoid;
- Candidate layer “C”—an NN with Tanh;
- Input Gate “I”—an NN with sigmoid;
- Output Gate “O”—an NN with sigmoid;
- Hidden state “H”—an vector;
- Memory state “C”—an vector.

Then the time step t is considered. The input to the LSTM cell is the current input vector \mathbf{X}_t , the previous hidden state H_{t-1} , and the previous memory state C_{t-1} . The cell outputs are the current hidden state H_t and the current memory state C_t . The following formulas are used to calculate outputs:

$$f_t = \sigma(\mathbf{X}_t * \mathbf{U}_f + H_{t-1} * \mathbf{W}_f), \quad (12)$$

$$\overline{C}_t = \tanh(\mathbf{X}_t * \mathbf{U}_c + H_{t-1} * \mathbf{W}_c), \quad (13)$$

$$I_t = \sigma(\mathbf{X}_t * \mathbf{U}_i + H_t * \mathbf{W}_i), \quad (14)$$

$$O_t = \sigma(\mathbf{X}_t * \mathbf{U}_o + H_{t-1} * \mathbf{W}_o), \quad (15)$$

where \mathbf{X}_t —the input vector; H_{t-1} —the hidden state of the previous cell; C_{t-1} —the memory state of the previous cell; H_t —the hidden state of the current cell; C_t —the memory state of the current cell at time t ; \mathbf{W} , \mathbf{U} are the weight vectors for the forget gate $f()$, the gate of candidates, i.e., an input and output gates; σ —sigmoidal function; \tanh —tangent function.

The most important role is the state of memory \bar{C}_t . It is the state in which the input context is stored. It changes dynamically depending on the need to add or remove information. If the value of the forget gate is 0, then the previous state is completely forgotten; if equal to 1, then it is completely transferred to the cell. With the current state of C_t memory, a new one can be calculated:

$$C_t = f_t * C_{t-1} + I_t * \bar{C}_t. \quad (16)$$

Then it is necessary to calculate the output from the hidden state H at time t . It will be based on memory state:

$$H_t = O_t * \tanh(C_t), \quad (17)$$

Received C_t and H_t are transferred to the next time step, and the process is repeated.

4.2.2. CNN with Attention

CNN consists of many convolutional layers and subsampling layers. Each convolutional layer uses filters with input and output dimensions D_{in} and D_{out} . The layer is parameterized by the four-dimensional nuclear tensor \mathbf{W} of the measurement and the displacement vector $D_{out} \rightarrow b_{out}$. Therefore, the output value for some word q :

$$Y_q = \sum_{\Delta} \mathbf{X}_{q+\Delta} \mathbf{W}_q + b, \quad (18)$$

where Δ —kernel change.

The main difference between the attention mechanism and CNN is that the new meaning of a word is determined by every second word of the sentence, since the receptive field of attention includes the full context and not just a grid of nearby words.

The attention mechanism takes as input a token feature matrix, query vectors, and several key-value pairs. Each of the vectors is transformed by a trainable linear transform, and then the inner product query vectors are calculated with each key in turn. The result is run through Softmax, and with the weights obtained from Softmax, all vectors values are summed into a single vector. As a result of applying the attention mechanism, a matrix is obtained where the vectors contain information about the value of the corresponding tokens in the context of other tokens.

4.2.3. Transformer

The mechanism of attention in its pure form can lose information and complicate the convergence, and therefore a solution is required to this problem. Therefore, it was decided to also try its more complex modification—a transformer.

The transformer consists of an encoder and a multi-head attention mechanism. Some of the transformer layers are fully connected, and part of a shortcut is connected. A mandatory component of the architecture is multi-head attention, which allows each input vector to interact with other tokens using the attention mechanism. The study uses a common combination of multi-head attention, a residual layer, and a fully connected layer. The depth of the model is created by repeating this combination 6 times.

A distinctive feature of multi-head attention is that there are several attention mechanisms and they are trained in parallel. The final result is concatenated, passed through the training linear transformation once again, and goes to the output. Formally, it can be described as follows. The attention layer is determined by the size of the key/query

D_k , the number of heads N_h , the size of the head D_h , and the output D_{out} . The layer is parametrized with the key matrix, the query matrix \mathbf{W}_{qry}^x , and the value matrix \mathbf{W}_{val}^x for each head, together with the protector matrix \mathbf{W}_{out} used to assemble all the heads together. Attention for each head is calculated as:

$$A_q = \mathbf{X}_q : \mathbf{W}_{qry} \mathbf{W}_{key}^T \mathbf{X}_k^T. \quad (19)$$

The actual head value is calculated as:

$$H_q^{(h)} = \sum_{k' \in [W] \times [H]} softmax(A_q^{(h)})_{k'} \mathbf{X}_{k'} \mathbf{W}_{val}^{(h)}. \quad (20)$$

And the output value is calculated as follows:

$$H_q = concat(H_q^{(1)}, \dots, H_q^{(N_h)}) \mathbf{W}_{out} + b_{out}, \quad (21)$$

where \mathbf{X} —output values, \mathbf{W}_{key} —the matrix of keys, T —the transposition operation, A_q —the attention value for a particular head, k —the key position, q —the query position, N_h —the number of heads, b_{out} —the bias coefficient of the measurement D_{out} .

5. Experiment Setup and Results

About 45 groups of different features of text were used to train the SVM classifier [1]. Vectors ranging in size from 33 to 5000 features were used, including characteristics of different levels of text analysis:

- Lexical (punctuation, special symbols, lexicon, slang words, dialectic, archaisms);
- Morphological (lemmas, morphemes, grammar classes);
- Syntactic (complexity, position of words, completeness, sentiments);
- Structural (headings, fragmentation, citation, links, design, mention of location);
- Content-specific (keywords, emoticons, acronyms and abbreviations, foreign words);
- Idiosyncratic stylistic features (spelling and grammatical errors, anomalies);
- Document metadata (steganography, data structures).

Even a carefully selected feature space does not guarantee high model efficiency, but equally important are the training parameters of the SVM model. In an early study [1], the following parameters were identified as the most appropriate:

- Learning algorithm—sequential optimization method;
- Kernel—sigmoid;
- Regularization parameter $C = 1$;
- Acceptable error level—0.00001;
- Normalization—included;
- Compression heuristic—included.

As stated earlier, deep NNs do not need a predetermined set of informative text features, as they are able to search for them on their own. However, these models are also extremely sensitive to learning parameters. These parameters have been selected based on the results of model experiments for related tasks [32,33]:

- Optimization algorithm—adaptive moment estimation (Adam);
- Regularization procedure—dropout (0.2);
- Loss function—cross-entropy;
- Hidden layer activation function—rectified linear unit (ReLU);
- Output layer activation function—logistic function for multi-dimensional case (Softmax).

A large number of data are required to train models. For this purpose, the corpus was collected from the Moshkov library [34]. The corpus includes 2086 texts written by 500 Russian authors. The minimum size of each text was 100,000 symbols.

As part of experiments with models, the number of training examples varied with needs in solving real-life authorship identification tasks (including when the training data

are limited). Therefore, the texts were divided into fragments ranging from 1000 to 100,000 characters (~ 200–20,000 words). We used three training examples for each author and one for testing.

Table 1 shows the accuracy of the SVM model for datasets of 2, 5, 10, and 50 candidate authors. Table 2 shows the results of applying SVM trained on statistical features and extracted aspects. Cross-validation for 10-folds was used as a procedure for evaluating the effectiveness of the models.

Table 1. Average accuracy of author identification using SVM.

The Length of Text, Symbols	2 Authors	5 Authors	10 Authors	50 Authors
1000	0.9	0.71	0.64	0.49
5000	0.91	0.79	0.77	0.54
10,000	0.93	0.85	0.81	0.59
20,000	0.98	0.97	0.94	0.78
40,000	0.99	0.99	0.97	0.82
60,000	0.99	0.97	0.97	0.89
80,000	0.99	0.98	0.98	0.93
100,000	1	0.99	0.99	0.95
Average accuracy	0.96	0.91	0.88	0.75

Table 2. Average accuracy of author identification using SVM with extracted aspects.

The Length of Text, Symbols	2 Authors	5 Authors	10 Authors	50 Authors
1000	0.92	0.74	0.68	0.53
5000	0.93	0.81	0.79	0.58
10,000	0.95	0.87	0.85	0.62
20,000	0.99	0.98	0.96	0.81
40,000	0.99	0.99	0.98	0.84
60,000	0.99	0.99	0.98	0.91
80,000	1	0.99	0.99	0.95
100,000	1	0.99	0.99	0.97
Average accuracy	0.97	0.92	0.90	0.78

It should be noted that the results presented in Tables 1 and 2 were obtained by joint application of SVM and the Laplace smoothing method, which gives a slight increase in accuracy (from 0.01 to 0.07) on small sample sizes. Experiments have also shown that the Good-Turing and Katz smoothing methods negatively affect the quality of identification, with an average accuracy 0.04–0.11 lower when using them.

Table 3 shows the accuracy of determining the author using the LSTM for datasets of similar size and obtained by 10-fold cross-validation, while Table 4 shows the CNN with Attention and Table 5, the Transformer.

Table 3. Average accuracy of author identification using LSTM.

The Length of Text, Symbols	2 Authors	5 Authors	10 Authors	50 Authors
1000	0.68	0.51	0.4	0.23
5000	0.75	0.53	0.45	0.3
10,000	0.82	0.59	0.49	0.37
20,000	0.88	0.64	0.55	0.41
40,000	0.91	0.73	0.58	0.46
60,000	0.95	0.77	0.64	0.56
80,000	0.97	0.82	0.68	0.62
100,000	0.98	0.89	0.74	0.66
Average accuracy	0.87	0.69	0.57	0.45

Table 4. Average accuracy of author identification using CNN with attention

The Length of Text, Symbols	2 Authors	5 Authors	10 Authors	50 Authors
1000	0.84	0.61	0.59	0.36
5000	0.89	0.69	0.68	0.4
10,000	0.92	0.75	0.76	0.5
20,000	0.95	0.78	0.79	0.56
40,000	0.96	0.83	0.85	0.62
60,000	0.96	0.88	0.86	0.68
80,000	0.99	0.93	0.91	0.73
100,000	0.99	0.95	0.95	0.78
Average accuracy	0.94	0.80	0.79	0.58

Table 5. Average accuracy of author identification using Transformer

The Length of Text, Symbols	2 Authors	5 Authors	10 Authors	50 Authors
1000	0.86	0.64	0.6	0.34
5000	0.88	0.68	0.69	0.4
10,000	0.91	0.74	0.77	0.52
20,000	0.94	0.8	0.81	0.59
40,000	0.95	0.86	0.83	0.66
60,000	0.95	0.88	0.86	0.72
80,000	0.98	0.94	0.92	0.76
100,000	0.98	0.95	0.96	0.8
Average accuracy	0.93	0.81	0.80	0.60

Obtained results allow one to form a conclusion about the special effectiveness of SVM trained on accurately selected parameters and features. The approach based on SVM demonstrates superior accuracy to modern deep NNs architectures, regardless of the number of the samples and their volume. It should also be noted that the SVM classifier is able to learn on large volumes of data 10 times faster than deep NNs architectures. The average training time for SVM was 0.25 machine-hours, while deep models were trained for an average of 50 machine-hours.

6. Attacks on the Method

SVM classifier showed excellent results in determining the author of a natural-language text. However, keep in mind that the above experiments were not complicated by deliberate modifications aimed at text anonymization. Anonymization may have a negative impact on the accuracy of authorship identification. This hypothesis was confirmed by an early study [35]. A text anonymization technique was proposed based on a fast correlation filter, dictionary synonymizing, and a universal transformer model with a self-attention mechanism. The results of the study showed that decision-making accuracy can be reduced by almost 50% due to the proposed method of anonymization, keeping the text in readable and understandable form for humans.

As part of the work, it was decided to evaluate the described anonymization technique on the developed approaches. The results are presented in Table 6. The results of the experiments confirm that deep models are much more resistant to the anonymization technique than the SVM classifier. This is due to their ability to extract unobvious features that are not controlled by the author on a conscious level, while SVM operates on the basis of pre-defined features manually found by experts, and therefore text may be exposed to deliberate confusion by anonymization techniques. It should be noted that in such cases, SVM with aspect analysis shows a bit higher accuracy than SVM without it.

Table 6. Average accuracy of author identification using Transformer.

Number of Authors	Model	Accuracy before Anonymization	Accuracy after Anonymization
10	SVM	0.97	0.46
	SVM (with aspects)	0.98	0.52
	LSTM	0.84	0.66
	CNN with attention	0.93	0.78
	Transformer	0.94	0.81
30	SVM	0.95	0.42
	SVM (with aspects)	0.97	0.49
	LSTM	0.72	0.63
	CNN with attention	0.91	0.71
	Transformer	0.93	0.74
50	SVM	0.93	0.39
	SVM (with aspects)	0.95	0.44
	LSTM	0.68	0.59
	CNN with attention	0.75	0.68
	Transformer	0.77	0.69

7. Discussion and Conclusions

During the course of the research, the authors analyzed modern approaches to determining the author of a natural-language text, implemented approaches of authorship attribution based on SVM and deep NNs architectures, evaluated the developed approaches on different numbers of authors and volumes of texts, and evaluated the resistance of the approaches to anonymization techniques. The results obtained allow us to draw several conclusions.

Firstly, despite the great popularity of deep NNs architectures, they are inferior to the traditional SVM machine learning algorithms in accuracy by more than 10% on average. This is due to the fact that NNs require more data for learning than SVM to extract informative features from the text. However, when solving real-life authorship identification tasks, the number of data could be not enough for accurate decision-making by the NN.

Secondly, the SVM classification is based on an accurately found set of features manually formed by experts. Such informative features are also obvious for anonymization techniques and therefore can be removed or significantly corrupted. Thus, to solve the problem of identification of the author of a natural language text, both the SVM-based approach and deep models proposed by authors are equally suitable. However, when choosing an approach, the researched data and available technical resources should be objectively evaluated. In the case of a lack of resources, an SVM approach should be used. If there are traces of use anonymization in the text, despite the longer processing time, deep NNs architectures are recommended because they can find both the obvious and unobvious dependences in the text.

Thirdly, when using SVM, we recommended using five of the most informative features of the author's style that may improve the authorship identification process: unigrams and trigrams of Russian letters, high-frequency words, punctuation marks, and distribution of words among parts of speech.

Finally, based on the results obtained, as well as on the experience of earlier research, the authors identified the important criteria to obtain accurate results when identifying the author of a natural language text:

1. Author's personality: the informative features extracted from the text should contain all important information about the writing style. In this case, the authorship attribution system will be able to distinguish between the same or different authors.
2. Invariance: the writing style's characteristics should be stable for certain reasons, e.g., author's mood, emotional state, and the subject of the text.

3. Stability: the writing style may be influenced by the imitation of another author's writing style or by deliberately distorting the author's own style for other reasons. The chosen approach should be resistant to such actions.
4. Adaptability to the style of text: the author adapts to the specificities of the selected style of text to follow it. Adaptability to the style of text leads to significant changes in the characteristics of the author's writing style. In addition, when writing official documents, many people use ready templates and just fill in their own data. As a result, it is quite problematic to identify the similarity of official documents and, for example, messages in social networks written by one person.
5. Distinguishing ability: the selected informative features of the text should be significantly different for various authors, greater than the possible difference between the texts written by the same author. Selecting a single parameter that clearly separates two authors is problematic. Therefore, it should be a complex set of features from different levels of text that are not controlled by the author at the conscious level. In this case, the probability of wrong identification for different authors is reduced.

Author Contributions: Supervision, A.S.; writing—original draft, A.K., A.F.; writing—review and editing, A.R., V.G., A.S.; conceptualization, A.K., V.G., A.S.; methodology, A.K., A.R.; software, A.K., A.F.; validation, A.K., A.R., A.R.; formal analysis, A.K., A.F.; resources, A.S.; data curation, A.K., A.R.; project administration, A.R.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of Russia, Government Order for 2020–2022, project no. FEWM-2020-0037 (TUSUR).

Informed Consent Statement: Not applicable.

Acknowledgments: The authors express their gratitude to the editor and reviewers for their work and valuable comments on the article.

Conflicts of Interest: The authors declare no conflict of interest.



References

1. Romanov, A.S.; Shelupanov, A.A.; Meshcheryakov, R.V. Development and Research of Mathematical Models, Methods and Software Tools of Information Processes in the Identification of the Author of the Text, Tomsk: V-Spektr, 2011.
2. Kurtukova, A.; Romanov, A.; Fedotova, A. De-Anonymization of the Author of the Source Code Using Machine Learning Algorithms. In Proceedings of the 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Novosibirsk, Russia, 21–27 October 2019; pp. 612–617.
3. Kurtukova, A.V.; Romanov, A.S. Identification author of source code by machine learning methods. *SPIIRAS Proc.* **2019**, *18*, 741–765. [CrossRef]
4. Rakhmanenko, I.A.; Shelupanov, A.A.; Kostyuchenko, E.Y. Automatic text-independent speaker verification using convolutional deep belief network. *Comput. Opt.* **2020**, *44*, 596–605. [CrossRef]
5. Kostyuchenko, E.Y.; Viktorovich, I.; Renko, B.; Shelupanov, A.A. User Identification by the Free-Text Keystroke Dynamics. In Proceedings of the 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, Russia, 18–25 August 2018; pp. 1–4.
6. PAN: Shared Tasks. Available online: <https://pan.webis.de/shared-tasks.html> (accessed on 18 November 2020).
7. Halvani, O.; Graner, L.; Regev, R. Cross-domain authorship verification based on topic agnostic features. In Proceedings of the Working Notes of CLEF, Thessaloniki, Greece, 22–25 September 2020.
8. Feature Vector Difference Based Neural Network and Logistic Regression Models for Authorship Verification. Available online: https://pan.webis.de/downloads/publications/slides/weerasinghe_2020.pdf (accessed on 18 November 2020).
9. Boenninghoff, B. Deep bayes factor scoring for authorship verification. *arXiv* **2020**, arXiv:2008.10105.
10. Boenninghoff, B.; Hessler, S.; Kolossa, D.; Nickel, R.M. Explainable Authorship Verification in Social Media via Attention-based Similarity Learning. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019.
11. Jafariakinabad, F.; Hua, K.A. A Self-Supervised Representation Learning of Sentence Structure for Authorship Attribution. *arXiv* **2020**, arXiv:2010.06786.
12. Mamgain, S.; Balabantaray, R.C.; Das, A.K. Author Profiling: Prediction of Gender and Language Variety from Document. In Proceedings of the 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 19–21 December 2019; pp. 473–477.

13. Barlas, G.; Stamatatos, E. Cross-Domain Authorship Attribution Using Pre-Trained Language Models. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5–7 June 2020; pp. 255–266.
14. Gomez-Adorno, H. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing* **2018**, *100*, 741–756. [CrossRef]
15. Custodio, J.E.; Paraboni, I. An ensemble approach to cross-domain authorship attribution. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Lugano, Switzerland, 9–12 September 2019; pp. 201–212.
16. Bartelds, M.; de Vries, W. Improving Cross-domain Authorship Attribution by Combining Lexical and Syntactic Features. In Proceedings of the CLEF (Working Notes), Lugano, Switzerland, 9–12 September 2019; Volume 24.
17. Isachenko, V.V.; Apanovich, Z.V. System of analysis and visualization for cross-language identification of authors of scientific publications. *NSU Vestnik Inf. Technol.* **2018**, *16*, 29–60. [CrossRef]
18. El Bakly, A.H.; Darwish, N.R.; Hefny, H.A. Using Ontology for Revealing Authorship Attribution of Arabic Text. *Int. J. Eng. Adv. Technol. (IJEAT)* **2020**, *4*, 143–151.
19. Iskhakova, A.O. Method and Software for Determining Artificially Created Texts. Available online: <https://tusur.ru/ru/nauka-i-innovatsii/podgotovka-kadrov-vysshey-nauchnoy-kvalifikatsii/ob-yavleniya-o-zaschitah-dissertatsiy/dissertatsiya-metod-i-programmnoe-sredstvo-opredeleniya-iskusstvenno-sozdannyh-tekstov> (accessed on 18 November 2020).
20. Uchendu, A. Authorship Attribution for Neural Text Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020; pp. 8384–8395. Available online: <http://www.cs.iit.edu/~kshu/files/emnlp20.pdf> (accessed on 25 December 2020).
21. Chashchin, S.V. Application of “supervised” machine learning methods for text attribution: Individual approaches and intermediate results in identifying authors of Russian-language texts. *Probl. Criminol. Forensic Sci. Forensic Exam.* **2018**, *1*, 139–147.
22. Dubovik, A.R. Automatic determination of the stylistic affiliation of texts by their statistical parameters. *Comput. Linguist. Comput. Ontol.* **2017**, *1*, 29–45.
23. Dmitrin, Y.V. Comparison of deep neural network architectures for authorship attribution of Russian social media texts. In Proceedings of the Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue, 2018; Available online: http://www.dialog-21.ru/media/4560/_-dialog2018scopus.pdf (accessed on 25 December 2020).
24. Kulakov, K.A. Attribution of texts using mathematical methods and computer technologies. *Digit. Technol. Educ. Sci. Soc.* **2019**, *3*, 121–125.
25. Huang, W.; Su, R.; Iwaihara, M. Contribution of Improved Character Embedding and Latent Posting Styles to Authorship Attribution of Short Texts. In Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Tianjing, China, 12–14 August 2020; pp. 261–269.
26. Gómez-Adorno, H.; Sidorov, G.; Pinto, D.; Vilariño, D.; Gelbukh, A. Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors* **2016**, *16*, 1374. [CrossRef] [PubMed]
27. Anwar, W.; Bajwa, I.S.; Choudhary, M.A.; Ramzan, S. An empirical study on forensic analysis of urdu text using LDA-based authorship attribution. *IEEE Access* **2018**, *7*, 3224–3234. [CrossRef]
28. Zhang, R.; Hu, Z.; Guo, H.; Mao, Y. Syntax encoding with application in authorship attribution. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2742–2753.
29. Keyrouz, Y.; Fonlupt, C.; Robilliard, D.; Mezher, D. Evolving a Weighted Combination of Text Similarities for Authorship Attribution. In Proceedings of the International Conference on Artificial Evolution (Evolution Artificielle), Mulhouse, France, 29–30 October 2018; pp. 13–27.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Gomez, A.N.J.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
31. Chang, W.-C.; Yu, H.-F.; Zhong, K.; Yang, Y.; Dhillon, I. Taming Pretrained Transformers for Extreme Multi-label Text Classification. *arXiv* **2019**, arXiv:1905.02331.
32. Kurtukova, A.; Romanov, A.; Shelupanov, A. Source Code Authorship Identification Using Deep Neural Networks. *Symmetry* **2020**, *12*, 2044. [CrossRef]
33. Romanov, A.S.; Kurtukova, A.V.; Sobolev, A.A.; Shelupanov, A.A.; Fedotova, A.M. Determining the Age of the Author of the Text Based on Deep Neural Network Models. *Information* **2020**, *11*, 589. [CrossRef]
34. Moshkov’s Library. Available online: <http://lib.ru/> (accessed on 18 November 2020).
35. Romanov, A.; Kurtukova, A.; Fedotova, A.; Meshcheryakov, R. Natural Text Anonymization Using Universal Transformer with a Self-attention. In Proceedings of the III International Conference on Language Engineering and Applied Linguistics, Saint Petersburg, Russia, 27 November 2019; pp. 22–37.

Article

A Data Augmentation Approach to Distracted Driving Detection

Jing Wang ^{1,2,3} , ZhongCheng Wu ^{1,2}, Fang Li ^{1,3,*} and Jun Zhang ^{1,3} 

- ¹ High Magnetic Field Laboratory, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China; wj2019@mail.ustc.edu.cn (J.W.); zcwu@iim.ac.cn (Z.W.); zhang_jun@hmf.ac.cn (J.Z.)
- ² Graduate School of Computer Applied Technology, University of Science and Technology of China, Hefei 230026, China
- ³ High Magnetic Field Laboratory of Anhui Province, Hefei 230031, China
- * Correspondence: lif@hmf.ac.cn

Abstract: Distracted driving behavior has become a leading cause of vehicle crashes. This paper proposes a data augmentation method for distracted driving detection based on the driving operation area. First, the class activation mapping method is used to show the key feature areas of driving behavior analysis, and then the driving operation areas are detected by the faster R-CNN detection model for data augmentation. Finally, the convolutional neural network classification mode is implemented and evaluated to detect the original dataset and the driving operation area dataset. The classification result achieves a 96.97% accuracy using the distracted driving dataset. The results show the necessity of driving operation area extraction in the preprocessing stage, which can effectively remove the redundant information in the images to get a higher classification accuracy rate. The method of this research can be used to detect drivers in actual application scenarios to identify dangerous driving behaviors, which helps to give early warning of unsafe driving behaviors and avoid accidents.

Keywords: distracted driving; driving behavior; driving operation area; data augmentation; feature extraction

Citation: Wang, J.; Wu, Z.; Li, F.; Zhang, J. A Data Augmentation Approach to Distracted Driving Detection. *Future Internet* **2021**, *13*, 1. <https://dx.doi.org/10.3390/fi13010001>

Received: 6 November 2020

Accepted: 20 December 2020

Published: 22 December 2020

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to the World Health Organization (WHO) global status report [1], road traffic accidents cause 1.35 million deaths each year. This is nearly 3700 people dying on the world's roads every day. The most heart-breaking statistic is that road traffic injury has become the leading cause of death among people aged 5 to 29 [2]. The investigation [3] for the cause of car collisions shows that 94% of road traffic accidents in the United States are caused by human operations and errors. Among them, distracted driving, which can reduce the driver's reaction speed, is the most dangerous behavior. In 2018 alone, 2841 people died in traffic collisions on United States roads due to driver distraction [4].

The impacts of distracted behavior of drivers are multifaceted [5], including visual behavior, operating behavior, driving pressure, and the ability to perceive danger, etc. According to the definition of the National Highway Traffic Safety Administration (NHTSA) [6], distracted driving refers to any activity that can divert attention away from driving, including (a) talking or texting on a phone, (b) eating and drinking, (c) talking to others in the vehicle, or (d) using radio, entertainment or navigation system.

Distracted driving detection can be used to give early warning of dangerous driving behavior, including using a mobile phone to call or send text messages, using navigation applications or choosing to play music, etc. [7]. Distracted driving detection methods are mainly based on the driver's facial expression, head operation, line of sight or body operation [8]. Through visual tracking, target detection, motion recognition and other technologies, the driver's driving behavior and physiological state can be detected.

In the early days, researchers mainly focused on behavior analysis based on the driver's line-of-sight direction through the eye, face, and head operation. In 2002, Liu et al. [9] proposed a method of tracking the driver's facial area and used the yaw direction angle to estimate the driver's facial operation to detect the driver's facial orientation. Eren et al. [10] successfully developed a driver's facial operation detection system based on binocular stereo vision, using hidden Markov models to predict the driver's facial operation in 2007.

Later, with the development of machine learning technology and the public of driving behavior datasets, increasing studies were added to analyze the driver's phone calling, drinking, eating and other unsafe driving behaviors. Southeast University driving posture (SUE-DP) dataset [11] was proposed in 2011. The experiment collected four types of distracted driving behaviors: "grasping the steering wheel", "operating the shift lever", "eating," and "talking on a cellular phone". A series of studies have been conducted based on the SUE-DP dataset: Zhao [12] used the multiwavelet transform method and the multilayer perceptron (MLP) classifier to recognize four predefined driving postures and obtained an accuracy of 90.61%. Zhao et al. [11] introduced a contourlet transform method for feature extraction and achieved 90.63% classification accuracy. Subsequently, Chihang [13] used support vector machines (SVM) algorithm with an intersection kernel for obtaining 94.25% accuracy. In 2013, The image pyramid histogram of oriented gradients (PHOG) features and edge features [14] were extracted comprehensively to increase the classification accuracy to 94.75% within MLP. In 2014, Yan [15] extracted the PHOG features of historical moving images containing time information and obtained 96.56% accuracy through the random forests (RF) algorithm.

Recently, with the development of deep learning classification and detection technology, increasing researchers analyzed driving behavior through convolutional neural networks (CNNs). More researchers have also begun to build their own research datasets. A combination of pre-trained sparse filters and convolutional neural networks [16] was used to increase the classification accuracy of the SUE-DP dataset to 99.78%. Yan [17] improved the conventional regions with CNNs features (R-CNN) framework by replacing traditional skin-like region extractor algorithms and obtained a classification accuracy of 97.76% on the SUE-DP dataset. Liu [18] used an improved dual-input deep three-dimensional convolutional network structure algorithm based on a three-dimensional convolutional neural network (3DCNN), achieving 98.41% accuracy on the rail transit dataset. Jin [19] trained two independent convolutional neural networks by optimizing the size and number of convolution kernels, which can effectively identify mobile phones and hands in real time achieving 144 fps and an accuracy of 95.7% for mobile phone usage on the self-built dataset. Multiscale attention convolutional neural network [20] was proposed driver action recognition.

More recent datasets and studies include: The StateFarm dataset was published in the 2016 Kaggle distracted driving recognition competition [21] with ten types of distracted driving behaviors. Alotaibi [22] used an improved multiscale Inception model with a classification accuracy of 96.23% on the StateFarm dataset. Lu et al. [23] used the improved deformable and dilated faster RCNN (DD-RCNN) structure to obtain an accuracy of 92.2%. Valeriano [24] and Moslemi [25] used the 3DCNN algorithm to recognize driving behavior and got 96.67% and 94.4% accuracy rates, respectively. In 2018, Eraqi and others [26, 27] proposed the American University in Cairo (AUC) distracted driving dataset with reference to the ten distracted postures defined in the StateFarm dataset. The ASUS ZenPhone close-range camera and DS325 Sony DepthSense camera were used to collect driving images and videos of 44 volunteers from 7 countries. The cameras were fixed to the roof handle on the top of the front passenger's seat. The resolution of the data is 1080×1920 or 640×480 . A total of 17,308 frames were collected for the dataset, which was finally annotated to ten kinds of distracted driving behaviors: safe driving (3686), texting using the right hand (1974), talking on the phone using the right hand (1223), texting using left hand (1301), talking on the phone using left hand (1361), operating the radio (1220), drinking (1612), reaching behind (1159), hair and makeup (1202), and talking

to the passenger (2568). The dataset was randomly divided into 75% for training and 25% for test data. A genetically weighted ensemble of convolutional neural networks combined with the face, hand, and skin regions was proposed to obtain an accuracy of 95.98% with the AUC dataset. Bhakti and others [28] used the AUC dataset to achieve 96.31% accuracy through improved visual geometry group 16 (VGG-16) with regularization methods.

The collection of the dataset mainly used the camera to obtain images of the driver's driving process. During the collection process, it was usually recommended that the driver perform distracting subtasks to simulate distracting driving. The distracted driving methods were mainly based on the driver's facial expression, head operation, line of sight, or body operation for feature extraction. Machine-learning methods and deep learning CNN methods were used for distracting driving recognition. However, the existing datasets and related analysis methods still encounter some problems in the research: on one hand, the current distracted driving research mainly judges driving behavior by the driver's facial and head direction, hand movements, or skin segmentation information. However, the judgment of single local information is prone to classification errors. On the other hand, due to the differences in the resolution, wide-angle, installation position, and installation angle of the camera in different datasets, or the differences in the position of the seat and steering wheel, the position and angle of each driver in the dataset will be different, leading to the images in the dataset have different redundant information.

Two-stage deep architecture methods [29] were usually used in image classification of remotely collected images. A common approach in the literature was employing CNNs for feature extraction to reduce the dimensionality [30]. Data enhancement methods [31] such as flip, rotation, crop, interpolation and color convert [32] were also often used in the first stage of processing to increase robustness. In order to build a more robust driver behavior analysis model and improve the accuracy of dataset classification, this paper designs a data augmentation preprocessing model for driver behavior key areas based on faster R-CNN [33] detection algorithms to improve the accuracy of the algorithm learned from the two-stage depth architecture method. The classification results with data augmentation are verified based on AlexNet [34], InceptionV4 [35] and Xception [36], respectively. To achieve the best performance, transfer learning [37] was applied in training. The American University in Cairo distracted driver (AUC) dataset is used for the experiments.

The main contributions of this paper are summarized in the following three parts:

- (1) First, the class activation mapping method [33] was used to analyze the driving operation key areas in the images.
- (2) Then, in order to enhance the dataset, the image detection algorithm faster R-CNN was used to generate the new driving operation area (DOA) dataset. The driving operation areas were labeled on 2000 images using the AUC dataset to establish the training driving operation areas detection dataset for faster R-CNN training. Within the trained faster R-CNN model, all the AUC dataset images were tested to obtain the preprocessed AUC new DOA classification dataset, which was consistent with the original AUC dataset at the classification storage method and naming method.
- (3) Next, a classification model was built to process the AUC original dataset and the DOA dataset. The experiments were tested with AlexNet, InceptionV4 and Xception separately to get the best result.
- (4) Finally, the trained classification method was used to test our own dataset, which was established with a wide-angle camera different from the close-range camera in the AUC dataset.

The framework of classification model with data augmentation method is shown in Figure 1. Experiments proved that the classification accuracy of the method proposed in this paper is up to 96.97%, which can improve the accuracy of classification.

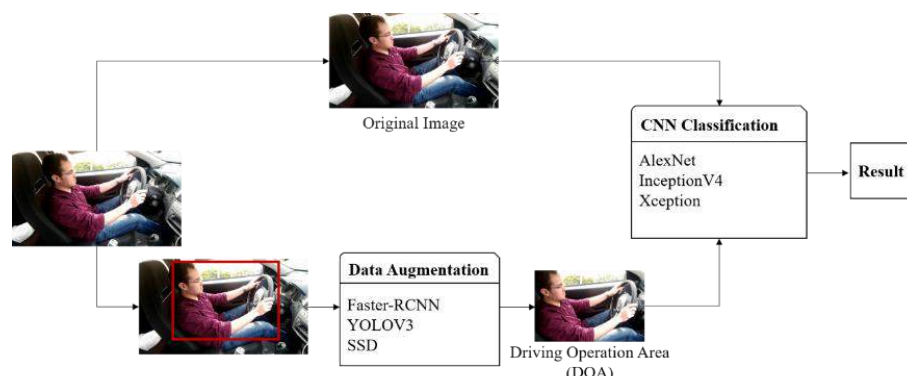


Figure 1. The framework of the classification model with the data augmentation method.

2. Materials and Methods

2.1. Driving Operation Area

In order to effectively observe which areas of the image the network focuses on, this paper used gradient-weighted class activation mapping (grad-CAM) [38] to visually display the features regions found by the Xception classification network, which displayed in the form of a class-specific saliency map or “heat map” for short. Figure 2 shows the grad-CAM result of ten different driving behaviors, which can be used to visually evaluate the key feature regions of the image. The distribution of the ribbon color from red to blue that you can see the mapping relationship between weight and color. The red area in the activation map represents the higher basis area for the model to make classification decisions.

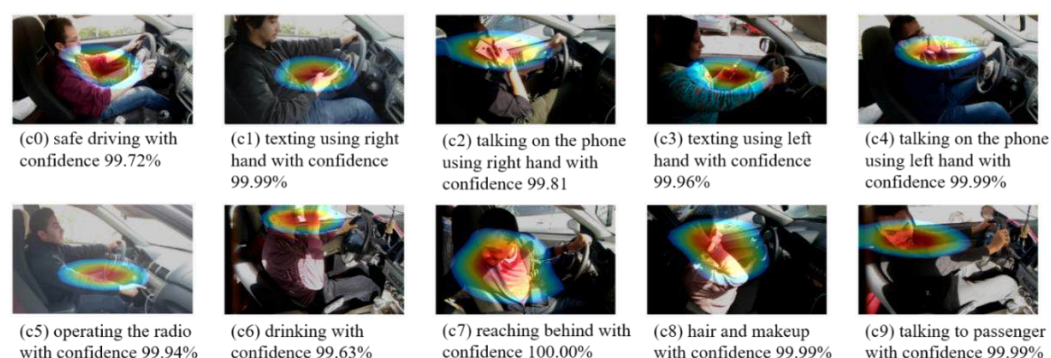


Figure 2. Grad-CAM activation maps of various distracted driving behaviors in the dataset.

According to the grad-CAM result, the driver’s upper body behaviors in the vehicle environment determine the distracted driving classification result, which means that the background and legs that are not related to the driver’s operation are all redundant information in the feature extraction. We proposed the concept of the driving operation area (DOA), including the steering wheel and the driver’s upper body, which include the head, torso, and arms, to describe the features related to the driver’s driving behavior. The area in the red box shown in Figure 1 is what we defined DOA.

2.2. Methods for Data Augmentation

Due to the fixed nature of the distracted driving background, traditional data augmentation methods (such as flipping, rotation, trimming, and interpolation) result in unrealistic scenes, which will cause information distortion and increase irrelevant data. This paper proposed the data augmentation method based on the key area of driving behavior. The mature image detection convolutional network model is used for the data augmentation method. The AUC dataset was enhanced based on the DOA to obtain a new

dataset. Classification modules were introduced to classify the original dataset and the new dataset.

According to the requirements of the image detection model for the dataset, we randomly selected 2000 images from the AUC dataset to relabel the driving operation area using “labellmg” software tool. The labeling area included the steering wheel and the driver’s upper body, including the head, torso, and arms. The annotation file was saved as an XML file in accordance with the Pascal visual object classes (PASCAL VOC) dataset format. The image detection convolutional network model was used to extract the driving operation area.

faster R-CNN [33] was chosen as the image detection model for feature extraction. faster R-CNN creatively used region proposal networks to generate proposals and shared the convolutional network with the target detection network, which can reduce the number of proposals from the original about 2000 to 300 and improve the quality of the suggested frames. The algorithm won many firsts in the ImageNet Large-scale visual recognition competition (ILSVRC) and the common objects in context (COCO) competitions that year, still used frequently by students now.

faster R-CNN model was trained with the labeled data; then, the trained faster R-CNN model was used to infer all the AUC dataset images to obtain the preprocessed new DOA classification dataset. The classification and naming of the DOA dataset were consistent with the original AUC dataset.

2.3. Methods for CNN Classification

This paper used the mature image detection convolutional network model for the data augmentation method. Classic models such as Alexnet [34], InceptionV4 [35], and Xception [36] had been widely used in image classification research in recent years. AlexNet successfully applied rectified linear units (ReLU), dropout and local response normalization (LRN) in CNN. The Inception network started from GoogLeNet in 2014, which had gone through several iterations of versions up to the latest InceptionV4. Xception was another improvement proposed by Google after Inception.

Transfer learning, whose initial weights of each model came from the weights obtained by pre-training on ImageNet, was used in our classification test to train the AUC dataset by optimizing the parameters to get the best result.

The AUC dataset was enhanced based on the DOA to obtain a new dataset. Classification modules were introduced to classify the original dataset and the new dataset. The classification framework with the data augmentation method is shown in Figure 1.

2.4. Wide-Angle Dataset

In order to further verify the generalization ability of our methods, A Wide-angle distracted driving dataset was collected for verification. Referring to the collection methods of the State Farm dataset and the AUC dataset, we fixed the camera to the car roof handle on top of the front passenger’s seat. Fourteen volunteers sat in the car to simulate distracted driving as required in both day and night scenes. Some volunteers participated in more than one collection session at different times of day, driving roads and wearing different clothes. The 360’s G600 recorder, which has a resolution of 1920×1080 and a wide-angle of 139 degrees, was used in the collection. In order to simulate a natural driving scene as much as possible; in some cases, there were other passengers in the car during the collection process.

The data were collected in a video format with the size of 1920×1080 and then cut into individual images. Our dataset finally collected 2200 pictures of ten kinds of distracted driving behaviors: safe driving (291), texting using the right hand (224), talking on the phone using the right hand (236), texting using left hand (218), talking on the phone using left hand (211), operating the radio (203), drinking (198), reaching behind (196), hair and makeup (182), and talking to the passenger (241). Part of the images of the wide-angle dataset is shown in Figure 3.



Figure 3. Part of the images in the wide-angle dataset.

3. Results

The experiments in this article were based on the PaddlePaddle framework and Python design, with the hardware environment using a Linux server with Ubuntu 16.04. A single NVIDIA GeForce GTX, 1080 Ti GPU with 12 GB RAM, was used in the experiments.

3.1. Results for Driving Operation Area Extraction

The labeled driving dataset with 2000 images was split into a training set and a validation set with a ratio of 8:2 for validating the detection model performance. Using the same training strategy as Detection, the dataset was trained with the batch size of 8, the learning rate of 0.001, and the training iterations of 50,000. The momentum 0.9 with a weight decay of 0.0001 for stochastic gradient descent (SGD) was used to converge the model. The Resnet was used for the backbone network. The Resnet weights pre-trained on ImageNet model was used for initialization.

Table 1 is the result of driving operation area extraction with the detection model. Faster R-CNN model was evaluated and compared with the other two models: you only look once (YOLO) [39] and single shot multibox detector (SSD) [40] models. According to the result in Table 1, the accuracy of faster R-CNN detection is 0.6271, and fps is 10.50 which can meet real-time requirements. Considering the accuracy requirements, the faster R-CNN was chosen as the detection model in our experiments. YOLOV3 and SSD models can be used as real-time detection system.

Table 1. Driving operation area detection result.

Model	fps	mAP (0.75)	mAR (0.75)
faster R-CNN	10.50	0.6271	0.6572
YOLOV3	37.21	0.5390	0.5568
SSD	49.52	0.5767	0.5812

Figure 4 shows ten different types of driving operation area detection results. Comparing Figures 2 and 4, the key regions are extracted by the image detection model.



Figure 4. Ten different types of driving operation area (DOA) detection results.

Then the trained weights of faster R-CNN were used to detect the key areas of driving behavior in the AUC dataset, and generate a dataset of driving operation area, which was recorded as the DOA dataset, which classification and naming methods were the same as the original AUC dataset.

3.2. Results for CNN Classification

In the experiment, the dataset of AUC and DOA were both 12,997 images of training set and 4331 images of test set. The image classification model AlexNet, InceptionV4 and Xception were used to train with image shape of $224 \times 224 \times 3$, the learning rate of 0.001, batch size of 32, and epoch of 100. The top-1 accuracy was selected to evaluate the performance of the models. We performed 3 rounds of verification. Table 2 summarizes the test results for loss and accuracy of three different convolutional network models: AlexNet, InceptionV4, and Xception.

Table 2. Image classification test result with AUC and DOA dataset.

Model	Source	Loss	Top-1 Acc.
AlexNet	AUC	0.3753	0.9314
	DOA	0.3402	0.9386
InceptionV4	AUC	0.3041	0.9506
	DOA	0.2771	0.9572
Xception	AUC	0.2320	0.9531
	DOA	0.2156	0.9655

As can be seen from Table 2, the test top-1 accuracy of the AlexNet, InceptionV4 and Xception on the AUC dataset are 0.9314, 0.9506 and 0.9531, respectively, and the test results on the DOA dataset are 0.9386, 0.9572 and 0.9655, which means the DOA dataset has higher detection accuracy and lower loss value than the original AUC dataset.

Figure 5 shows the change for loss and accuracy of each method in each epoch stage. When the epoch is 10, the loss and accuracy of the DOA dataset with the Xception model begin to stabilize, and when the epoch is 14, the original AUC dataset loss and accuracy with the Xception model begin to stabilize. Moreover, The loss values in the DOA-based results are lower than original AUC dataset. It can be seen from the testing loss and accuracy curves with varying epochs, the loss of DOA dataset corresponding to the key areas of driving behavior converges faster than the original AUC dataset, and the detection accuracy rises faster too.

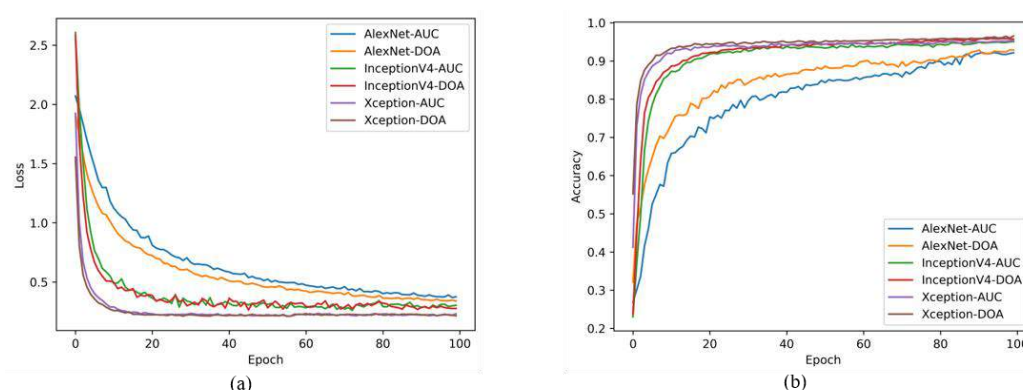


Figure 5. Accuracy and loss of image classification result. (a) the loss of each method in each epoch stage; (b) the accuracy of each method in each epoch stage.

Finally, the DOA training set obtained through data augmentation and the original AUC training set were merged to expand the dataset. The final classification accuracy is shown in Table 3. Among the three classification models, the baseline with Xception has the smallest fluctuation, the lowest loss result, and the highest accuracy, which is the most suitable for the benchmark model of this classification. For more evaluation, Figure 6 is the confusion matrix for the classification results of the ten distracting behaviors with Xception. Using the given confusion matrix, one can check that many categories can

easily be mistaken for (c0) “safe driving”. Moreover, the most confusing operation is (c8) “hair and makeup”. It may be due to the position of “hands on the wheel” in both classes.

Table 3. Final result after confidence comparison.

Model	Top-1 Acc.
AlexNet	0.9396
InceptionV4	0.9603
Xception	0.9697

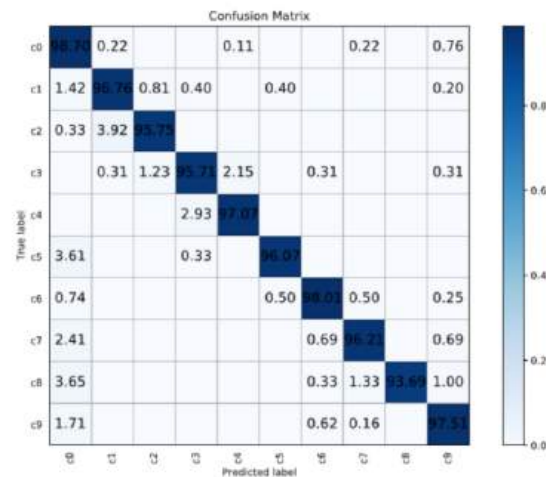


Figure 6. Confusion matrix of results Xception classification.

Our distracted driver detection result was compared with earlier methods in the literature. Compared with some early methods, our method can be applied to the preprocessing stage. We achieve the best accuracy than earlier methods as shown in Table 4. Among them, the top-1 accuracy of our module based on Xception is finally 0.9697, which is 1.66% higher than the classification accuracy of origin AUC dataset.

Table 4. Comparison with earlier methods from literature on AUC dataset.

Model	Top-1 Acc.
GA weighted ensemble of all 5 [26]	0.9598
VGG [28]	0.9444
VGG with regularization [28]	0.9631
ResNet + HRNN + modified Inception [22]	0.9236
Our method	0.9697

3.3. Tests on Wide-Angle Dataset

Due to the high correlation between the training and test data of the AUC dataset, this makes the detection of driving distraction an easier problem. Therefore, the newly collected wide-angle dataset was used to verify the generalization ability of our method. The wide-angle dataset contains 14 drivers (2200 samples). The wide-angle dataset was used to verify the feasibility of our proposed method, especially for datasets with a relatively small proportion of drivers. The trained model on the AUC dataset was used in the verification for the wide-angle dataset directly. Referring to the performance of the previous experiment with Xception-based model, this paper used the Xception-based model to verify the generalization ability.

Table 5 shows the verification result of the dataset captured by the wide-angle camera. The classification top-1 accuracy of the model is greater than 80%, which verifies a relatively good generalization ability. In addition, the classification results after extracting

the key areas of the driver operation are significantly better than the original data classification results. It proves the necessity of extracting key areas of drivers in distracted driving detection.

Table 5. Result on wide-angle dataset.

Model	Source	Top-1 Acc.
Xception	Wide-angle Dataset	0.8131
	DOA of Wide-angle Dataset	0.8394

4. Discussion

In practical applications, due to the difference in the installation position and resolution of the camera, and the difference in the position of the driver's seat and steering wheel, the driver's distribution position and angle in the image will be different. The difference in the proportion of the driver's operating area in the image will cause many pixels in the image of the collected dataset to be redundant information. This article focuses on improving the robustness and accuracy of distracted driving detection.

First, with the labeled data, faster R-CNN was used to detect the key areas of driving behavior. The extraction of DOA was a large target detection for CNN, and the general faster R-CNN has been able to achieve good accuracy. It can be seen from the experimental results that this method can extract key information and can be used in the first stage of distracted driving detection. Comparing with grad-CAM activation maps, it can be seen that our method was especially helpful for driving behavior detection in complex backgrounds.

Second, the convolutional neural network classification model was used to test the loss and accuracy of the AUC dataset and the DOA dataset. It can be seen from the result that the DOA dataset has higher detection accuracy and lower loss value than the original AUC dataset. Testing with the combined dataset of AUC and DOA, the experiment got a 96.97% top-1 accuracy. Compared with some early methods in the literature, our method can extract the overall characteristics of key areas of driving behavior. The loss of InceptionV4 and Xception dropped to a better result when the epoch was 4, and reached relatively stable when the epoch reached 40. The results showed the effectiveness of transfer learning for CNN models.

Third, The wide-angle dataset collected by actual scene was used to verify our method. Our results demonstrated that detect the key areas of driving behavior has a great significance for driving behavior analysis of wide-angle camera shooting and long-range shooting.

It can find that if the extracted features come from the entire image, which means all the information in the image (regardless of whether it is related to driving behavior) are used as a training input, the result will lead to more redundant information and larger calculation. Considering the diversity of the driver's position and the complexity of the cab environment, our method is suitable for practical application fields.

5. Conclusions

Distracted driving detection has become a major research in transportation safety due to the increasing use of infotainment components in vehicles. This paper proposed a data augmentation method for driving position area with the faster R-CNN module. The convolutional neural network classification model was used to identify ten distracting behaviors in the AUC dataset, reaching the top-1 accuracy of 96.97%. Extensive results carried out show that our method improves the accuracy of the classification and has strong generalization ability. The experimental results also showed that the proposed method was able to extract key information. This provided a path for the preprocessing stage of driving behavior analysis.

In the future, the following aspects can be continued for further research:

First, more distracted driving datasets with multi-angle and night scenarios should be collected and published for more comprehensive research. We need to verify our model on more practical large-scale datasets.

Second, the current classification algorithm divides dangerous driving behaviors into multiple categories, but in actual driving behaviors, multiple dangerous behaviors may co-exist, such as watching around when making a call. We can use detection modes such as YOLO (or any other object detector) to detect the face, hand, and other information on the basis of the work of DOA for more driving behavior identification.

Author Contributions: Conceptualization, J.W. and J.Z.; methodology, J.W.; software, J.W.; validation, J.W.; resources, J.W.; writing—original draft preparation, J.W.; writing—review and editing, F.L.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This project is supported by the Internet of Vehicles Shared Data Center and Operation Management Cloud Service Platform of Anhui Province.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. *Global Status Report on Road Safety 2018: Summary*; World Health Organization: Geneva, Switzerland, 2018.
2. Peden, M. Global collaboration on road traffic injury prevention. *Int. J. Inj. Control Saf. Promot.* **2005**, *12*, 85–91. [CrossRef]
3. Singh, S. *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*; National Highway Traffic Safety Administration: Washington, DC, USA, 2015.
4. Vasilash, G.S. Distraction and Risk. *Automot. Des. Prod.* **2018**, *130*, 6.
5. Kaber, D.B.; Liang, Y.; Zhang, Y.; Rogers, M.L.; Gangakhedkar, S. Driver performance effects of simultaneous visual and cognitive distraction and adaptation behavior. *Transp. Res. Part F-Traffic Psychol. Behav.* **2012**, *15*, 491–501. [CrossRef]
6. Strickland, D. How Autonomous Vehicles Will Shape the Future of Surface Transportation. 2013. Available online: <https://www.govinfo.gov/content/pkg/CHRG-113hhrg85609/pdf/CHRG-113hhrg85609.pdf> (accessed on 21 December 2020).
7. Liu, D. Driver status monitoring and early warning system based on multi-sensor fusion. In Proceedings of the 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Vientiane, Laos, 11–12 January 2020. [CrossRef]
8. Yanfei, L.; Yu, Z.; Junsong, L.; Jing, S.; Feng, F.; Jiangsheng, G. Towards Early Status Warning for Driver's Fatigue Based on Cognitive Behavior Models. In Proceedings of the Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management: 4th International Conference, DHM 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, 21–26 July 2013. [CrossRef]
9. Liu, X.; Zhu, Y.D.; Fujimura, K. Real-time pose classification for driver monitoring. In Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, Singapore, 3–6 September 2002; pp. 174–178.
10. Eren, H.; Celik, U.; Poyraz, M. Stereo vision and statistical based behaviour prediction of driver. In Proceedings of the 2007 IEEE Intelligent Vehicles Symposium, Istanbul, Turkey, 13–15 June 2007; pp. 657–662.
11. Zhao, C.H.; Zhang, B.L.; He, J.; Lian, J. Recognition of driving postures by contourlet transform and random forests. *IET Intell. Transp. Syst.* **2012**, *6*, 161–168. [CrossRef]
12. Zhao, C.; Gao, Y.; He, J.; Lian, J. Recognition of driving postures by multiwavelet transform and multilayer perceptron classifier. *Eng. Appl. Artif. Intell.* **2012**, *25*, 1677–1686. [CrossRef]
13. Chihang, Z.; Bailing, Z.; Jie, L.; Jie, H.; Tao, L.; Xiaoxiao, Z. Classification of Driving Postures by Support Vector Machines. In Proceedings of the 2011 Sixth International Conference on Image and Graphics, Hefei, China, 12–15 August 2011; pp. 926–930. [CrossRef]
14. Zhao, C.H.; Zhang, B.L.; Zhang, X.Z.; Zhao, S.Q.; Li, H.X. Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers. *Neural Comput. Appl.* **2013**, *22*, S175–S184. [CrossRef]
15. Yan, C.; Coenen, F.; Zhang, B.L. Driving Posture Recognition by Joint Application of Motion History Image and Pyramid histogram of Oriented Gradients. *Int. J. Veh. Technol.* **2014**, 846–847. [CrossRef]
16. Yan, C.; Zhang, B.; Coenen, F. Driving Posture Recognition by Convolutional Neural Networks. In Proceedings of the 2015 11th International Conference on Natural Computation (Icnc), Zhangjiajie, China, 15–17 August 2015; pp. 680–685.
17. Yan, S.; Teng, Y.; Smith, J.S.; Zhang, B. Driver Behavior Recognition Based on Deep Convolutional Neural Networks. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (Icnc-Fskd), Changsha, China, 13–15 August 2016; pp. 636–641.
18. Liu, Y.Q.; Zhang, T.; Li, Z. 3DCNN-Based Real-Time Driver Fatigue Behavior Detection in Urban Rail Transit. *IEEE Access* **2019**, *7*, 144648–144662. [CrossRef]

19. Jin, C.C.; Zhu, Z.J.; Bai, Y.Q.; Jiang, G.Y.; He, A.Q. A Deep-Learning-Based Scheme for Detecting Driver Cell-Phone Use. *IEEE Access* **2020**, *8*, 18580–18589. [CrossRef]
20. Hu, Y.C.; Lu, M.Q.; Lu, X.B. Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network. *Signal Process. Image Commun.* **2020**, *81*. [CrossRef]
21. Kaggle. State Farm Distracted Driver Detection. Available online: <https://www.kaggle.com/c/state-farm-distracted-driver-detection/data> (accessed on 21 December 2020).
22. Alotaibi, M.; Alotaibi, B. Distracted driver classification using deep learning. *Signal Image Video Process.* **2019**. [CrossRef]
23. Lu, M.Q.; Hu, Y.C.; Lu, X.B. Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. *Appl. Intell.* **2020**, *50*, 1100–1111. [CrossRef]
24. Valeriano, L.C.; Napoletano, P.; Schettini, R. Recognition of driver distractions using deep learning. In Proceedings of the 2018 IEEE 8th International Conference on Consumer Electronics, Berlin, Germany, 2–5 September 2018.
25. Moslemi, N.; Azmi, R.; Soryani, M. Driver Distraction Recognition using 3D Convolutional Neural Networks. In Proceedings of the 2019 4th International Conference on Pattern Recognition and Image Analysis, Tehran, Iran, 6–7 March 2019; pp. 145–151. [CrossRef]
26. Eraqi, H.M.; Abouelnaga, Y.; Saad, M.H.; Moustafa, M.N. Driver Distraction Identification with an Ensemble of Convolutional Neural Networks. *J. Adv. Transp.* **2019**. [CrossRef]
27. Abouelnaga, Y.; Eraqi, H.M.; Moustafa, M.N. Real-time Distracted Driver Posture Classification. *arXiv* **2017**, arXiv:abs/1706.09498.
28. Baheti, B.; Gajre, S.; Talbar, S.; IEEE. Detection of Distracted Driver using Convolutional Neural Network. In Proceedings of the 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, Utah, USA, 18–22 June 2018; pp. 1145–1151. [CrossRef]
29. Petrovska, B.; Zdravevski, E.; Lameski, P.; Corizzo, R.; Štajduhar, I.; Lerga, J. Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification. *Sensors* **2020**, *20*, 3906. [CrossRef]
30. Petrovska, B.; Atanasova-Pacemska, T.; Corizzo, R.; Mignone, P.; Lameski, P.; Zdravevski, E. Aerial scene classification through fine-tuning with adaptive learning rates and label smoothing. *Appl. Sci.* **2020**, *10*, 5792. [CrossRef]
31. Zhao, Z.; Luo, Z.; Li, J.; Chen, C.; Piao, Y. When Self-Supervised Learning Meets Scene Classification: Remote Sensing Scene Classification Based on A Multitask Learning Framework. *Remote Sens.* **2020**, *12*, 3276. [CrossRef]
32. Izadpanahkakhk, M.; Razavi, S.M.; Taghipour-Gorjilaie, M.; Zahiri, S.H.; Uncini, A. Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning. *Appl. Sci.* **2018**, *8*, 1210. [CrossRef]
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
35. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First Aaai Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
36. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
37. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
38. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
39. Redmon, J.; Farhadi, A. Yolo v3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
40. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.

Article

Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce

Laith T. Khrais 

Department of Business Administration, College of Applied Studies and Community Service,
Imam Abdulrahman Bin Faisal University, Dammam 34212, Saudi Arabia; lakhris@iau.edu.sa

Received: 21 October 2020; Accepted: 3 December 2020; Published: 8 December 2020

Abstract: The advent and incorporation of technology in businesses have reformed operations across industries. Notably, major technical shifts in e-commerce aim to influence customer behavior in favor of some products and brands. Artificial intelligence (AI) comes on board as an essential innovative tool for personalization and customizing products to meet specific demands. This research finds that, despite the contribution of AI systems in e-commerce, its ethical soundness is a contentious issue, especially regarding the concept of explainability. The study adopted the use of word cloud analysis, voyance analysis, and concordance analysis to gain a detailed understanding of the idea of explainability as has been utilized by researchers in the context of AI. Motivated by a corpus analysis, this research lays the groundwork for a uniform front, thus contributing to a scientific breakthrough that seeks to formulate Explainable Artificial Intelligence (XAI) models. XAI is a machine learning field that inspects and tries to understand the models and steps involved in how the black box decisions of AI systems are made; it provides insights into the decision points, variables, and data used to make a recommendation. This study suggested that, to deploy explainable XAI systems, ML models should be improved, making them interpretable and comprehensible.

Keywords: artificial intelligence; automation; e-commerce; machine learning; big data; customer relationship management (CRM)

1. Introduction

Technological advancement continues to create new opportunities for people across a variety of industries [1]. Technology helps to improve the efficiency, quality, and cost-effectiveness of the services provided by businesses. However, technological advancements can be disruptive when they make conventional technologies obsolete. Neha et al. [1] assert that cloud computing, blockchain, and AI are the current developments that may create new opportunities for entrepreneurs. The computer systems are also influencing and improving interactions between consumers and business organizations. Thus, the shift towards the improved use of technology has led to the creation of intelligent systems that can manage and monitor business models with reduced human involvement. AI systems that demonstrate an ability to meet consumers' demands in different sectors are necessary for the current economy [2]. AI plays a critical role in monitoring the business environment, identifying the customers' needs, and implementing the necessary strategies without or with minimal human intervention. Thus, it bridges the gap between consumers' needs and effective or quality services.

Therefore, AI is modifying the economic landscape and creating changes that can help consumers and entrepreneurs to reap maximum benefits. It is gaining popularity in businesses, especially in business administration, marketing, and financial management [3]. AI creates new opportunities that result in notable transformations in the overall economic systems. For instance, it causes the rapid unveiling of big data patterns and improved product design to meet customers' specifications and

preferences [1]. E-commerce is the major beneficiary of the increased use of AI to improve services' efficiency and quality.

AI helps in reducing complications that may result from human errors. Thus, although AI may reduce employment opportunities, its benefits to organizations are immense.

Notably, AI is a formidable driving force behind the development and success of e-commerce. In e-commerce, AI systems allow for network marketing, electronic payments, and management of the logistics involved in availing products to the customers. Di Vaio et al. [3] note that AI is becoming increasingly vital in e-commerce food companies because it maintains the production sites' hygienic conditions and ensures safe food production. It also helps in maintaining high levels of cleanliness of the food-producing equipment.

The automated systems collect, evaluate, and assess data at a rapid rate compared to human beings. AI helps e-commerce to capture the business trends and the changing market needs of customers. Therefore, the customers' increased convenience leads to increased satisfaction and balancing of the demand and supply mechanisms.

Kumar and Trakru [4] report that AI allows e-commerce to develop new ideas on satisfying the consumers' needs and keep up with the changing preferences and choices. Human intelligence may often be limited in carrying out some tasks in e-commerce, including predicting demand and supply chain mechanisms. AI simulates and extends human intelligence to address the rising challenges in e-commerce [5]. For instance, Soni [6] established that AI helps e-commerce platforms to manage and monitor their customers. Through AI, a business can gather a wide range of information and evaluate customers to ensure that quality services are offered to them. This helps e-commerce platforms understand the factors that influence their current and potential clients' purchasing behaviors. It improves the interactions between the e-commerce companies and their customers through chatbots and messengers. E-commerce companies use automation processes to eliminate redundancies in their operations. Kitchens et al. [7] state that AI allows for automated responses to questions raised by the customers. However, Kumar and Trakru [4] warn of potential threats and challenges to customers and e-commerce that limit the efficiency and effectiveness of AI in meeting the business expectations. Consequently, it is necessary to explore opportunities and challenges in light of changing consumer demands in e-commerce.

1.1. Statement of the Problem

Even though AI systems have revolutionized e-commerce, courtesy of the wide range of functionalities such as video, image and speech recognition, natural language, and autonomous objects, a range of ethical concerns have been raised over the design, development, and deployment of AI systems. Four critical aspects come into play: fairness, auditability, interpretability and explainability, and transparency. The estimation process via such systems is considered the 'black box' as it is less interpretable and comprehensible than other statistical models [8]. Bathaee [9] identifies that there are no profound ways of understanding the decision-making processes of AI systems. The black box concept implies that AI predictions and decisions are similar to those of humans. However, they fail to explain why such decisions have been made. While AI processes may be based on perceivable human patterns, it can be imagined that understanding them is similar to trying to determine other highly intelligent species. Because little can be inferred about the conduct and intent of such species (AI systems), especially regarding their behavior patterns, issues of the degree of liability and intent of harm also come into play. The bottom line revolves around how best to guarantee transparency to redeem trust among the targeted users within e-commerce spaces [9].

1.2. Proposed Solution

The main aim of the current study is to lay the foundation for a universal definition of the term explainability. AI systems need to adopt post hoc machine learning methods that will make them more interpretable [10]. While there is a wide range of taxonomic proposals on the definition and

classification of what should be considered ‘explainable’ in regards to AI systems [10], it is contended that there is a need for a uniform blueprint of the definition of the term and its components. To address the main objective, which is solving the black box problem, this research proposes the employment of XAI models. XAI models feature some level of explainability approached from various angles, including interpretability and transparency. Even though the current research recognizes the prevalence of studies under the topic and consults widely within the area, it goes a step ahead to offer a solution to the existing impasse, as seen in divided scientific contributions concerning what constitutes the concept of ‘explainability’. In this regard, the article will contribute to the state of the art by establishing a more uniform approach towards research that seeks to create evidence based XAI models that will address ethical concerns and enhance business practices.

1.3. Overview of the Study

The current article provides a critical outline of key tenets of AI and its role in e-commerce. The project is structured into six main sections, which are the introduction, review of the literature, proposed method, results, discussion, and conclusion. The introduction part of the project provides an overview of the research topic ‘role of AI in shaping consumer in E-commerce,’ a statement of the problem and the proposed solution. Section 2 examines the available literature on artificial intelligence techniques, including sentimental analysis and opinion mining, deep learning, and machine learning. It also examines the AI perspective in the context of shaping the marketing strategies that have been adopted by businesses. Section 3 discusses the methodology used to explore the research question. It identifies the research approach (word cloud analysis), sources of data (Neural Information Processing Systems (NIPS), and Cognitive Science Society (CSS)). Section 3 also details data analysis techniques, which include corpus analysis and concordance analysis. The results section provides the outcomes of the analyzed data, particularly the corpus generated from the word cloud. Some of the software, such as the Voyant Tools, was used to reveal the most prominent words. A concordance analysis table was also generated, in the results section, for the term ‘explainability’. Section 5 provides a detailed discussion of the outcome of the research as well as the key observations made regarding opaque systems, interpretable systems, and comprehensible systems. The overall outcome of the study and implications for future research are provided in the conclusion section.

2. Artificial Intelligence Techniques

2.1. Sentimental Analysis and Opinion Mining

Communication between a business organization and customers occurs through user-generated content on websites and social media. The views of clients expressed on Twitter, Instagram, and Facebook can impact the service providers’ image and reputation. The sentimental analysis involves methodologies that help companies evaluate the meaning of online content and develop strategies that can improve customer loyalty [7]. Thus, sentimental analysis refers to an automation process that helps customers to extract emotions from customers’ online content by processing unstructured data and creating models to extract information from it [11,12]. It helps managers or decision-makers to understand customers’ responses towards particular topics, and it helps to determine if the event is neutral, positive, or negative. There is a minor difference between sentimental analysis and opinion planning. Opinion planning focuses on extraction and analyzing customers’ opinions while searching for clients’ expressions or words and analyzing them. Opinion mining involves using predefined rules to establish rules based on automated systems that analyze data [13]. It helps in the classification and investigation of customers’ behavior and attitudes towards an event, a brand, company services, and products. Twitter is one platform that contains relevant information and hashtags that have been followed by large numbers of people [14].

Sentimental analysis or opinion mining classifies customer emotions into three levels: the sentence level, document level, and entity level.

Natural language processing (NLP) is a fundamental technique for promoting sentimental analysis. It is a sub-domain in AI that focuses on understanding the unstructured content found on social media platforms and organizing it to make it easier for sentimental analysis [13]. It helps the computer-aided system to assess and understand the various languages spoken by customers. NLP is also a key component in opinion mining [13]. It helps people process a large amount of information from unstructured data by analyzing the sentence structures and computing it into the sentence or document level using linguistic databases such as treebanks and WordNet [14]. Besides, sentiment polarity from user-generated texts can be classified using various approaches: a semantic approach, machine learning, lexical-based analysis, and a statistical approach [14]. These approaches help to eliminate unnecessary noise or information such as URLs, slang, and abbreviations, thus, they decrease the size of the dataset. Therefore, through sentimental analysis and opinion mining, customers can understand the type of shopping experience they are likely to have in a particular enterprise based on customers' views or opinions on social media platforms.

2.2. Related Work

2.2.1. Deep Learning

The rapid increase in cross-border e-commerce (CBEC) and demand for international logistics have rendered third-party logistics (3PL) obsolete because they can meet the current requirements of CBEC [12]. The random arrival of goods or products creates a big challenge. The deep-learning method avoids these challenges through the optimization and demand forecasting process [12]. It helps e-commerce companies identify the best advertisements to use, understand customers' intentions, and optimize products' delivery. Deep learning also supports decision making in a logistics service capacity.

2.2.2. Machine Learning

Machine learning techniques use machine learning algorithms to control data processing by classifying linguistic data and presenting it in vector form [14]. The machine learning techniques help e-commerce businesses to detect anomalies in the products and prices. They also help the companies track assets in warehouses and improve their pages or websites' rankings. Machine learning algorithms help e-commerce companies to learn from the derived data and create solutions to challenges they may be experiencing. Thus, machine learning models are used to solve various economic issues.

2.3. AI Perspective

Davenport et al. [15] assert that AI has great potential in shaping the marketing strategies adopted by businesses. AI redefines business models and sales processes to address the changing macro-business environment. The impact of AI is substantial in customer behavior. Soni et al. [5] argue that the changes brought about by AI in e-commerce aim at enhancing customer experience through improved systems that bridge the gap between the business and consumers. Companies in e-commerce face the challenge of addressing the continually changing customer expectations. Businesses are crafting ways to remain competitive by going beyond merely delivering what customers demand. Soni [5] postulates that the appropriateness of the time and channels of delivery are central in customer service. A business that seems to be eliminating the customer's pain associated with purchasing processes will attract more consumers by expanding its market share. Such a business relies on machine learning to develop a model that enhances efficiency based on available information regarding the market and nature of the competition.

Marketing is a crucial component of the business as it convinces customers to prefer some products from individual enterprises to others. Tussyadiah and Miller [16] observe that AI has great potential in redefining marketing in the present and future. AI comes with new and dynamic marketing strategies with entirely new and improved ways of reaching out to customers. AI allows for attitudinal

segmentation making marketing strategies more sustainable. Additionally, AI ensures a comprehensive approach to understanding customer behavior hence tailoring marketing to win over individual buyers in e-commerce [17]. For instance, AI, as applied in marketing, supports changing user interfaces across platforms accessed by the customers, hence increasing their likelihood to purchase a given product. The hyper-personalization of marketing supported by AI is one critical aspect that influences customer's demand in e-commerce. Davenport et al. [15] stress the association between marketing and customer's purchasing behavior. On the contrary, Kachamas et al. [18] warn against a blanket belief that AI in marketing leads to increased consumer demand. Instead, the authors point out the complexities associated with AI in marketing. Customers remain exposed to malicious activities that may lead to a breach of information, thereby harming them. On the other hand, AI is quickly evolving, and businesses' strategies in e-commerce to attract customers become obsolete fast. The use of AI in marketing may mislead marketers, leading to unrealistic expectations. The study points out that AI comes with the concepts of perceived usefulness (PU) and perceived ease of use (PEOU), which may not necessarily be the case. Therefore, it is necessary to note the issues in using macro-levels of business operations before rolling out the market operations [16]. Adadi and Berrada [19] argue that, with the advent of AI, society is shifting towards a more algorithmic system characterized by the wide use of ML techniques. With this unprecedented advancement, an essential aspect of AI systems is that they have been deemed to lack transparency. Indeed, the BlackBox nature of Machine Learning (ML) systems allows unexplainable but powerful predictions to occur. When AI systems make predictions without explanations, they are bound to cause difficulties when there is a need to detect inappropriate decisions. Thus, organizations are exposed to biased data, incorrect decision making, and unsuitable modeling techniques in the algorithm lifecycle.

As AI systems are getting powerful, sophisticated and pervasive, the techniques for troubleshooting and monitoring them lag behind their adoption, leading to many risks that affect consumers. Marketing measurements are still one-dimensional. AI analytics continue to run statistical methods that are single-sided. They are based on unidirectional feedback and responses such as click-through purchase, customer responses, etc. While such traditional metrics are useful for gauging customer behavior, they are static and slow. Additionally, they do not reveal the real picture of consumers—that is, they are shallow. For example, in most cases, drilling down into consumer behavior is usually done through segmentation and clustering, which tends to be descriptive and outdated. While such analytics tend to reveal interesting consumer findings, the outcomes are always non-dynamic and contextually unsafe, and inappropriate for use to inspire actionable insight to influence the customer's daily activities and how they interact with the business. Arrieta et al. [10] echo similar findings; the dangers of using black-box systems are that they elicit unjustifiable decisions or those that do not allow the acquisition of detailed explanations of the target customer's behavioral patterns. This becomes an issue in a world where customers expect relevance, since they are not static in their thinking, interests, and needs.

Even though AI systems are highly precise, if they cannot explain why a given customer responds the way they did, it becomes difficult to gain insight that can generate contextually significant action. XAI presents as a solution to the problem of static AI data. Notably, with fully explainable AI, businesses can conduct dynamic analytics generating contextually relevant data. XAI is concerned with the issues revolving around trustworthiness. XAI mostly revolves around the evaluation of the moral and ethical standards of an ML. Thus XAI seeks to overcome the dangerous practice of “accepting” outcomes blindly, whether by necessity or by choice. Therefore, by implementing XAI models, analysts provide an output that can be understood to support rational human decision-making.

Additionally, analysts can easily assess decisions made by clients and users of AI systems to determine if they are rational—that is, if the decisions incorporate reasoning and do not conflict with legal or ethical norms. As echoed in the chorus voices across ML studies, there is an increased demand for ethical AI, thus pointing to the need for XAI models. However, as argued by Arrieta et al. [10], there is a lack of a novel framework that brings together previous works in the field as well as covering conceptual

propositions with a unified focus on the users of XAI. Thus, this research proposes a comprehensive framework for defining what comprises XAI, paving the way for large-scale implementation.

3. Proposed Method

Figure 1 below is an overview of the proposed method used in the study. In the research approach, the researcher utilized word cloud analysis. Data were collected from two databases, namely Cognitive Science Society and Neural Information Processing Systems. Finally, the collected data were analyzed through normalization of the frequency of the ‘explainability’ term, Voyant analysis, and concordance analysis.

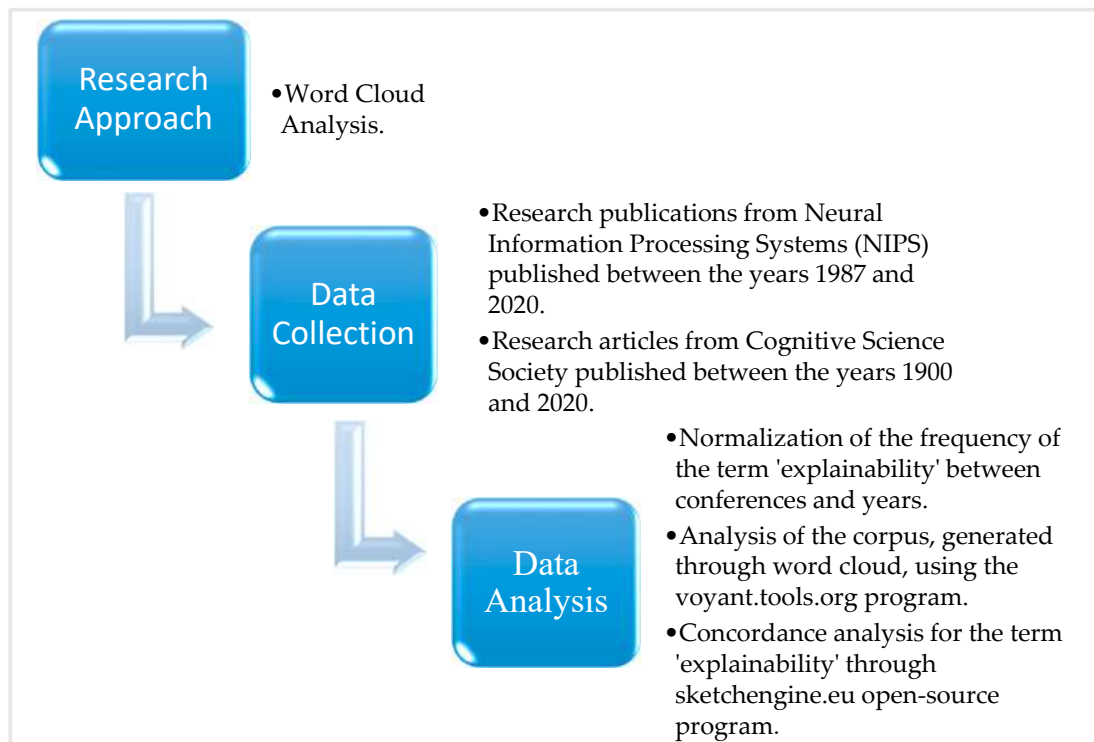


Figure 1. A flowchart of the proposed method.

3.1. Research Approach

This research utilized a word cloud analysis to investigate the research problem and question. A word cloud analysis is a research approach where a visual representation of a word is generated based on its frequency [20]. In this approach, the image of the word generated becomes larger the more the term appears within the analyzed text. This research approach is appropriate for this study, given that the main purpose was to investigate the significance of the term ‘explainability’ within AI research databases and determine how the concept has been defined. According to Doran et al. [21], terms such as interpretability have been widely applied in research publications despite unclear definitions. Thus, through the word cloud analysis, the researcher was able to assess relevant terms in AI databases whose researchers rely on machine-learning-driven techniques to approach their study objectives. Therefore, the method was effective in identifying the focus of the written materials being assessed by highlighting the frequency of word use for a basic understanding of data as described by McNaught and Lam [22].

Additionally, compared to other methods of textual analysis, the word cloud analysis was the most appropriate for this study. For instance, in comparing word clouds to the user interface with a search box, Sinclair and Cardew-Hall [23] found that participants tended to prefer the search box to fill specific terms; however, they favored word clouds more when it came to dealing with open-ended

tasks. Kuo et al. [24] found similar results and argued that word clouds are an essential technique for giving impressions of the information in a query list. The overall conclusion is that word clouds are a useful visualization tool for communicating the overall textual picture [20].

3.2. Data Collection

Data used in this study were the linguistic corpora obtained from conference proceedings whereby research articles and peer-reviewed papers were evaluated. Data were gathered from the Neural Information Processing Systems (NIPS), with the target being a range of research publications from 1987 to 2020. Additionally, research papers from the Cognitive Science Society from between 1900 and 2020 (accessed via an open-source managed by the University of California) were obtained for analysis. These two ML communities were chosen for their richness in AI-based information [21]. Moreover, the wide information available makes it possible to analyze the trends in using the term explainability and the related concepts.

3.3. Data Analysis

The data analysis involved conducting a shallow assessment for the term ‘explainability’. The frequencies were normalized between conferences and years, after which the plots in Figure 2a,b were derived. Figure 2a is a frequency plot for the term explainability obtained from the Cognitive Science Society database. Figure 2b is the frequency plot obtained from the Neural Information Processing Systems (NIPS) database searches. The voyant.tools.org program was used to analyze the corpus, generated through the word cloud analysis, to establish a connection between the words and, consequently, generate meaning. The researcher also utilized the sketchengine.eu open-source program to generate the concordance of the term ‘explainability’.

4. Results

The word cloud plots (Figure 2a,b) are an easy way of understanding the composition of the ‘explainability’ concept and the related semantic meaning across the ML-driven databases chosen for this study. Here, essential words were perceived as those that first appeared in a 20-word window following a search of the term ‘explainability’; such words also had a frequency above the average level. In Figure 2a, it is evident that the corpus reveals the prominent words as use, explanation, model, and emotion. Other notable words are learn and the system. The corpus (in Figure 2b) shows well-known words as model, learn, use, and data. Other prominent words are method, task, infer, and image. There are also essential words such as decision and prediction appearing within the 20-word window. Thus, the AI community in Figure 2a describes the term explainability as being related to words such as use, explanation, and model, among others, suggesting an emphasis on using system models that enhance learning, explanation, decision making, and prediction. The implication (in Figure 2b) is that the term explainability could be taken to mean using models that allow learning, ease of inference, and prediction, among others.

The corpus generated from the word cloud (in Figure 2a) reveals that the 20 most prominent words are as follows: system, explanation, model, emotion, learn, use, study, train, method, predict, data, estimate, control, result, show, decision, interpret, design, variance, image). Analyzing this corpus to formulate a connection between the words and generate meaning using voyant-tools.org (an open-source corpus analysis program) revealed that the most prominent words, as seen in Figure 3, were as follows: control; data; decision; design; emotion.

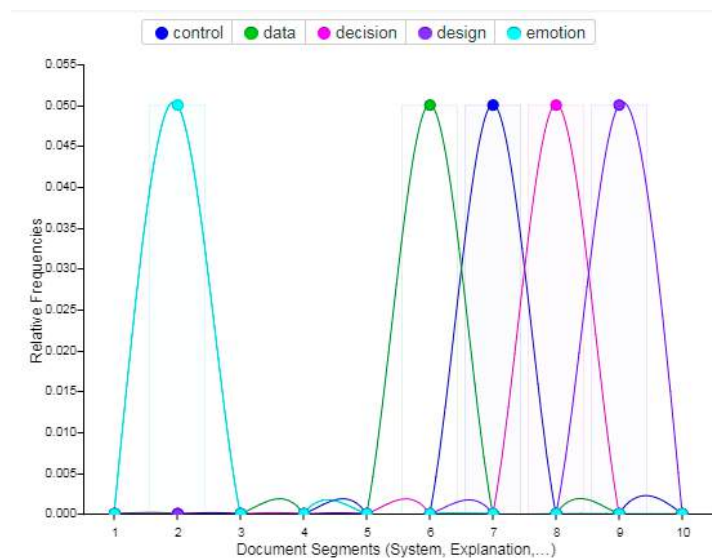


Figure 3. Word cloud 1 analysis.

Many meaningful sentences in the context of responsible XAI can, therefore, be formed from combining these words. For example, the words could be brought together to imply ‘a design of data control that enhances decision making’ or ‘designing and controlling data in such a way that enhances emotion and decisions’. The corpus from word cloud 2 reveals 20 words as the most prominent (model, use, learn, data, method, predict, behavior, task, perform, base, show, infer, image, algorithm, propose, optimal, object, general, explain, network). Voyant analysis revealed that five words were the most prominent words: action; algorithm, base, behavior, data, and explain, as seen in Figure 4 below.

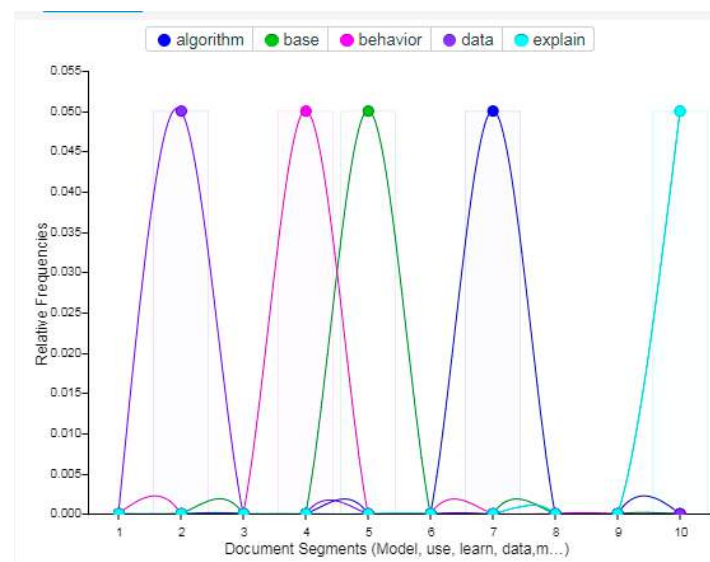


Figure 4. Word cloud 2 analysis.

Possible sentences from this word combination include: ‘an algorithm that explains data behavior’ or ‘an algorithm behavior that is based on data explanation’. Additionally, upon looking for the concordance of the term ‘explainability’ in sketchengine.eu (an open-source program that analyses how real users of a given language use certain words), some terms emerged as critical definitional terms of the word: predictability, verifiability, user feedback, information management, data insights, analytics, determinism, understanding, and accuracy, as seen in Figure 5 below.

	Details	Left context	KWIC	Right context
1	evo-art.org	ur primary levels and explicate their nonhomomorphic interlevel relations. </s><s> Explainability	of emergence in relation to determinism and predictability is considered. </s><s> R	
2	juli.net	i called empathy), as that's the precondition of finding help for people. </s><s> But explainability	does not mean that every kind of behavior is "good", i.e. has a right to stay unchan	
3	iapa.org.au	s that it takes to actually apply analytics – data restrictions, model accuracy versus explainability	, inter-departmental politics etc. </s><s> I subsequently started my tenure at SAS i	
4	jsmith.org	'the most profound truths and one that we should meditate seriously on. </s><s> " Explainability	" and "rationality" are both good things...at its core, the irrationality and inexplicabili	
5	usemp-project.e...	limedia content shared by OSN users and is designed for interactive use and with explainability	as a core requirement. </s><s> User assistance tools – that provide the users with	
6	usemp-project.e...	pective, translated into user control over automated processing and work towards explainability	of obtained results. </s><s> As explained in Section 1.1.1, we discriminate between	
7	wise.io	and your next best action. </s><s> And in some industries, such as credit scoring, explainability	of predictions can even be a regulatory requirement. </s><s> Most advanced mact	
8	usemp-project.e...	<s> Emphasis will be put on integrating user feedback in the USEMP tools and on explainability	of the extraction processes in order to facilitate the adoption of the tools by the use	
9	usemp-project.e...	ds adapted to personal information management, with focus on user feedback and explainability	in order to ensure fast adoption by the users. </s><s> S4 . Improve the manage	
10	global-systems-...	h mappings should enjoy a number of features, including: clarity and conciseness, explainability	, formal verifiability, and the ability of adapting to an enormous number of possible	
11	vukcevic.net	ients. </s><s> In one step of their modelling process, they do a trade-off between ' explainability	' and 'predictability'. </s><s> Specifically, they chose a model that was easier to int	
12	josemalvarez.es	nmenders based on optimizing a global cost function keeping advantages such as explainability	and handling new users while improving accuracy </s><s> a set of extensions to e.	
13	degreedays.net	the point of the site is to generate degree days for you , making the simplicity and explainability	of the calculation process irrelevant. </s><s> So, when we can't use the Integrator	
14	opossem.org	ncy curves are trivial. </s><s> And if a social scientist felt compelled for reasons of explainability	to discretize the analysis, then I certainly agree that doing an extreme-groups anal	
15	blackstone.name	s quite willing to replace the judgment of doctors with this net. </s><s> This lack of explainability	is a practical problem for applications involving people, but it is not an argument de	
16	blackstone.name	n argument demonstrating that the net cannot understand. </s><s> Indeed, lack of explainability	is to be expected if the net does understand. </s><s> The point is, understanding c	
17	bostoncommons.n...	s><s> At times it can be difficult to meet a company's Data Science goal of model explainability	– or data insights provided by the model – if the Data Scientist has not done a goo	
18	witgensteinrep...	and necessary condition; to these relations are added the claims of reducibility and explainability	. </s><s> A supervening property or fact can be reduced to the property or fact on	
19	gfk-leipzig.de	d by Klaus Werner </s><s> In closed systems, concepts like responsibility, justice, explainability	create the desire to withdraw from instrumentalized inventions. </s><s> Abundanc	
20	madm.eu	-further, the utilization of SentBank features shows high potential for detection and explainability	of such content. </s><s> Overall, multi-modal feature fusion can achieve an improv	

Figure 5. Concordance analysis for the term 'Explainability'.

5. Discussion

The findings from the results highlight critical tenets regarding the concept of explainability in AI. From the results, it is evident that specific keywords were prominently featured compared to others. Notably, words such as ‘use’, ‘explanation’, and ‘model’, as highlighted in Figure 2, confirm increased usability. The findings are key in addressing the shortcoming of the whole concept of explainability as reported by Soni et al. [5]. Similar results are reported in the case of Figure 3, which identifies notable words such as ‘use’, ‘model’, ‘learn’, and ‘data’. The analysis reveals that the AI communities approach the concept of explainability differently, but with some general resemblance. The term is used to assist in evaluating the mechanism of machine language systems, such as understanding the working of the system, yet, at other times, to relate to concepts of particular inputs like determining how a given input was mapped to an output. From the analysis, certain key observations were made.

5.1. Opaque Systems

Mainly, this refers to a system where mapping inputs to outputs are not visible to the user. It can be perceived as a mechanism that only predicts an input, without stating why and how the predictions are made. Opaque systems exist, for example, when closed-source AI systems are licensed by an organization where the licensor prefers to keep their workings private. Additionally, black-box methodologies are categorized as opaque because they use algorithms that do not give insight into the actual reasoning for the system mapping of inputs to outputs [21].

5.2. Interpretable Systems

This refers to AI systems where clients cannot see, understand, and study how mathematical concepts are used to map inputs to the output. Thus, related issues include transparency and understanding. Regression models are considered interpretable, for instance, in comparing covariate heights to determine the significance of each aspect to the mapping. However, deep neural networks’ functioning may not be interpretable, especially regarding the fact that input features are mostly based on automatic non-linear models [21].

5.3. Comprehensible Systems

Comprehensible AI models emit symbols without the use of inputs. These symbols are mostly words and allow the user to relate input properties to the outputs. Thus, the user can compile and understand the symbols relying on personal reasoning and knowledge about them. Thus, comprehensibility becomes a graded notion, with the extent of clarity being the difficulty or ease of use. The required form of knowledge from the user’s side relates to the cognitive intuition concerning how the output, inputs, and symbols relate to each other [21]. Interpretable and comprehensible systems are improvements over opaque systems. The concept of interpretation and comprehension are separate: interpretation mostly relies on the transparency of system models, whereas comprehension requires emitting symbols that allow a user to reason and think [21]. Most present-day AI systems can produce accurate output but are also highly sophisticated if not outright opaque, making their workings incomprehensible and difficult to interpret. As part of the efforts to renew user trust in AI systems, there are calls to increase system explainability.

There is an increasing need to enhance user confidence in AI systems [25]. User confidence and trust are cited as the primary reasons for pursuing explainable AI. System users seek explanations for various reasons—to manage social interactions, assign responsibility, and persuade others. A critical aspect of explainable AI is that it creates an opportunity for personalized human–computer interaction. Instead of the brick and mortar models of decision-making techniques that cannot be understood or interpreted by humans, explainability ensures that the customer journey through machine learning and AI systems is modeled in a way that mimics human interactions. There is a need for machine learning to explain why and how individual recommendations or decisions are made [26]. The implication is that

consumers' activity on AI systems will be complemented to make better and accurate decisions faster. Thus, organizations will be able to leverage customer value. When giving advice, recommendations, or even descriptive aspects, looking for justifications and reasons behind the recommended action is important. It is not enough to predict or suggest outcomes as the best or the preferred action without showing connections between the data and the factors involved to arrive at the decisions [26].

McNaught and Lam [22] state that most people have the perception that a doctor is 'a kind of black box' who transforms symptoms of illnesses and related test outcomes into the best diagnosis for treatment. Doctors deliver diagnostic recommendations to patients without explaining why and how they arrived at such decisions. Mostly, doctors use high-level indicators or symptoms (which, in the context of AI systems, denote system symbols). However, when handling a patient, the doctor should be like a comprehensible model. When interacting with medical staff and other physicians, the doctor may be like an interpretable model. Other doctors will interpret the technical analysis like an analyst would do to a ML system to ensure that the decisions arrived at are backed up by the reasonable assessment of the evidence provided.

Thus, XAI ensures that the information provided is accurate and personalized, to engage with targeted users in the most optimal manner [26]. In this way, XAI will increase the offer's relevance, thus enhancing user engagement and interest. Additionally, XAI will offer aspects that drive predicted outcomes, allowing real-time adjustment of primary business aspects to optimize gains and corporate outcomes. Transparency and reasoning will contribute to a decrease in abandoned products and an increase in order value, and thus higher revenue and conversion rates. Therefore, this research proposes implementing XAI frameworks that possess the following features: interpretability and comprehensibility. However, responsible XAI is more than just the ML features: user-related aspects of external AI features that should be considered are trust, fairness, confidence, ethics, and safety, as highlighted in So [27]. Moreover, the actual meaning of XAI is dependent on the perspective of the user, as highlighted by So [27] in Figure 6. As illustrated in Figure 6, XAI utilizes customer data to support the process of decision-making, with the impact being witnessed in improved business outcomes.

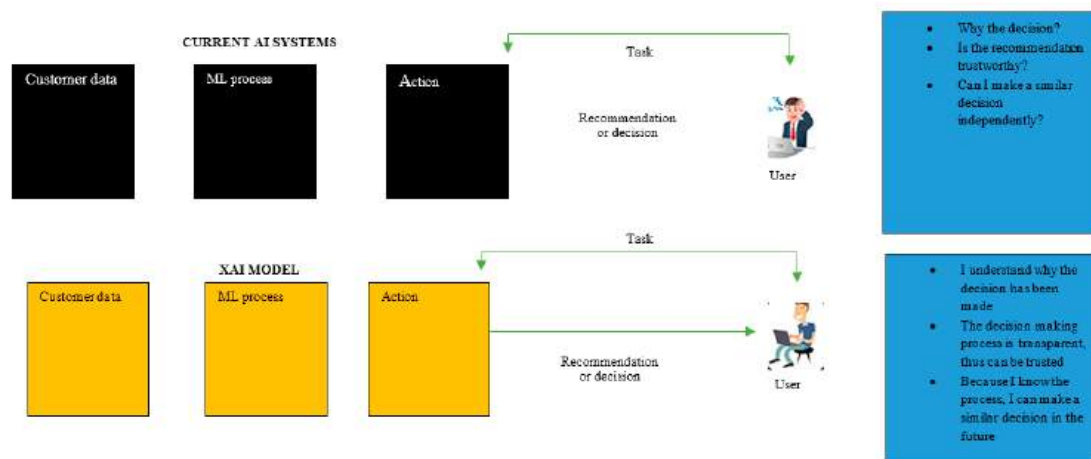


Figure 6. An analogy showing the differences between present-day AI and XAI models.

6. Conclusions

The study's main purpose was to lay the foundation for a universal definition of the term 'explainability'. The analyzed data from the word cloud plots revealed that the term 'explainability' is mainly associated with words such as model, explanation, and use. These were the most prominent words exhibited in the corpus generated from the word cloud. In the corpus from word cloud 1, they include emotion, design decision, data, control, image, variance, interpret, decision, show, result, control, estimate, data, predict, method, train, study, use, learn, model, explanation, and system.

In corpus word cloud 2, they include model, use, learn, data, method, predict, behavior, task, perform, base, show, infer, image, algorithm, propose, optimal, object, general, explain, network. After the Voyant analysis was conducted, the most prominent words that appeared on word cloud 1 included control, data, decision, design, and emotion, while in word cloud 2, they included algorithm, base, behavior, data, and explain. When the words obtained from the Voyant analysis were combined, they provided specific meanings of the word ‘explainability’. The main definitions obtained from the combination of the most frequent words include ‘an algorithm that explains data behavior’ or ‘an algorithm behavior that is based on data explanation’ or ‘a design of data control that enhances decision making’ or ‘designing and controlling data in such a way that enhances emotion and decisions’.

The application of AI in e-commerce stands to expand in the future, as businesses are appreciating their role in influencing consumer demands. The rapid development of research technology and increased access to the internet present e-commerce businesses with an opportunity to expand their various platforms. Notably, the influence of AI in e-commerce spills over to customer retention and satisfaction. The customers are at the center of the changes and adoption of AI in e-commerce. Hence, e-commerce can further develop contact with customers and establish developed customer relationship management systems.

The researcher of this study has made an effort to provide a critical outline of the key tenets of AI and its role in e-commerce, as well as a comprehensive insight regarding the role of AI in addressing the needs of consumers in the e-commerce industry. Even though the study has attempted to provide a universal meaning of ‘explainability’, the actual impact of AI on consumers’ decisions is not yet clear, considering the notion of the “black box”, i.e., if the decisions arrived at cannot be explained and the reasons behind such actions given, it will be difficult for people to trust AI systems. Therefore, there is a need for future studies further to examine the need for explainable AI systems in e-commerce and find solutions to the ‘black box’ issue.

For future research, this study may serve as a ‘template’ for the definition of an explainable system that characterizes three aspects: opaque systems where users have no access to insights that define the involved algorithmic mechanism; interpretable systems where users have access to the mathematical algorithmic mechanism, and comprehensible systems where users have access to symbols that enhance their decision making.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Soni, N.; Sharma, E.K.; Singh, N.; Kapoor, A. Artificial Intelligence in Business: From Research and Innovation to Market Deployment. *Procedia Comput. Sci.* **2020**, *167*, 2200–2210. [CrossRef]
2. Di Vaio, A.; Palladino, R.; Hassan, R.; Escobar, O. Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. *J. Bus. Res.* **2020**, *121*, 283–314. [CrossRef]
3. Di Vaio, A.; Boccia, F.; Landriani, L.; Palladino, R. Artificial Intelligence in the Agri-Food System: Rethinking Sustainable Business Models in the COVID-19 Scenario. *Sustainability* **2020**, *12*, 4851. [CrossRef]
4. Kumar, T.; Trakru, M. The Colossal Impact of Artificial Intelligence. E-Commerce: Statistics and Facts. *Int. Res. J. Eng. Technol. (IRJET)* **2020**, *6*, 570–572. Available online: <https://www.irjet.net/archives/V6/i5/IRJET-V6I5116.pdf> (accessed on 3 December 2020).
5. Soni, N.; Sharma, E.K.; Singh, N.; Kapoor, A. Impact of Artificial Intelligence on Businesses: From Research, Innovation, Market Deployment to Future Shifts in Business Models. *arXiv* **2019**, arXiv:1905.02092.
6. Soni, V.D. International Journal of Trend in Scientific Research and Development. *Int. J. Trend Sci. Res. Dev.* **2019**, *4*, 223–225. [CrossRef]
7. Ahmed, A.Z.; Rodríguez-Díaz, M. Significant Labels in Sentiment Analysis of Online Customer Reviews of Airlines. *Sustainability* **2020**, *12*, 8683. [CrossRef]
8. Castelvechi, D. Can we open the black box of AI? *Nat. Cell Biol.* **2016**, *538*, 20–23. [CrossRef]

9. Bathaee, Y. The artificial intelligence black box and the failure of intent and causation. *Harv. J. Law Technol.* **2017**, *31*, 889. Available online: <https://heinonline.org/HOL/LandingPage?handle=hein.journals/hjlt31&div=30&id=&page=> (accessed on 3 December 2020).
10. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
11. Păvăloaia, V.-D.; Teodor, E.-M.; Fotache, D.; Danileț, M. Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences. *Sustainability* **2019**, *11*, 4459. [CrossRef]
12. Ren, S.; Choi, T.-M.; Lee, C.; Lin, L. Intelligent service capacity allocation for cross-border-E-commerce related third-party-forwarding logistics operations: A deep learning approach. *Transp. Res. Part E Logist. Transp. Rev.* **2020**, *134*, 101834. [CrossRef]
13. Yi, S.; Liu, X. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex Intell. Syst.* **2020**, *6*, 621–634. [CrossRef]
14. Patel, R.; Passi, K. Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. *IoT* **2020**, *1*, 14. [CrossRef]
15. Davenport, T.; Guha, A.; Grewal, D.; Bressgott, T. How artificial intelligence will change the future of marketing. *J. Acad. Mark. Sci.* **2020**, *48*, 24–42. [CrossRef]
16. Tussyadiah, I.P.; Miller, G. Perceived Impacts of Artificial Intelligence and Responses to Positive Behaviour Change Intervention. In *Information and Communication Technologies in Tourism 2019*; Springer Science and Business Media LLC: Berlin, Germany, 2018; pp. 359–370.
17. Nadimpalli, M. Artificial Intelligence? Consumers and Industry Impact. *Int. J. Econ. Manag. Sci.* **2017**, *6*. [CrossRef]
18. Kachamas, P.; Akkaradamrongrat, S.; Sinthupinyo, S.; Chandrachai, A. Application of Artificial Intelligent in the Prediction of Consumer Behavior from Facebook Posts Analysis. *Int. J. Mach. Learn. Comput.* **2019**, *9*, 91–97. [CrossRef]
19. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
20. Heimerl, F.; Lohmann, S.; Lange, S.; Ertl, T. Word Cloud Explorer: Text Analytics Based on Word Clouds. In Proceedings of the 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 6–9 January 2014; pp. 1833–1842.
21. Doran, D.; Schulz, S.; Besold, T.R. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *arXiv* **2017**, arXiv:1710.00794. Available online: <https://openaccess.city.ac.uk/id/eprint/18660/> (accessed on 3 December 2020).
22. McNaught, C.; Lam, P. Using Wordle as a supplementary research tool. *Qual. Rep.* **2010**, *15*, 630–643.
23. Sinclair, J.; Cardew-Hall, M. The folksonomy tag cloud: When is it useful? *J. Inf. Sci.* **2008**, *34*, 15–29. [CrossRef]
24. Kuo, B.Y.-L.; Hentrich, T.; Good, B.M.; Wilkinson, M.D. Tag clouds for summarizing web search results. In Proceedings of the 16th international conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 1203–1204.
25. Babu, K.J. Business Intelligence: Concepts, Components, Techniques and Benefits. *SSRN Electron. J.* **2012**, *9*, 60–70. [CrossRef]
26. Siau, K.; Wang, W. Building trust in artificial intelligence, machine learning, and robotics. *Bus. Inf. Technol. Fac. Res. Creat. Works* **2020**, *31*, 47–53. Available online: https://scholarsmine.mst.edu/bio_inftec_facwork/325/ (accessed on 3 December 2020).
27. Kenn, S. Why Explainable AI is Exciting to VCs. Towards Data Science. 2019. Available online: <https://towardsdatascience.com/investor-view-explainable-ai-5ba66b31cd82> (accessed on 3 December 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

About Rule-Based Systems: Single Database Queries for Decision Making

Piotr Artiemjew ^{1,*} , Lada Rudikova ² and Oleg Myslivets ²

¹ Faculty of Mathematics and Computer Science, University of Warmia and Mazury in Olsztyn, 10-710 Olsztyn, Poland

² Faculty of Mathematics and Computer Science, Grodno State Yanka Kupala University, Street. Ozheshko 22, 230023 Grodno, Belarus; lada.rudikowa@gmail.com (L.R.); myslivec.oleg@yandex.ru (O.M.)

* Correspondence: artem@matman.uwm.edu.pl

Received: 10 November 2020; Accepted: 25 November 2020; Published: 27 November 2020

Abstract: One of the developmental directions of Future Internet technologies is the implementation of artificial intelligence systems for manipulating data and the surrounding world in a more complex way. Rule-based systems, very accessible for people's decision-making, play an important role in the family of computational intelligence methods. The use of decision-making rules along with decision trees are one of the simplest forms of presenting complex decision-making processes. Decision support systems, according to the cross-industry standard process for data mining (CRISP-DM) framework, require final embedding of the learned model in a given computer infrastructure, integrated circuits, etc. In this work, we deal with the topic concerning placing the learned rule-based model of decision support in the database environment—exactly in the SQL database tables. Our main goal is to place the previously trained model in the database and apply it by means of single queries. In our work we assume that the decision-making rules applied are mutually consistent and additionally the Minimal Description Length (MDL) rule is introduced. We propose a universal solution for any IF THEN rule induction algorithm.

Keywords: decision systems; rule based systems; databases; rough sets

1. Introduction

The last, very important stage of the CRISP-DM (Cross-industry standard process for data mining) framework [1] is the final implementation of the learned models at the disposal of end users. The ways in which models are placed, their security, and quick access for many users is a significant problem for computer system administrators and developers of decision-support devices. In this work, we have developed a way of placing decision-making rules in database tables and a method of making decisions that is based on individual database queries. This is a very simple and effective way to use decision support systems based on IF THEN type rules. The users can send queries to the Database and get the answer for their decision problem without local computations. It was possible to build a model under the assumption that the rules are consistent and built on the Riisanen's Minimal Description Length (MDL) rule [2]. The MDL rule, among other things, refers to the exclusion of longer rules, which contain previously accepted shorter rules generated by a specific method.

In the next section we define the goal set in the work.

1.1. The Aim of the Work

The main objective of the research was to develop a method of classification using rules collected in database tables (i.e., a previously learned rule-based model), which works through single queries to database tables. This has been achieved by assuming that we use non-contradictory rules that belong entirely to only one decision-making class and meet the condition set by the MDL rule; long rules cannot contain the correct short rules. In the following subsections, we introduce our solution and finally show an example of the method in Section 4.3.

In the next section we present a brief approach to the methods.

1.2. Brief Approach to the Methods

To achieve the goal, we needed to design a database system supported by a data mining module assessing the usefulness of a rule-based model. To assess the usefulness, we proposed the Cross Validation method, which allows us to estimate the quality of the rules generated in the model. This means that the model, which has been trained and embedded in a database table, works with a specific, real efficiency. This element can be seen in the diagram Figure 1.

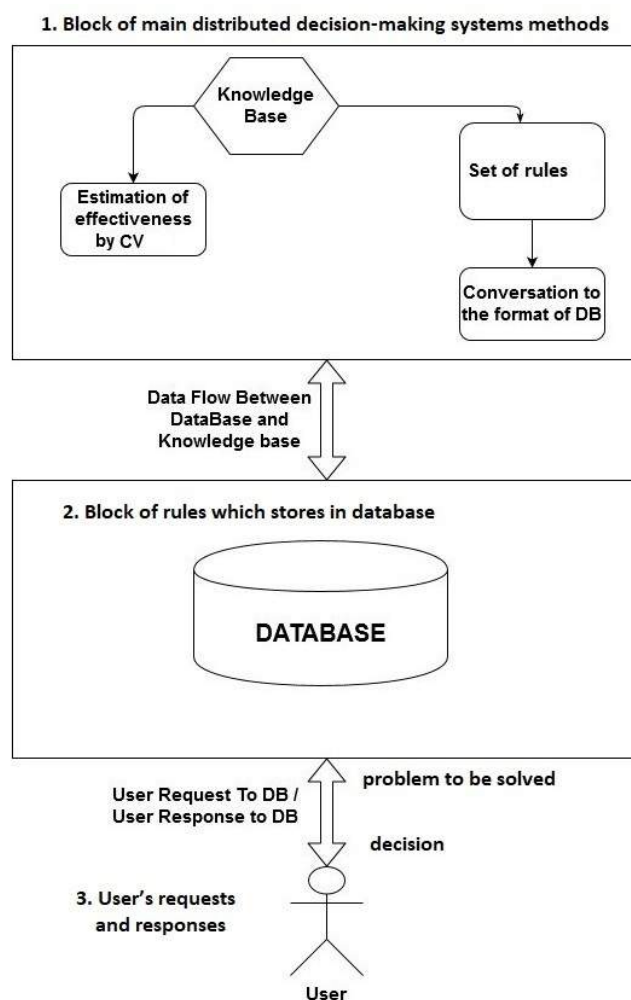


Figure 1. Basic scheme of the database system vs. rule-based knowledge based system.

As the heart of the data mining model, we have chosen the most popular methods of generating decision-making rules-exhaustive rules, sequentially covering and LEM2 rules usually applied in medical diagnosis solutions. We presented these methods in examples (see Section 4.1) and we have given examples of effectiveness in models based on these rules in Figure 2.

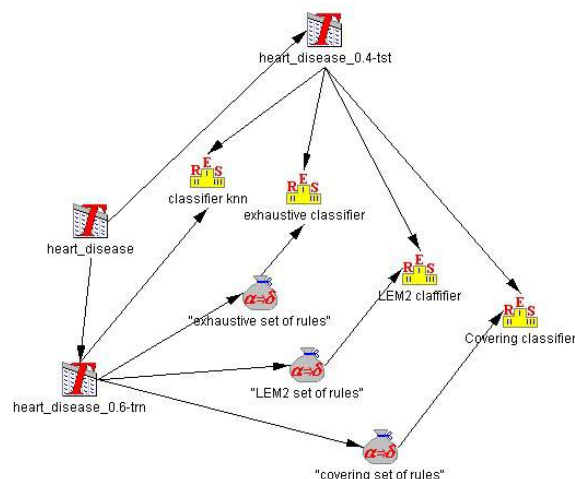


Figure 2. Considering Train and Test with split ratio 0.4 for Statlog (Heart) Data Set (270 objects, 14 attributes) [3], the Training system has 162 objects, the test system 108. Exemplary classification result; in case of knn we have total and balanced $accuracy = 0.833$, $coverage = 1$; in case of an exhaustive classifier we have total and balanced $accuracy = 0.824$, $coverage = 1$, $number_of_rules = 2588$; in case of an LEM2 classifier we have total $accuracy = 0.887$, balanced $accuracy = 0.871$, $coverage = 0.491$, $number_of_rules = 65$; in case of a Covering classifier we have total $accuracy = 0.554$, balanced $accuracy = 0.56$, $coverage = 0.852$, $number_of_rules = 188$.

As a container containing learned models we used SQL database tables. We have worked out how to put the rules in the proposed form in the database. The general scheme is in Figure 1. And the detailed concept approach in Section 4.2.

In order to achieve the correct classification, which is described in Section 4.3, we have proposed a specific form of rules, and a method for extracting the parameters that determine the classification, by creating an appropriate query to the database. Our queries (consisting in passing the problem to the database to be solved) to the database directly return the name of the decision class that is most likely according to the available rules.

The rest of the paper is constructed as follows. In Section 2, we introduce basic information about rule-based systems and we mention the existing similar solutions to the one presented in this work. In Section 3 we present the basic assumptions about the database system. In Section 4 we provide an introduction to the steps needed to build the system. In the following subsections we present the next steps in detail with examples. In Section 4.1 we discuss selected techniques of rule generation, in Section 4.2 we introduce the way of communication with the database and the representation of rules, in Section 4.3 we have an introduction to the sample classifier used on rules downloaded from the database, and we present an additional test showing the potential effectiveness of the system. In Section 5 we conduct a critical analysis of the features of rule generation techniques. In Section 6 we benchmark our system, verifying linear computational complexity. In Section 7 we summarize the results of the work and discuss future plans.

Let us move on to the matter of rule-based decision-making models.

2. Background

We are thinking about rule-based systems through the prism of the rough set theory-see Pawlak [4]. Where we start from the definition of the decision system as a triple (U, A, d) , where U is a universe of objects, A is a set of conditional attributes, d is a decision attribute fulfilling a condition ($d \notin A$). To clarify this definition, see Table 1 for a sample decision system. In our example, the universe of objects $U = \{person_1, person_2, person_3\}$, the set of conditional attributes A is $\{color, body_temperature\}$ and the decision attribute $d \in D = \{Yes, No\}$.

Table 1. The exemplary decision system (U, A, d) .

	<i>Color</i>	<i>body_temperature</i>	<i>d</i>
<i>person₁</i>	red	40.	Yes
<i>person₂</i>	green	36.6	No
<i>person₃</i>	blue	36.5	No

For the formal representation of decision rules, we need to introduce a descriptor notation ($a_i = value$). For example ($a_1 = 1$) is a general reference to all objects of the decision system that have the value 1 on the conditional attribute a_1 . Exemplary IF THEN rules can be defined as:

$$IF(a_1 = 1) THEN (d = YES)$$

$$IF(color = red) OR (temperature > 100) THEN (d = potential_danger)$$

$$IF(time = free) AND (attitude = good) THEN (d = go_to_the_park)$$

There are several basic types of rule generation algorithms, for instance an exhaustive set of rules [5] is dedicated to problems, where the decision must be made with the greatest possible raw knowledge available-this is a Brute Force technique. The sequential covering type [6] gives brief solutions to all problems by using short rules, which fully cover the universe of objects. The last important family of techniques worth mentioning is based on Modules of data (non conflicting data blocks), an example method is the LEM2 algorithm [7], which allows to create long characteristic, most trusted patterns from data. Such methods are used for medical data. Most other methods use elements of these basic three approaches. If someone would like to use decision rules in uncertain conditions, the indiscernibility degree of descriptors should be defined. It is not our goal to review all available variants of rule generation algorithms, but to present system operation on a representative set of rules.

Let us move on to a discussion of selected work on the application of decision-making rules in databases. The seemingly simple subject of using rule-based systems in databases is a specific niche, in which there is not much research or many technical reports. We will mention the following examples of work on this subject. In [8] authors use a database (ORACLE) to implement the functionality of an expert system shell by a simple rule language which is automatically translated into SQL. In [9] we have a presentation on how to maintain rule-based knowledge in temporal databases. In the paper [10] we have implementation of a rule based system at MTBC, for applying billing compliance rules on medical claims. In article [11] the author considers the problem of searching for rules in relational databases for potential use in data mining. Then in the article [12] we can see discussion of generation of Apriori-Based Rules in SQL. Finally, at work [13] authors apply an active database system in an academic information system, to plan academic business rules.

3. The Basic Assumptions for Our Distributed Decision Support System

In the Figure 1 represents the general view of our distributed decision system. The whole system can be divided into 3 logical components:

1. Block of main distributed decision-making systems methods,
2. Block of rules which stores in database,
3. User's requests and responses.

The first block has the following components: the knowledge base is a set of resolved problems in a fixed context. The set of decision rules is the set of rules created by a selected algorithm. Estimation of effectiveness by CV (Cross Validation) is the block in which we have the internal estimation of the decision process effectiveness. The estimation is needed to show how the stored rules possibly work on new data. It is just the quality of the decision-making product. Conversion to the format for DB it's the converter, which lets you put the rules into the Database. DB is simply Database. USER is the person who asks the questions to the database and expects the solution for the decision problem. In the block USER (Question) the user asks the question to DB. In the block USER (Answer from DB), the USER gets an answer from the DB to his question.

Allow us to introduce the design of our system.

4. Stages in the Design of a Rule-Based Database System

- STEP 1: We calculate (IF THEN) decision rules for the selected technique, (1st block of Figure 1)
- STEP 2: We create a database table containing all conditional attributes, decision attribute, rule support and rule length, where empty attribute values (not used in the rule) are marked as '-', (2nd block of Figure 1)
- STEP 3: On the client's side we load the problem to be solved. The decision is made by a single database query. (3rd block of Figure 1)

4.1. STEP 1 in Detail-Discussion of Selected Rule Generation Techniques

Allow us to present some toy examples of selected rule generation techniques based on the system from Table 2.

Table 2. The exemplary decision system-for rule induction [14].

Day	Weather	Temperature	Humidity	Wind	Play_Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cold	Normal	Weak	Yes
D6	Rain	Cold	Normal	Strong	No
D7	Overcast	Cold	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

The implementation of the methods described below can be tested with the RSES (Rough Set Exploration System) [5] tool. We will begin by presenting the exhaustive method.

4.1.1. Exhaustive Set of Rules

Calculation of an exhaustive set of rules consists of browsing through combinations without repetition of attributes, starting from the shortest, and checking their consistency against the decision. The MDL (Minimal description length) concept is used in this method. MDL consists in not accepting longer rules, which contain the correct shorter ones. This method is from the Brute Force family of techniques. Despite the fact that it uses the greatest possible knowledge from decision systems, it is at risk of using rules calculated from data noise. The method is quantitative, not necessarily giving the best possible results for a specific subgroup of objects. The exponential computational complexity limits usability on large data. It is worth knowing that a tool useful in calculating exhaustive rules is the relative indiscernibility matrix-see [15].

Considering the decision system from Table 2, the exhaustive set of rules is as follows:

- $(Weather = Overcast) \Rightarrow (Play_Tennis = Yes[4])$
- $(Humidity = Normal) \& (Wind = Weak) \Rightarrow (Play_Tennis = Yes[4])$
- $(Weather = Sunny) \& (Humidity = High) \Rightarrow (Play_Tennis = No[3])$
- $(Weather = Rain) \& (Wind = Weak) \Rightarrow (Play_Tennis = Yes[3])$
- $(Weather = Sunny) \& (Temperature = Hot) \Rightarrow (Play_Tennis = No[2])$
- $(Temperature = Cold) \& (Wind = Weak) \Rightarrow (Play_Tennis = Yes[2])$
- $(Weather = Rain) \& (Wind = Strong) \Rightarrow (Play_Tennis = No[2])$
- $(Weather = Sunny) \& (Humidity = Normal) \Rightarrow (Play_Tennis = Yes[2])$
- $(Temperature = Mild) \& (Humidity = Normal) \Rightarrow (Play_Tennis = Yes[2])$
- $(Temperature = Hot) \& (Wind = Strong) \Rightarrow (Play_Tennis = No[1])$
- $(Weather = Sunny) \& (Temperature = Mild) \& (Wind = Weak) \Rightarrow (Play_Tennis = No[1])$
- $(Weather = Sunny) \& (Temperature = Cold) \Rightarrow (Play_Tennis = Yes[1])$
- $(Weather = Sunny) \& (Temperature = Mild) \& (Wind = Strong) \Rightarrow (Play_Tennis = Yes[1])$
- $(Temperature = Hot) \& (Humidity = Normal) \Rightarrow (Play_Tennis = Yes[1])$

Now allow us to present a toy example of a sequential covering type algorithm.

4.1.2. Sequential Covering Variant

The sequential type of methods comes from Mitchell [16] works. It consists of searching for the shortest possible rules, superficially covering the decision system. Covered objects are removed from consideration. An exemplar of implementation can be seen in the RSES tool [5]. This type of technique is used when the data set is large and the solution is needed quickly.

- $(Weather = Overcast) \Rightarrow (Play_Tennis = Yes[4])$
- $(Weather = Sunny) \& (Temperature = Hot) \Rightarrow (Play_Tennis = No[2])$
- $(Weather = Rain) \& (Wind = Weak) \Rightarrow (Play_Tennis = Yes[3])$
- $(Weather = Rain) \& (Wind = Strong) \Rightarrow (Play_Tennis = No[2])$
- $(Weather = Sunny) \& (Humidity = High) \Rightarrow (Play_Tennis = No[3])$
- $(Weather = Sunny) \& (Temperature = Cold) \Rightarrow (Play_Tennis = Yes)$
- $(Weather = Sunny) \& (Humidity = Normal) \Rightarrow (Play_Tennis = Yes[2])$

Now allow us to discuss a sample technique particularly designed for medical data.

4.1.3. Minimal Sets of Rules: LEM2

The methods consist in searching for the smallest set of rules representing the most common patterns in data. An illustrative technique is LEM2 (Learn from Examples by Modules 2) introduced by Grzymala-Busse [7,17,18]. LEM2 is based on a general heuristic principle, cf, Michalski's AQ [19], Clark's CN2 [20], or PRISM, Cendrowska [21], by which, rule descriptors are selected based on an established criterion (e.g., frequency) and rule-covered objects removed from consideration. We only remember to include them always in the rule support. The calculation is continued until the system is covered by rules, or until there are completely contradictory combinations in the system.

The LEM2 set of rules from the Table 2 is as follows:

- $(Humidity = Normal) \& (Wind = Weak) \Rightarrow (Play_Tennis = Yes[4])$
- $(Weather = Overcast) \Rightarrow (Play_Tennis = Yes[4])$
- $(Humidity = High) \& (Weather = Sunny) \Rightarrow (Play_Tennis = No[3])$
- $(Weather = Rain) \& (Wind = Strong) \Rightarrow (Play_Tennis = No[2])$
- $(Temperature = Mild) \& (Weather = Rain) \& (Humidity = High) \& (Wind = Weak) \Rightarrow (Play_Tennis = Yes[1])$

In order to present the effectiveness of the techniques described in the section, we conducted a demonstration experiment using the RSES system on selected data from UCI Repository [3]-see Figure 2.

Let us present samples of real rules and related statistics, generated from the heart disease system [3]. The meaning of the individual attribute names [3]: *attr0* = age, *attr1* sex, *attr2* = chest pain type, *attr3* = resting blood pressure, *attr4* = serum cholesterol in mg/dL, *attr5* = fasting bloodsugar > 120 mg/dL, *attr6* = resting electrocardiographic results, *attr7* = maximum heart rate achieved, *attr8* = exercise induced angina, *attr9* = oldpeak, ST depression induced by exercise relative to rest, *attr10* = the slope of the peak exercise ST segment, *attr11* = number of major vessels (0–3) colored by flourosopy, *attr12* = thal : 3 = normal; 6 = fixed defect; 7 = reversable defect.

Using the defined notation of *attr1*, ..., *attr12* attributes of the rules, an example of the most supported exhaustive rules is in Table 3-based on RSES tool [5]. A similar example for the LEM2 algorithm can be seen in Table 4. And finally for the sequential covering technique in Table 5.

Table 3. Examples of real rules-exhaustive algorithm, Statlog (heart) Data Set [3], Rule statistics: 5352 rules, Support of rules Minimal: 1, Maximal: 32, Mean: 1.7, Length of rule premises, Minimal: 1, Maximal: 6, Mean: 2.6, Distribution of rules among decision classes, Decision class 1: 2732, Decision class 2: 2620.

$(attr1=0) \& (attr5=0) \& (attr6=0) \& (attr8=0) \& (attr12=30) \Rightarrow (attr13=1[32])$
$(attr1=0) \& (attr2=30) \& (attr12=30) \Rightarrow (attr13=1[29])$
$(attr1=0) \& (attr6=0) \& (attr10=10) \Rightarrow (attr13=1[26])$
$(attr2=40) \& (attr6=20) \& (attr10=20) \& (attr12=70) \Rightarrow (attr13=2[25])$
$(attr1=0) \& (attr6=0) \& (attr8=0) \& (attr11=0) \Rightarrow (attr13=1[25])$
$(attr1=0) \& (attr2=30) \& (attr11=0) \Rightarrow (attr13=1[24])$
$(attr2=20) \& (attr5=0) \& (attr10=10) \& (attr12=30) \Rightarrow (attr13=1[22])$
$(attr1=0) \& (attr5=0) \& (attr8=0) \& (attr10=20) \& (attr12=30) \Rightarrow (attr13=1[20])$
$(attr1=0) \& (attr2=30) \& (attr10=10) \Rightarrow (attr13=1[20])$

Table 4. Examples of real rules-LEM2 algorithm, Statlog (heart) Data Set [3], Rule statistics: 99 rules, Support of rules Minimal: 1, Maximal: 23, Mean: 3.1, Length of rule premises, Minimal: 4, Maximal: 10, Mean: 7.2, Distribution of rules among decision classes, Decision class 1: 43, Decision class 2: 56.

(attr5=0)&(attr8=0)&(attr12=30)&(attr1=0)&(attr2=30)=>(attr13=1[23])
(attr5=0)&(attr8=0)&(attr12=30)&(attr11=0)&(attr1=0)&(attr6=0)=>(attr13=1[22])
(attr5=0)&(attr8=0)&(attr12=30)&(attr10=10)&(attr6=0)&(attr1=0)=>(attr13=1[21])
(attr5=0)&(attr1=10)&(attr2=40)&(attr12=70)&(attr8=10)&(attr10=20)&(attr6=20)=>(attr13=2[13])
(attr5=0)&(attr1=10)&(attr2=40)&(attr8=10)&(attr6=20)&(attr11=10)=>(attr13=2[12])
(attr8=0)&(attr1=10)&(attr11=0)&(attr5=10)=>(attr13=1[11])
(attr5=0)&(attr8=0)&(attr12=30)&(attr11=0)&(attr6=20)&(attr1=0)&(attr10=20)=>(attr13=1[9])

Table 5. Examples of real rules-Sequential covering algorithm, Statlog (heart) Data Set [3], Rule statistics: 199 rules, Support of rules Minimal: 1, Maximal: 7, Mean: 1.6, Length of rule premises, Minimal: 1, Maximal: 1, Mean: 1, Distribution of rules among decision classes, Decision class 1: 106, Decision class 93: 56.

(attr7=1720)=>(attr13=1[7])
(attr7=1780)=>(attr13=1[5])
(attr4=2040)=>(attr13=1[4])
(attr4=2110)=>(attr13=1[4])
(attr4=2260)=>(attr13=1[4])
(attr4=2820)=>(attr13=2[4])
(attr7=1510)=>(attr13=1[4])
(attr7=1790)=>(attr13=1[4])

4.2. STEP 2 in Detail-Rule Converter for Database

We need to write the rules in the form, which can be uploaded into the Database. The rules listed in point 1 are inserted into the database table. Creating an array looks like this.

```
CREATE TABLE rules(Weather TEXT, Temperature TEXT, Humidity TEXT, Wind TEXT, Play_Tennis TEXT, length int, support int);
```

Considering the training decision system (U_{trn}, A, d) , $|A|$ as the cardinality of set A . In general, the common form or all SQL-queries can be represented with the following syntax:

INSERT INTO table($a_1, a_2, a_3, \dots, a_{|A|}$, support, length)
VALUES($v_1, v_2, v_3, \dots, v_{|A|+2}$);

attributes not used in the rule are initialized with a value ‘-’. A demonstration of the representation of the true rule calculated from the system 2 is shown in Table 6.

Table 6. An example of how to upload rules into the database.

Original Rule	SQL QUERY
(Weather = Overcast) => (Play_Tennis = Yes[4])	INSERT INTO rules(Weather, Temperature, Humidity, Wind, Play_Tennis, length, support) VALUES('Overcast', '-', '-', '-', '-', 'Yes', 4, 1);
(Humidity = Normal) &(Wind = Weak) => (Play_Tennis = Yes[4])	INSERT INTO rules(Weather, Temperature, Humidity, Wind, Play_Tennis, length, support) VALUES('-', '-', '-', 'Normal', 'Weak', 'Yes', 4, 2);

4.3. STEP 3: Decision-Making Based on a Database Query

Allow us now to present a query to the trained database, which will return the decision to our test object tst_i .

Basically there are many ways to perform classification based on decision rules. One of the simplest techniques is to find out which rules fit the test object one hundred percent and use their supports and lengths to perform the competition between classes. The support is the number of training objects, which fits the rule. Length is the number of conditional attributes of a rule. Considering the set of rules, which fits to the exemplary object tst . Assuming that $size(rule_i)$ is the number of conditional attributes of i -th rule and $support(rule_i)$ is the support of this rule. The scheme of classification inside the Database system can be as follows.

$$Param_{class_j}(tst) = \sum_{\text{rules, which fits } tst \text{ and belong to } j\text{-th class}} support(rule_i) * length(rule_i)$$

Allow us to go to a sample database query. Let's assume that our test object (problem to be solved) is in Table 7.

Table 7. The demonstration problem to be solved.

Test Object
Is it worth playing tennis? when Weather is Sunny Temperature is Mild Humidity is Normal Wind is Weak

The database query that performs the classification process described above on the selected test object is in Table 8.

Table 8. The demonstration of the query used to classify the test object.

Decision-Making by Single Query
<pre>SELECT Play_Tennis FROM rules WHERE Weather in('Sunny',' -') AND Temperature in('Mild',' -') AND Humidity in('Normal',' -') AND Wind in('Weak',' -') GROUP BY Play_Tennis ORDER BY SUM(support * length) DESC LIMIT 1;</pre>

In this way, we select only the rules that fit the test object one hundred percent and carry out the classification based on them. The query returns the decision for the test object.

For noteworthy variants of classifiers based on rules we refer the reader to works [22,23].

5. Critical Analysis

Making decisions based on rules is a very natural process, but people are not able to use too complex rules. In complex problems and large amounts of data, rule-based decision support systems come in handy. Each of the techniques of rule generation has their own advantages and disadvantages. The exhaustive method, for example, has exponential complexity and it is often difficult to count rules in a reasonable

time. The LEM2 method gives very accurate rules, but often does not give decisions to objects. Sequential covering techniques, on the other hand, give too general and short rules that have little classification efficiency when there is little data. That is, the application of a set of rules in the database is associated with the problems mentioned above.

6. Verification of System Speed by Benchmarking

The technical details concerning our method. As a computational environment we have used phpMyAdmin a free software tool written in PHP, intended to handle the administration of MySQL over the Web. We were making a series of automatically generated queries inside. In the experiments we used automatically generated data, the samples of which we can see in Tables 9, 10 and 11.

Figures 3–5 show the speed of our technique in different variants. Examples of random queries used in benchmarking can be found in Table 9 and 11. The code used to test the query speed can be seen in Table 12. The presented verification of the speed of our classification method shows its linear computational complexity.

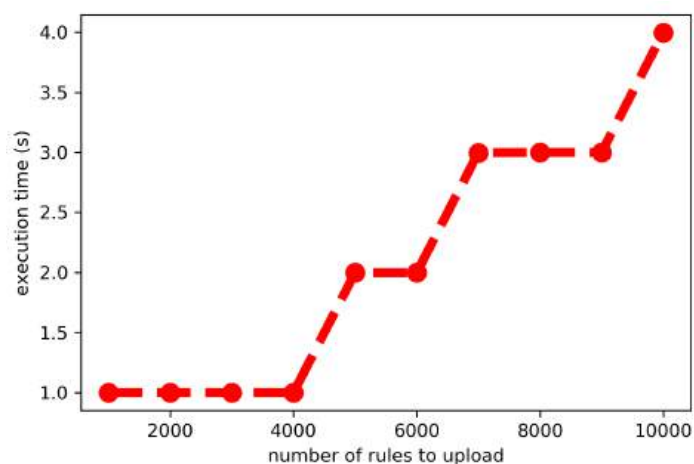


Figure 3. The speed of loading rules into the database for batches of 1000, 2000, ..., 10,000 rules.

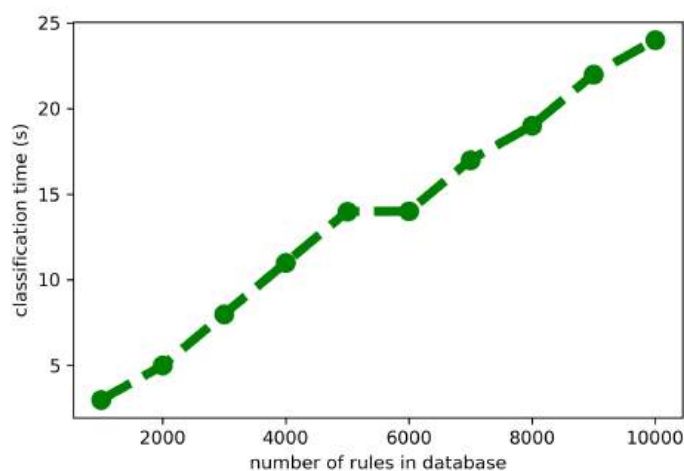


Figure 4. Speed of classification of 1000 random test objects in databases containing from 1000 to 10,000 rules.

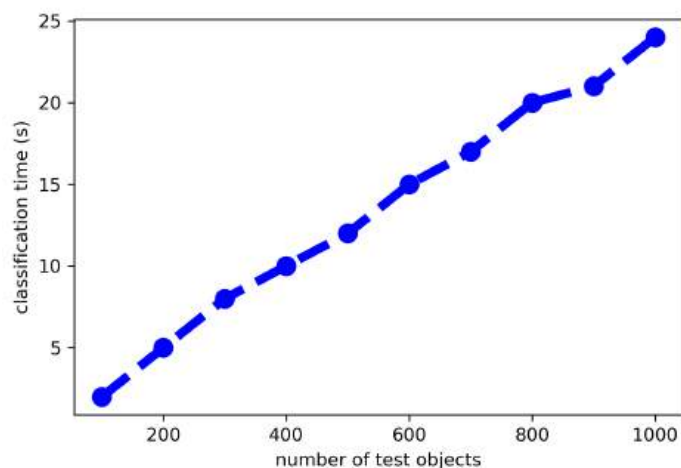


Figure 5. The classification speed of 100 to 1000 test objects by 10,000 rules in the database table.

Table 9. Sample of random rules created by the generator for benchmarking purposes. In experiments we use 1000, 2000, ..., 10,000 randomly generated rules.

Rules in the Form of a Database Query
INSERT INTO rules(Weather,Temperature,Humidity,Wind,Play_Tennis,length,support)VALUES('Rain','-'','-'','-'','Yes',4,5);
INSERT INTO rules(Weather,Temperature,Humidity,Wind,Play_Tennis,length,support)VALUES('Overcast','Mild','-'','Strong','No',1,4);
INSERT INTO rules(Weather,Temperature,Humidity,Wind,Play_Tennis,length,support)VALUES('Overcast','Cold','Normal','Strong','Yes',2,4);

Table 10. Creating tables to test the classification process, In experiments we upload to Tables 1000, 2000, ..., 10000 randomly generated rules.

Creating Database Tables
CREATE TABLE rules1000(Weather TEXT, Temperature TEXT, Humidity TEXT, Wind TEXT, Play_Tennis TEXT, length int, support int);
CREATE TABLE rules2000(Weather TEXT, Temperature TEXT, Humidity TEXT, Wind TEXT, Play_Tennis TEXT, length int, support int);
CREATE TABLE rules3000(Weather TEXT, Temperature TEXT, Humidity TEXT, Wind TEXT, Play_Tennis TEXT, length int, support int);

Table 11. Sample of random generated classification query.

Classification Based on Single Query
SELECT Play_Tennis FROM rules10000
WHERE Weather in('Sunny','-'')
AND Temperature in('Mild','-'')
AND Humidity in('Normal','-'')
AND Wind in('Weak','-'')
GROUP BY Play_Tennis
ORDER BY SUM(support*length) DESC LIMIT 1;

Table 12. The code for calculating the speed of execution of database operations.

```

SET @start_time = (TIMESTAMP(NOW()) * 1000000 + MICROSECOND(NOW(6)));
select @start_time;

...

SET OF DATABASE QUERIES

...

SET @end_time = (TIMESTAMP(NOW()) * 1000000 + MICROSECOND(NOW(6)));
select @end_time;
select (@end_time-@start_time)/1000;

```

7. Conclusions

This work is about rule-based decision support systems designed for a Database environment. We present the project in a detailed way, step by step, starting from the generation of decision rules and inserting them into the database with appropriate queries. The project is concluded with the presentation of a query allowing an appropriate classification using the rules stored in the database. The presented scheme allows us to build a real-time working (IF THEN) rule-based classifier. The use of our solution is limited to methods generating IF THEN type rules, whose sets of rules do not contain contradictions or longer rules containing correct shorter ones. Other limitations are mainly due to computational complexity of individual methods and the level of accuracy of covering the knowledge contained in the decision-making systems by the rules generated.

Despite initial results that proved to be very promising a great amount of work is left to be done to evaluate the final performance and determine the application of our new method.

This work calls for several perspectives. In future implementation of the presented system in selected mobile devices is planned. We are also going to use the system in real-life data mining projects.

Author Contributions: Conceptualization, P.A., L.R. and O.M.; methodology, software, validation, formal analysis and investigation, P.A. and O.M.; data acquisition and classification P.A., L.R.; resources, P.A. and L.R.; writing—original draft preparation, P.A. and L.R.; writing—review and editing, P.A. and L.R.; funding acquisition, P.A. and L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The research has been supported by grant 23.610.007-300 from Ministry of Science and Higher Education of the Republic of Poland.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shearer, C. The CRISP-DM model: the new blueprint for data mining. *J. Data Warehous.* **2000**, *5*, 13–22.
2. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 445–471. [CrossRef]
3. UCI (University of California at Irvine) Repository. Available online: <http://archive.ics.uci.edu/ml/> (accessed on 2 August 2020).
4. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
5. RSES. Available online: <http://mimuw.edu.pl/logic/~rses/> (accessed on 1 April 2014).
6. Skowron, A. Boolean reasoning for decision rules generation. In *Proceedings of the ISMIS'93. Lecture Notes in Artificial Intelligence*; Springer: Berlin, Germany, 1993; Volume 689, pp. 295–305.
7. Grzymala-Busse, J.W. LERS—A system for learning from examples based on rough sets. In *Intelligent Decision Support: Handbook of Advances and Applications of the Rough Sets Theory*; Słowiński, R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1992; pp. 3–18.

8. Skarek, P.; László, Z.V. Rule-Based Knowledge Representation Using a Database. In Proceedings of the Conference on Artificial Intelligence Applications, Paris, France, 21–22 October 1996.
9. Lorentzos, N.A.; Yialouris, C.P.; Sideridis, A.B. Time-evolving rule-based knowledge bases. *Data Knowl. Eng.* **1999**, *29*, 313–335. [CrossRef]
10. Abdullah, U.; Sawar, M.J.; Ahmed, A. Design of a Rule Based System Using Structured Query Language. In Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Chengdu, China, 12–14 December 2009; pp. 223–228. [CrossRef]
11. Spits Warnars, H.L.H. *Classification Rule with Simple Select SQL Statement*; National Seminar University of Budi Luhur: Jakarta, Indonesia; University of Budi Luhur: Jakarta, Indonesia, 2010.
12. Liu, C.; Sakai, H.; Zhu, X.; Nakata, M. On Apriori-Based Rule Generation in SQL—A Case of the Deterministic Information System. In Proceedings of the 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS), Sapporo, Japan, 25–28 August 2016; pp. 178–182. [CrossRef]
13. Miftakul Amin, M.; Andino, M.; Shankar, K.; Eswaran Perumal, R.; Vidhyavathi, M.; Lakshmanaprabu, S.K. Active Database System Approach and Rule Based in the Development of Academic Information System. *Int. J. Eng. Technol.* **2018**, 95–101. [CrossRef]
14. Salzberg, S.L. *C4.5: Programs for Machine Learning by J. Ross Quinlan*; Morgan Kaufmann Publishers: Burlington, MA, USA, 1993; Mach Learn 16, 235–240 (1994). [CrossRef]
15. Artiemjew, P. On Strategies of Knowledge Granulation with Applications to Decision Systems. Ph.D. Thesis, Polish–Japanese Institute of Information Technology, Warszawa, Polish, 2009.
16. Mitchell, T. *Machine Learning*; McGraw-Hill: Englewood Cliffs, NJ, USA, 1997.
17. Chan, C.C.; Grzymala-Busse, J.W. On the two local inductive algorithms: PRISM and LEM2. *Found. Comput. Decis. Sci.* **1994**, *19*, 185–204.
18. Grzymala-Busse, J.W.; Lakshmanan, A. LEM2 with interval extension: An induction algorithm for numerical attributes. In *Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96)*; The University of Tokyo: Tokyo, Japan, 1996; pp. 67–76.
19. Michalski, R.S.; Moztetic, I.; Hong, J.; Lavrac, N. The multi-purpose incremental learning system AQ15 and its testing to three medical domains. In *Proceedings of AAAI-86*; Morgan Kaufmann: San Mateo, CA, USA, 1986; pp. 1041–1045.
20. Clark, P.; Niblett, T. The CN2 induction algorithm. *Mach. Learn.* **1989**, *3*, 261–283. [CrossRef]
21. Cendrowska, J. PRISM, an algorithm for inducing modular rules. *Int. J. Man Mach. Stud.* **1987**, *27*, 349–370. [CrossRef]
22. Artiemjew, P. On clasification of data by means of rough mereological granules of objects and rules. In Proceedings of the International Conference on Rough Set and Knowledge Technology RSKT'08, Chengdu, China, 17–19 May 2008; LNAI. Springer: Berlin, Germany, 2008; Volume 5009, pp. 221–228.
23. Polkowski, L.; Artiemjew, P. Granular Computing in Decision Approximation, An Application of Rough Mereology. In *Series: Intelligent Systems Reference Library*; Springer: Berlin, Germany, 2015; Volume 77, 452p.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Predicting Activities of Daily Living with Spatio-Temporal Information

Sook-Ling Chua ^{1,*} , Lee Kien Foo ¹  and Hans W. Guesgen ² 

¹ Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya 63100, Selangor, Malaysia; lkfoo@mmu.edu.my

² School of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand; h.w.guesgen@massey.ac.nz

* Correspondence: slchua@mmu.edu.my; Tel.: +60-3-8312-5579

Received: 30 October 2020; Accepted: 26 November 2020; Published: 27 November 2020

Abstract: The smart home has begun playing an important role in supporting independent living by monitoring the activities of daily living, typically for the elderly who live alone. Activity recognition in smart homes has been studied by many researchers with much effort spent on modeling user activities to predict behaviors. Most people, when performing their daily activities, interact with multiple objects both in space and through time. The interactions between user and objects in the home can provide rich contextual information in interpreting human activity. This paper shows the importance of spatial and temporal information for reasoning in smart homes and demonstrates how such information is represented for activity recognition. Evaluation was conducted on three publicly available smart-home datasets. Our method achieved an average recognition accuracy of more than 81% when predicting user activities given the spatial and temporal information.

Keywords: prediction by partial matching; spatio-temporal; activity recognition; smart homes

1. Introduction

Almost every country in the world is experiencing a growing and aging population. The smart home is considered a viable solution to address living problems, typically the elderly or those with diminished cognitive capabilities. An important part of the functioning of smart homes is to monitor user daily activities and detect any alarming situations (e.g., skipping meals several days in a row). Sensors attached to objects of daily use (e.g., fridge, light, etc.) are often deployed in the smart home to collect information about user daily activities. These sensors are activated when the user performs their activities (e.g., opening the fridge, turning on the light, etc.). The recognition system uses the sensory outputs from the home to learn about user activity patterns and predict the next probable event.

The majority of the probabilistic graphical models such as the hidden Markov model and its variants, and deep learning methods for activity recognition, can predict where the user will go next or what activity is the user doing given the sequence of sensor readings [1–5]. However, for a smart home to react intelligently and support its users, the recognition system should not only recognize their activities but also have the ability to reason, e.g., at what time and in which room did a particular event occur.

Knowing the contexts of the user such as when and where a particular event occurred are important for detecting any unusual or abnormal events and issue warnings to caregivers or family members. For example, the recognition system learns that the user always goes to bed at 10 p.m. If something is happening after 10 p.m. then it is more likely that the user is sleeping. If the user were in the kitchen doing laundry at 1 a.m., this would be something that the recognition system

will recognize as unusual. Both space and time play an important role in activity recognition, and to represent and fuse all this information in the smart home poses a challenge.

An individual's pattern of daily activities is likely the same everyday. For example, a morning routine could consist of making the bed, grooming, making coffee and having breakfast. It turns out that compression can be used to identify repeated 'patterns' that represent user activities. To illustrate how compression can be used in this study, we first describe the form in which we expect the data to appear. When the user performs an activity in the smart home, each user-interaction event contains information about (1) the time when the activity is performed, (2) the location of where the activity is performed and (3) the sensor that is being activated. Each user-interaction event is annotated (usually by the user themselves) with an activity name and a label stating the starts and ends of an activity. Table 1 shows an example of activity events in a smart home.

Table 1. An example of user-interaction events in a smart home.

Time	Location	Sensor ID	Activity	
8.02 a.m.	bathroom	S22	toileting	start
8.06 a.m.	bathroom	S3	toileting	end
8.31 a.m.	kitchen	S18	making coffee	start
8.33 a.m.	kitchen	S7	making coffee	
8.38 a.m.	kitchen	S5	making coffee	end
9.05 a.m.	bathroom	S22	toileting	start
9.17 a.m.	bathroom	S3	toileting	end

Since each activity event consists of information about time and location, this information can be incorporated when compressing the data stream. This paper extends the prediction by partial matching (PPM), an adaptive statistical data compression technique, to include spatial and temporal information. Our aim in this paper is to improve the activity recognition process by incorporating spatial and temporal context information, as illustrated in Figure 1. In particular, this paper aims to answer questions such as 'where will the user most likely be at a given time' or 'given that the user is in a particular location at a given time, what activity is the user most likely doing'. Evaluation was performed on three publicly available smart-home datasets.

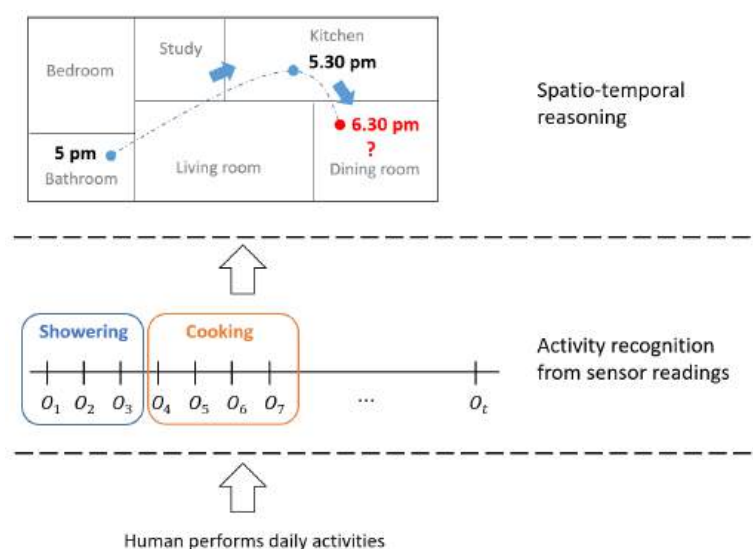


Figure 1. Spatio-temporal reasoning in smart homes: **(Bottom)** User in a smart home performs their daily activities. **(Middle)** Sensor readings are mapped to user activities. **(Top)** Using location and time information for reasoning.

This paper is structured as follows: Section 2 provides reviews of related work. Section 3 contains description of our approach. Section 4 details the datasets and evaluation method used in this study. Section 5 describes the experiments and Section 6 contains the experimental results and discussion. Section 7 concludes our findings.

2. Related Work

The work of Das, Cook and Bhattacharya [6] is among the earlier work that used compression for activity recognition. They partitioned the home into different zones, where each zone is represented by a symbol. A dictionary of user movements is trained using the LZ78 compression algorithm. They then applied prediction by partial matching (PPM) to predict user next location based on the phrases in the dictionary. Gopalratnam and Cook [7] proposed a sequential prediction algorithm called Active LeZi, to predict the user's next action based on an observed sequence of user-home interactions. Similar to the approach taken in [6], they built a representation of user-home interactions with LZ78 and used PPM to calculate the probability of the next most probable action.

In the work of Alam, Reaz and Ali [8], they introduced an adaptive sequence prediction algorithm that generates a finite-order Markov model and makes predictions based on PPM. To better capture the boundary between two opposite events, they applied an episode evaluation criteria that makes use of sensor ON and OFF states to indicate the start and end of an episode. An enhancement to this work is seen in [9] where they used a prefix tree-based data model to learn and predict user actions.

A variation to the work described above is the work of Chua, Marsland and Guesgen [10]. They used compression as an unsupervised learning method to identify activities in an unlabeled sensor stream. The sensor stream is represented as a sequence of characters. The Lempel–Ziv–Welch dictionary-based compression algorithm and edit distance are used to identify repeated sequences from sensor readings that represent user activities.

Most of these studies attempt to infer user activities from a sequence of sensor readings to either predict the user's next location or action. Our work differs from these studies as the spatio-temporal information and user activities are represented in tuples. The PPM model is built on these tuples and the trained model is used to make prediction.

There are works that used PPM to predict user locations using GPS trajectories. In the work of Neto, Baptista and Campelo [11], they combined Markov model and PPM for predicting route and destination. PPM is used to learn about the road segments traversed by the user, while Markov model is used to model the transitions between locations. When predicting user route and destination, Markov model first predicts the next user location based on current location, and the routes and destination are retrieved from PPM. Burbey and Martin [12] extends the PPM algorithm to include temporal and location information to predict user next location based on movement traces obtained from wireless access points. Our work is closely related to the work proposed in [12]. The main difference lies in the way the PPM is trained. Our approach trained the PPM based on user-interaction events rather than on the entire sequence of movement traces.

Another variation to the methods pointed out above is to use frequent pattern mining. Liu et al. [13] mined frequent temporal patterns from a sequence of user actions. The mined patterns are used to characterize activities. Nazerfard [14] combined association rule mining and expectation-maximization clustering to discover temporal features and relations of activities from sensor-based data. However, these methods do not use any spatial information to infer user activities. In the work of Liu et al. [15], they attempt to extract spatial and temporal features from sensor data. They first applied *k*-means to cluster the temporal features and then used spatial features, which include a set of sensors and their frequency, to build a prediction model in each temporal cluster. Such methods, however, require the temporal features obtained from sensor readings to be clustered first before any classification can be performed.

3. Our Proposed Method

This section describes the statistical-based text compression approach based on prediction by partial matching, and our approach in predicting user activities based on spatio-temporal information.

3.1. Prediction by Partial Matching (PPM)

The main idea of PPM is to use the last few characters to predict the next character in the input sequence [16]. PPM builds several finite-context models of order k adaptively, where k is the number of preceding characters used.

Table 2 illustrates the operation of PPM after input string ‘sensorsensor’ has been processed. All the previously seen contexts in each model ($k = 2, 1, 0$ and -1) are shown with their respective predictions, frequency counts c and the probabilities p . The lowest-level model is $k = -1$, which predicts all characters equally, where A refers to the set of distinct characters used. PPM records the frequency counts of each character seen for each context in the model, which is used to calculate the prediction probabilities.

By default, the model with the largest k is used when the PPM is queried, which in this example $k = 2$. When string ‘rs’ is seen, the likely next character is e , with a probability of 0.5. If a novel character is observed in this context, then an escape (‘esp’) event is activated, signifies a switch to a lower order model. This process is repeated until it reaches a model where the context is matched or the lowest model ($k = -1$) is reached. Suppose that o followed the string ‘rs’, which is not predicted from the model $k = 2$ in the context rs . An escape event occurs and $k = 1$ model of context s is used, i.e., through the prediction $s \rightarrow o$.

Table 2. PPM model after processing input string ‘sensorsensor’.

Order $k = 2$				Order $k = 1$				Order $k = 0$				Order $k = -1$			
Predictions		c	p	Predictions		c	p	Predictions		c	p	Predictions		c	p
en	$\rightarrow s$	2	$\frac{2}{3}$	e	$\rightarrow n$	2	$\frac{2}{3}$	$\rightarrow e$	2	$\frac{2}{17}$		$\rightarrow A$	1	$\frac{1}{ A }$	
	$\rightarrow esp$	1	$\frac{1}{3}$		$\rightarrow esp$	1	$\frac{1}{3}$		$\rightarrow n$	2	$\frac{2}{17}$				
									$\rightarrow o$	2	$\frac{2}{17}$				
ns	$\rightarrow o$	2	$\frac{2}{3}$	n	$\rightarrow s$	2	$\frac{2}{3}$		$\rightarrow r$	2	$\frac{2}{17}$				
	$\rightarrow esp$	1	$\frac{1}{3}$		$\rightarrow esp$	1	$\frac{1}{3}$		$\rightarrow s$	4	$\frac{4}{17}$				
									$\rightarrow esp$	5	$\frac{5}{17}$				
or	$\rightarrow s$	1	$\frac{1}{2}$	o	$\rightarrow r$	2	$\frac{2}{3}$								
	$\rightarrow esp$	1	$\frac{1}{2}$		$\rightarrow esp$	1	$\frac{1}{3}$								
rs	$\rightarrow e$	1	$\frac{1}{2}$	r	$\rightarrow s$	1	$\frac{1}{2}$								
	$\rightarrow esp$	1	$\frac{1}{2}$		$\rightarrow esp$	1	$\frac{1}{2}$								
se	$\rightarrow n$	2	$\frac{2}{3}$	s	$\rightarrow e$	2	$\frac{1}{3}$								
	$\rightarrow esp$	1	$\frac{1}{3}$		$\rightarrow o$	2	$\frac{1}{3}$								
					$\rightarrow esp$	2	$\frac{1}{3}$								
so	$\rightarrow r$	2	$\frac{2}{3}$												
	$\rightarrow esp$	1	$\frac{1}{3}$												

3.2. Description of Our Approach

Since our interest is to predict activities based on time and location, the sensor information, i.e., sensors that are being activated (column 3 in Table 1), are not considered. Each user-interaction event a_i is therefore represented by a triplet (X, Y, Z) in ASCII character, where X is the time, Y is the location and Z is the activity performed.

Referring to the example of input string ‘sensorsensor’, PPM builds the context models based on the number of preceding characters used. Assume that $a_1 = (s, e, n)$, $a_2 = (s, o, r)$, $a_3 = (s, e, n)$ and \dots , we want the highest context model ($k = 2$) to predict user activity (n in a_1) based on time and location

(s and e in a_1). To do this, PPM is trained on each a_i rather than on the entire sequence of input string. With this, the $k = 2$ model will have two predictions: (1) $se \rightarrow n$ and (2) $so \rightarrow r$.

One of the issues with temporal resolution is that the scale on which it is measured can in fact change the analysis. If a representation would to train on data where the user was in the bathroom showering at 9.15 a.m. and then head to the kitchen to make coffee at 9.50 a.m., the training data for these two events would be:

(9.15 a.m., bathroom, showering), (9.50 a.m., kitchen, making coffee), ...

If the test data indicated that the user was showering in the bathroom at 9.05 a.m., the model would not be able to make a prediction of the user location at 9.05 a.m. since the model does not have any information of the user showering in the bathroom precisely at 9.05 a.m. In view of this, the time is processed as 30 min intervals, i.e., mapping each ASCII character for each time interval of 30 min. When the PPM model is queried for a prediction of where the user is likely to be at 9.05 a.m., which falls within '9 a.m.–9.29 a.m.' interval, it can predict that the user is in the bathroom. Figure 2 shows the representation of Table 1 in ASCII characters and processed based on 30 min time intervals.

Time	Location	Activity		Time	Location	Activity
8.02 am	bathroom	toileting		a	x	p
8.06 am	bathroom	toileting		a	x	p
8.31 am	kitchen	making coffee		b	y	q
8.33 am	kitchen	making coffee		b	y	q
8.38 am	kitchen	making coffee		b	y	q
9.05 am	bathroom	toileting		c	x	p
9.17 am	bathroom	toileting		c	x	p

Figure 2. Representation of Table 1 in ASCII characters and processed based on 30 min time intervals.

4. Description of the Data and Evaluation Method

This study was run on three publicly available smart-home datasets:

- MIT PlaceLab Dataset [17] This dataset contains 16 days of user activities with 1805 user-interaction instances. The activities are grooming/dressing, doing/putting away laundry, toileting/showering, cleaning, preparing meal/beverages and washing/putting away dishes. These activities are performed in 4 different locations of the home i.e., kitchen, bathroom, bedroom and office/study.
- van Kasteren Dataset [18] This dataset contains 24 days of user activities with 1318 user-interaction instances. The activities are toileting/showering, going to bed, preparing meal/beverages and leaving house. These activities are performed in 4 different locations of the home i.e., living room, bedroom, bathroom and kitchen.
- Aruba Dataset [19] For this dataset, the November 2010 data is used, which contains 27 days of user activities with 3569 user-interaction instances. The activities are meal preparation, eating, working, sleeping, washing dishes and bed to toilet. These activities are performed in 7 different locations of the home i.e., kitchen, dining, living room, bedroom-1, bedroom-2, bathroom and office.

In all the datasets, the users annotated the activities themselves meaning that there were ground-truth annotations. It was observed that some activities were not repeated daily such as 'doing/putting away laundry', 'washing dishes' and 'cleaning' on MIT PlaceLab dataset. Since it is important that these activities are seen in the test set, leave-two-days-out cross-validation method is used, i.e., training on 14 days data and testing on the remaining 2 days. As for van Kasteren dataset,

since there is a small number of activity examples per activity, leave-four-days-out cross-validation method is used, i.e., training on 20 days data and testing on the remaining 4 days. Since the activities in Aruba dataset are repeated frequently, leave-nine-days-out cross-validation method is used, i.e., training on 18 days data and testing on the remaining 9 days. Figure 3 shows the cross-validation method applied in this study.

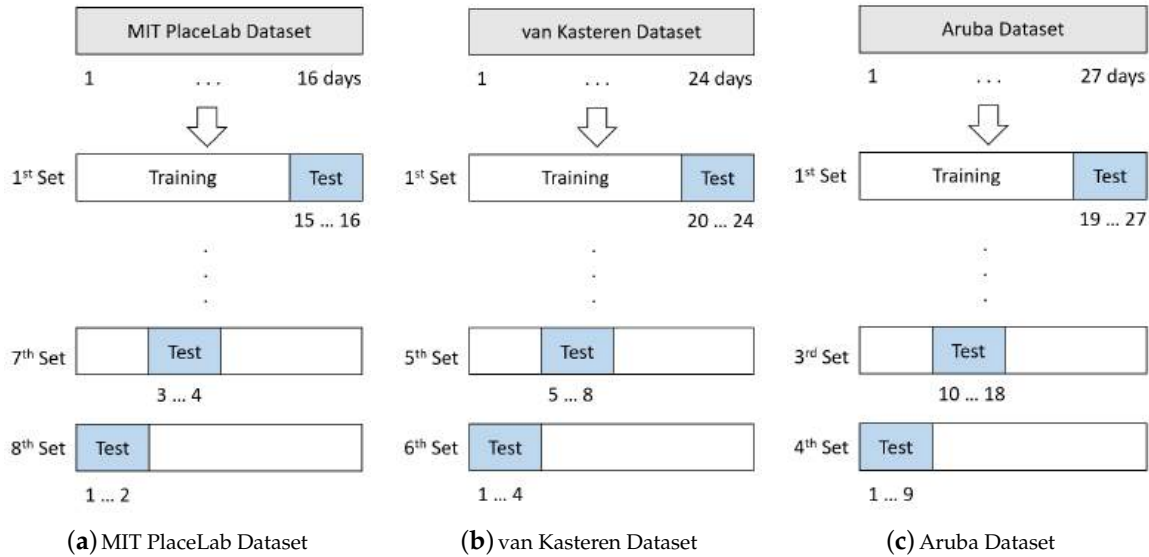


Figure 3. Evaluation method: (a) Leave-two-days-out cross-validation, (b) Leave-four-days-out cross-validation and (c) Leave-nine-days-out cross-validation.

5. Experiments

To evaluate the effectiveness of our approach, four experiments were conducted. The first experiment used $k = 1$ model for prediction, while the second experiment used $k = 2$ model. The third experiment evaluated the amount of training data needed to train the PPM model. For each evaluation, a new PPM model is created on each training set. During testing, context information such as user location and time will be fed to the PPM for prediction. The fourth experiment evaluated the computational performance.

5.1. Experiment 1: Prediction Based on First-Order Model

In this experiment, the first-order model (i.e., $k = 1$) is used for prediction. Two types of predictions were evaluated. The first predicts the location of the user given the time of the day ($time \rightarrow location$), while the second predicts user activity given the location of the user ($location \rightarrow activity$). For example, if the test data consists of the string 'sen', using the first-order model, it will first given 's' and see if it predicts 'e' and then given the context 'e' and see if it predicts 'n'.

5.2. Experiment 2: Prediction Based on Second-Order Model

This experiment aims to predict user activity based on time and location ($(time, location) \rightarrow activity$). Using the example of the string 'sen' (discussed in Section 5.1), the PPM model would be given the string 'se' and see if it predicts 'n'. In this experiment, the second-order model will be used for prediction.

5.3. Experiment 3: Training Size

To ensure that the PPM model acquires a good representation of user activity, it is important that each event is seen several times in the training set. The aim of this experiment is to determine the

amount of training data required to train the PPM. Different splits of the data were evaluated. Table 3 shows the number of days used for training and testing. These data were cross-validated.

Table 3. Different training splits on (a) MIT PlaceLab, (b) van Kasteren and (c) Aruba datasets.

Split	Number of Days	
	Training Set	Test Set
(a) MIT PlaceLab Dataset		
1	12 days	4 days
2	8 days	8 days
3	4 days	12 days
(b) van Kasteren Dataset		
1	16 days	8 days
2	12 days	12 days
3	8 days	16 days
4	4 days	20 days
(c) Aruba Dataset		
1	16 days	11 days
2	12 days	15 days
3	8 days	19 days
4	4 days	23 days

Both experiments 1 and 2 were repeated in this experiment on different training-test splits. Three types of predictions were analyzed. First is to predict user location based on time of the day ($time \rightarrow location$), while the second predicts user activity based on location ($location \rightarrow activity$). The third predicts user activity based on time and location, i.e., $(time, location) \rightarrow activity$.

5.4. Experiment 4: Computational Performance

The aim of this experiment is to evaluate the computational performance of our method in terms of training time and prediction time. The training time is computed based on the time it requires to build the PPM, while the prediction time is computed based on the time it takes to predict from the trained PPM. The performance is evaluated on a desktop computer with an Intel(R) Core(TM) CPU i7-7700K @ 4.2 GHz and 64 GB memory.

6. Results and Discussion

The first experiment was conducted using the first-order model for prediction. The main purpose is to test the model's ability in predicting user location given the time of the day ($time \rightarrow location$) and predicting user activity given the location of the user ($location \rightarrow activity$). The recognition accuracy is calculated based on the number of times the model correctly makes the prediction.

Table 4 shows the recognition results for ($time \rightarrow location$) prediction. Our method achieved an average recognition of 95.40% on MIT PlaceLab dataset, 90.41% on van Kasteren dataset and 84.04% on Aruba dataset when predicting user location based on time. Referring to Table 4, a low recognition accuracy of 71.94% was observed in the 2nd set of the van Kasteren dataset. In this test set, the user came home early around 1 pm, which was the only time and day that this was observed in the entire dataset. The PPM model is not able to make prediction since this event was not learned during training. Although the Aruba dataset has the lowest average recognition performance, only 18 days out of 27 days data were used for training, which is the lowest training:testing ratio as compared to the other two smart-home datasets.

Table 4. Recognition performance in predicting user location based on time for each test set.

Test Sets	Recognition Accuracy (%)		
	MIT PlaceLab Dataset	van Kasteren Dataset	Aruba Dataset
1st Set	88.07	92.82	96.79
2nd Set	92.78	71.94	75.24
3rd Set	97.18	98.32	79.94
4th Set	94.63	92.67	84.17
5th Set	96.31	96.69	–
6th Set	100	90.03	–
7th Set	98	–	–
8th Set	96.24	–	–
Average	95.40	90.41	84.04

Table 5 shows the recognition results in predicting user activity given location. A high average recognition accuracy was observed when predicting user activity based on location across all the datasets (i.e., 99.77% on MIT PlaceLab, 99.86% on van Kasteren and 98.89% on Aruba). This was expected since there are certain locations in the home where an activity usually takes place. For example, showering usually occurs in the bathroom, while cooking usually occurs in the kitchen. The results from this experiment showed that location provides important context for inferring user activity. The results from experiment 1 showed that our approach is effective in predicting (*time* \rightarrow *location*) and (*location* \rightarrow *activity*) with more than 84% average recognition performance for (*time* \rightarrow *location*) and more than 98% average recognition performance for (*location* \rightarrow *activity*).

Table 5. Recognition performance in predicting user activity based on location for each test set.

Test Sets	Recognition Accuracy (%)		
	MIT PlaceLab Dataset	van Kasteren Dataset	Aruba Dataset
1st Set	99.39	100	100
2nd Set	100	100	95.57
3rd Set	100	100	100
4th Set	99.17	100	100
5th Set	99.59	99.17	–
6th Set	100	100	–
7th Set	100	–	–
8th Set	100	–	–
Average	99.77	99.86	98.89

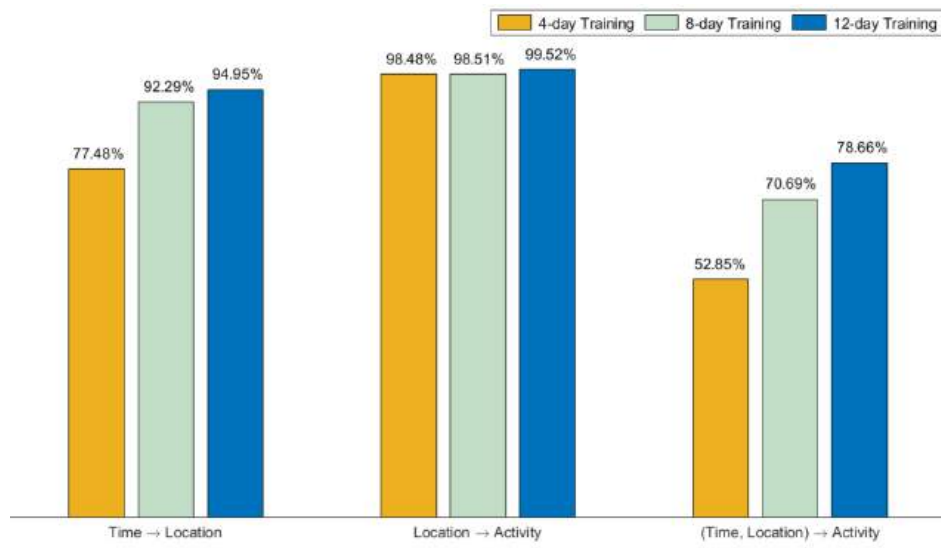
The second experiment tests the model's ability to predict user activity given the time and location (*(time, location) \rightarrow activity*). This experiment used the second-order model for prediction. The results are shown in Table 6. An average accuracy of 81.74% is achieved on MIT PlaceLab dataset, 88.14% on van Kasteren dataset and 81.05% on Aruba dataset. A lower recognition accuracy was observed in this experiment compared to the first experiment. This was due to variations in the activities performed on different time of the day. Although the accuracy was lower compared to the first experiment, our method still able to achieve an average recognition of more than 81% on all the three datasets. The 1st test set of MIT PlaceLab dataset has a low recognition accuracy of 63%. This was due to the variations in the way that the user performed his activity in this test set. For example, the second-order model learned that the user will be in the kitchen between 9.30 a.m. and 10 a.m. to 'prepare meal'. In this particular test set, the user was 'doing laundry'.

Table 6. Recognition performance in predicting activity based on time and location on each test set.

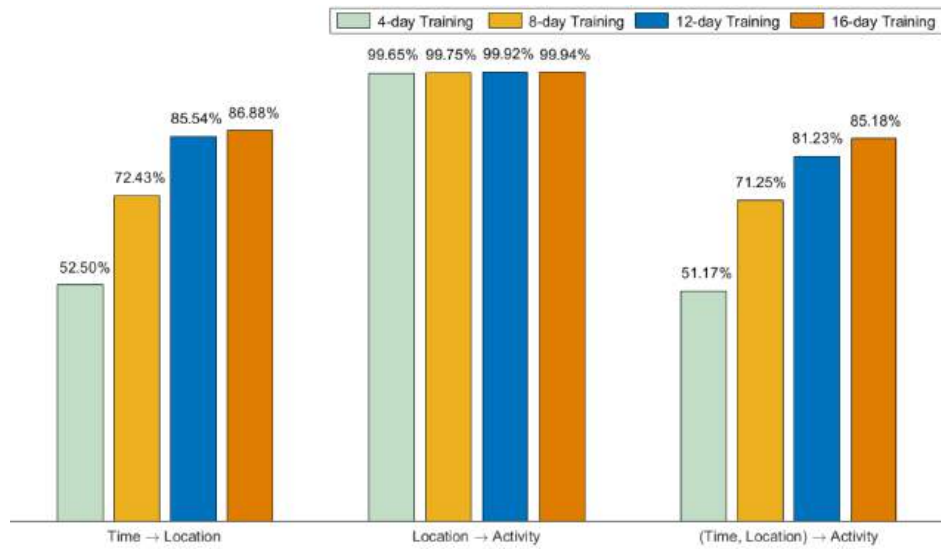
Test Sets	Recognition Accuracy (%)		
	MIT PlaceLab Dataset	van Kasteren Dataset	Aruba Dataset
1st Set	63	91.16	96.20
2nd Set	82.22	68.96	72.62
3rd Set	91.55	97.19	76.17
4th Set	79.75	89.66	79.22
5th Set	81.56	92.56	–
6th Set	94.02	89.30	–
7th Set	81.38	–	–
8th Set	80.45	–	–
Average	81.74	88.14	81.05

The third experiment was conducted to determine the amount of training data required to train the PPM. Various training-test splits were examined. The results presented in Figure 4 showed that the size of training data does have an impact on recognition performance. Such results are expected since compression is more effective when patterns are repeated frequently. When trained on 4 days, an average accuracy of 77.48% is achieved on MIT PlaceLab, 52.50% on van Kasteren dataset and 60.49% on Aruba dataset for *time* \rightarrow *location* prediction. A lower accuracy was observed in van Kasteren dataset. This was due to variations in the time when the user performed the activities. Such variations were not repeated frequent enough for PPM to learn the representations. For *(time, location)* \rightarrow *activity* prediction, an average accuracy of 52.85% is achieved on MIT PlaceLab, 51.17% on van Kasteren dataset and 57.15% on Aruba dataset. However, for both *time* \rightarrow *location* and *(time, location)* \rightarrow *activity* predictions, the average accuracy increases when trained with more training data, as shown in the first and third grouped bar chart in Figure 4. For *location* \rightarrow *activity* prediction, our method achieved a high recognition accuracy across the three datasets. A consistent performance of more than 91% average recognition accuracy was observed across the different training-tests splits. The results showed that the size of training data does not have much impact on *location* \rightarrow *activity* prediction. The high recognition performance achieved in this experiment (and also in the first experiment) showed that knowing which room that the user is in provides a better inference of what activity the user might be involved in.

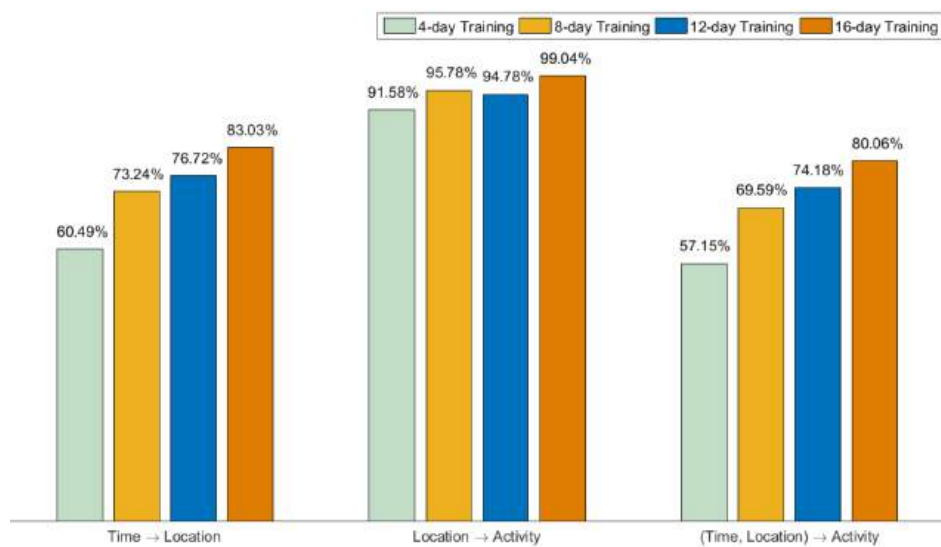
The fourth experiment evaluates the time it takes to train the PPM model and the time to predict from the trained PPM. Table 7 shows the computational performance in terms of (a) training time (in minutes) and (b) prediction time (in seconds). The value in parentheses shows the number of user-interaction instances in each respective training/test set. The training time across the number of instances presented in Figure 5 clearly show that the training time grows approximately linearly with the number of instances. As for the prediction time, it takes 0.1216 s to predict 1377 instances on the second test set of Aruba dataset (see Table 7b), which has the highest number of instances among all the test sets. The prediction time reduces when test on a smaller number of instances.



(a) MIT PlaceLab Dataset



(b) van Kasteren Dataset



(c) Aruba Dataset

Figure 4. Recognition performance on different training-test splits.

Table 7. Computational time: (a) Training time (in minutes) and (b) Prediction time (in seconds). The value in parentheses represents the number of user-interaction instances in each respective training/test set.

(a) Training Time (In Minutes)			
Training Sets	MIT PlaceLab Dataset	van Kasteren Dataset	Aruba Dataset
1st Set	11.34 (1478)	5.68 (1137)	53.88 (2728)
2nd Set	14.38 (1625)	3.90 (983)	30.14 (2192)
3rd Set	13.77 (1592)	5.84 (1140)	31.17 (2218)
4th Set	12.95 (1563)	5.21 (1086)	38.95 (2419)
5th Set	13.05 (1561)	6.63 (1197)	–
6th Set	16.08 (1688)	4.80 (1047)	–
7th Set	11.31 (1456)	–	–
8th Set	15.36 (1672)	–	–
Average	13.53 (1579)	5.34 (1098)	38.54 (2389)
(b) Prediction Time (In Seconds)			
Test Sets	MIT PlaceLab Dataset	van Kasteren Dataset	Aruba Dataset
1st Set	0.0275 (327)	0.0146 (181)	0.0674 (841)
2nd Set	0.0154 (180)	0.0248 (335)	0.1216 (1377)
3rd Set	0.0172 (213)	0.0152 (178)	0.1191 (1351)
4th Set	0.0234 (242)	0.0197 (232)	0.0842 (1150)
5th Set	0.0212 (244)	0.0098 (121)	–
6th Set	0.0080 (117)	0.0216 (271)	–
7th Set	0.0406 (349)	–	–
8th Set	0.0100 (133)	–	–
Average	0.0204 (226)	0.0176 (220)	0.0981 (1180)

In this study, the temporal information is processed based on a 30 min interval. This time interval is a preliminary choice to validate our approach. However, a more adequate interval could be determined by preprocessing the data and finding the most suitable interval for each activity. Methods that could potentially be applied include rough and fuzzy sets [20].

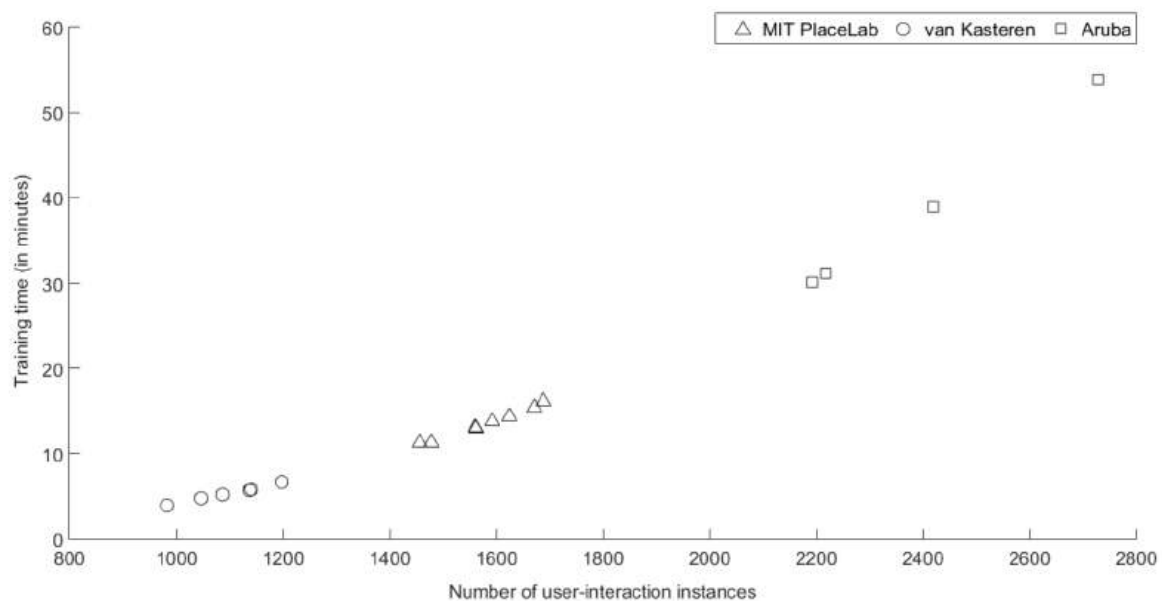


Figure 5. Training time (in minutes) across the number of user-interaction instances on three datasets (MIT PlaceLab, van Kasteren and Aruba).

7. Conclusions

This paper shows the importance of spatial and temporal information in interpreting human activity and how such information can be represented for activity recognition. The prediction by partial matching method is extended to capture spatio-temporal information by exploiting the repetitions from activity events. Evaluation was performed on three publicly available smart-home datasets. Our method can achieve an average accuracy of more than 84% for *time* \rightarrow *location* prediction. For *location* \rightarrow *activity* prediction, our method achieved more than 98% average accuracy across all the three datasets. Although the *(time, location)* \rightarrow *activity* has a lower recognition performance, our method can achieve an average accuracy of more than 81%. The results showed that the size of training data has an impact on the recognition performance for *time* \rightarrow *location* and *(time, location)* \rightarrow *activity* predictions. Compression tends to be more effective when trained with more data and the training time grows approximately linearly with the number of instances. The results from the experiments showed that location provides useful context information for inferring user activity. As future work, the plan is to extend our approach for abnormality detection. The learned PPM can be used to identify inputs that do not fit into the contexts.

Author Contributions: Conceptualization, S.-L.C. and H.W.G.; Data curation, S.-L.C. and L.K.F.; Formal analysis, S.-L.C. and L.K.F.; Methodology, S.-L.C. and L.K.F.; Writing—original draft, S.-L.C. and L.K.F.; Writing—review and editing, L.K.F., H.W.G. and S.-L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Du, Y.; Lim, Y.; Tan, Y. A novel human activity recognition and prediction in smart home based on interaction. *Sensors* **2019**, *19*, 4474. [CrossRef] [PubMed]
2. Lu, L.; Cai, Q.-L.; Zhan, Y.-J. Activity recognition in smart homes. *Multimed Tools Appl.* **2017**, *76*, 24203–24220. [CrossRef]
3. Chua, S.-L.; Marsland, S.; Guesgen, H. A supervised learning approach for behaviour recognition in smart homes. *J. Ambient Intell. Smart. Environ.* **2016**, *8*, 259–271. [CrossRef]

4. Gochoo, M.; Tan, T.-H.; Liu, S.-H.; Jean, F.-R.; Alnajjar, F.S.; Huang, S.-C. Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 693–702. [CrossRef]
5. Singh, D.; Merdivan, E.; Hanke, S.; Kropf, J.; Geist, M.; Holzinger, A. Convolutional and Recurrent Neural Networks for Activity Recognition in Smart Environment. In *Towards Integrative Machine Learning and Knowledge Extraction*; Holzinger, A., Goebel, R., Ferri, M., Palade, V., Eds.; Springer: Cham, Switzerland, 2017; pp. 194–205.
6. Das, S.K.; Cook, D.J.; Bhattacharya, A. The role of prediction algorithms in the MavHome smart home architecture. *IEEE Wirel. Commun.* **2002**, *9*, 77–84. [CrossRef]
7. Gopalratnam, K.; Cook, D.J. Online sequential prediction via incremental parsing: The active LeZi algorithm. *IEEE Intell. Syst.* **2007**, *22*, 52–58. [CrossRef]
8. Alam, M.R.; Reaz, M.B.I.; Ali, M.A.M. SPEED: An inhabitant activity prediction algorithm for smart homes. *IEEE Trans. Syst. Man. Cybern. A Syst. Hum.* **2012**, *42*, 985–990. [CrossRef]
9. Farayez, A.; Reaz, M.B.I.; Arsad, N. SPADE: Activity prediction in smart homes using prefix tree based context generation. *IEEE Access* **2019**, *7*, 5492–5501. [CrossRef]
10. Chua, S.-L.; Marsland, S.; Guesgen, H. Unsupervised learning of human behaviours. In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011; pp. 319–324.
11. Neto, F.D.N.; Baptista, C.S.; Campelo, C.E.C. Combining Markov model and prediction by partial matching compression technique for route and destination prediction. *Knowl. Based Syst.* **2018**, *154*, 81–92. [CrossRef]
12. Burbey, I.; Martin, T.L. Predicting Future Locations Using Prediction-by-Partial-Match. Available online: <https://dl.acm.org/doi/abs/10.1145/1410012.1410014> (accessed on 21 July 2020).
13. Liu, Y.; Nie, L.; Liu, L.; Rosenblum, D.S. From action to activity: Sensor-based activity recognition. *Neurocomputing* **2016**, *181*, 108–115. [CrossRef]
14. Nazerfard, E. Temporal Features and Relations Discovery of Activities from Sensor Data. Available online: <https://link.springer.com/article/10.1007/s12652-018-0855-7> (accessed on 12 November 2020).
15. Liu, Y.; Ouyang, D.; Liu, Y.; Chen, R. A novel approach based on time cluster for activity recognition of daily living in smart homes. *Symmetry* **2017**, *9*, 212. [CrossRef]
16. Cleary, J.G.; Teahan, W.J.; Witten, I.H. Unbounded length contexts for PPM. In Proceedings of the DCC'95 Data Compression Conference, Snowbird, UT, USA, 28–30 March 1995; pp. 52–61.
17. Tapia, E.M.; Intille, S.S.; Larson, K. Activity recognition in the home using simple and ubiquitous sensors. In Proceedings of the 2nd International Conference on Pervasive, Vienna, Austria, 21–23 April 2004; pp. 158–175.
18. van Kasteren, T.; Noulas, A.; Engleblenne, G.; Kröse, B. Accurate activity recognition in a home setting. In Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, Korea, 21–24 September 2008; pp. 1–9.
19. Cook, D.J. Learning setting-generalized activity models for smart spaces. *IEEE Intell. Syst.* **2012**, *27*, 32–38. [CrossRef] [PubMed]
20. Guesgen, H.W. Using rough sets to improve activity recognition based on sensor data. *Sensors* **2020**, *20*, 1779. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Monitoring and Support for Elderly People Using LoRa Communication Technologies: IoT Concepts and Applications

José Paulo Lousado ^{1,*}  and Sandra Antunes ² ¹ Research Centre in Digital Services (CISeD), Polytechnic Institute of Viseu, 3504-510 Viseu, Portugal² Center for Studies in Education and Innovation (CI & DEI), Polytechnic Institute of Viseu, 3504-510 Viseu, Portugal; santunes@estgl.ipv.pt

* Correspondence: jlousado@estgl.ipv.pt; Tel.: +351-254-615-477

Received: 11 October 2020; Accepted: 18 November 2020; Published: 20 November 2020

Abstract: The pandemic declared by the World Health Organization due to the SARS-CoV-2 virus (COVID-19) awakened us to a reality that most of us were previously unaware of—isolation, confinement and the massive use of information and communication technologies, as well as increased knowledge of the difficulties and limitations of their use. This article focuses on the rapid implementation of low-cost technologies, which allow us to answer a fundamental question: how can near real-time monitoring and follow-up of the elderly and their health conditions, as well as their homes, especially for those living in isolated and remote areas, be provided within their care and protect them from risky events? The system proposed here as a proof of concept uses low-cost devices for communication and data processing, supported by Long-Range (LoRa) technology and connection to The Things Network, incorporating various sensors, both personal and in the residence, allowing family members, neighbors and authorized entities, including security forces, to have access to the health condition of system users and the habitability of their homes, as well as their urgent needs, thus evidencing that it is possible, using low-cost systems, to implement sensor networks for monitoring the elderly using the LoRa gateway and other support infrastructures.

Keywords: internet of things; LoRaWAN; COVID-19; ICT; The Things Network; ESP32 microcontroller

1. Introduction

Since the beginning of the SARS-CoV-2 pandemic, a virus discovered in 2019 [1], one of the fundamental concerns was the elderly population, namely due to the impact that the disease caused by the new coronavirus could have on the population in this age group (65 years old or more) [2].

Until then, several solutions for the surveillance of the elderly in a residential context had been advanced by the electronic industry, information technology and entities related to the protection of property and the security of people, allied to well-being and home automation [3–8]. However, the reality, in the current context, shows us that the proliferation of proposals in computer and telecommunications systems for monitoring and supporting the elderly population falls far short of what is desired. The ageing population, living in remote regions, has been exposed to the cruelest conditions of abandonment, without access to medicines, without means of communication, exacerbated by the fact that, in many areas of Portugal, there is no mobile network coverage or, if there is, it has a deficient signal. For these citizens, everything became more distant. Thus, based on this reality as a motivation for the present work, the following question arose: how can the health status and living conditions of the elderly population, dispersed in rural areas with low or no mobile network coverage at all, be remotely monitored using low-cost technologies? To answer that question, several situations need to be considered.

The emergence of the Internet of Things (IoT), currently present in several home systems, notably in small devices for regular use, such as a blood pressure meter, but also in larger equipment, including photovoltaic panels, household appliances, consumption and energy efficiency controllers, among others [9], has extended the application spectrum of data communication networks to other sectors. The health and well-being area is also one of the areas that has benefited the most from this type of technology [10], which is why its exploitation and use in the current context of the pandemic for the benefit of the most disadvantaged populations, in particular the elderly population living in rural areas, becomes imperative.

With the technology currently available, it is possible to combine devices with heterogeneous systems, such as smartphones with mobile networks (3G/4G and, in the future, 5G), Bluetooth devices, wireless networks, sensors, among others, allowing these devices to interact with one another and provide fully automated, adaptive operating environments, taking advantage of these infrastructures and being able to contribute to the improvement of people's quality of life. In [11,12], the authors present some models for the use of low-consumption and long-range networks for home and industry automation, respectively, using Long-Range (LoRa) communication technologies. These communication networks are essential to disseminate data and for its analysis, without necessarily resorting to the Internet, as they are able to collect data from various sensors, maintaining their activity for a long period of time, since the consumption of devices and sensors is reduced. The communication of these data uses a LoRa Wide Area Network (LoRaWAN) gateway, which can be up to 15 km away from the LoRa node that sends the data to the respective Internet connection device [13].

The emergence and expansion of smart cities is an excellent example of the use of the Internet of Things and the use of Artificial Intelligence, in which ubiquitous computing systems are collecting and generating huge amounts of data daily that not require only a storage location, using Cloud Storage, but also immediate processing, helping citizens to take advantage of these data [14]. Decision support systems can also complement the analysis of people's health status, such as infrared body temperature screening at airports and other places of public circulation, being able to detect people who may be suspected of suffering from some pathology that poses a danger of contagion, as repeatedly observed in the media in the current context of the pandemic [15].

Several reports published before the COVID-19 pandemic show that the proportion of those classed as part of the ageing population is increasing in Portugal, without support units being able to provide an efficient and timely response to all requests [16]; thus, one of the solutions will be to keep people in their homes as long as they can be properly followed and monitored, keeping them in their comfort zone. In this way, elderly people will feel more comfortable in their residence, maintaining their habits and routines, a situation that contributes to active and healthy aging. Nevertheless, recent studies show that those suffering or who will suffer from some type of mental disorder is growing considerably throughout the world, so it is imperative to assess the existing technologies for the benefit of the people, minimizing the negative impact that these pathologies have on their quality of life [17,18]. In the context of social isolation and confinement caused by COVID-19, this became even more evident, with several studies proving that these measures are risk factors for the health of the population and for the elderly population in particular [2,19].

Based on this reality, using the concepts already defined in other monitoring and follow-up environments based on miniaturized sensors and telecommunications equipment, namely the work referred to in [20], we present, in this article, an answer to our question, with a system model for following, monitoring and protecting old people who are in a stable state of health, allowing them to maintain their autonomy within their homes and eventually abroad.

The main goal of the proposed system is, therefore, the monitoring of the health status of elderly people who are in their homes, sometimes a few kilometers away from support centers and local community centers (Parish/County), in a discreet and non-intrusive way, through information and communication technologies. Thus, both support institutions, as well as family members, friends, or other entities, can monitor the status of these people in real time. Considering that systems aimed at

monitoring and managing people's data are currently undergoing great evolution, and that pervasive and ubiquitous computing is already part of everyday life, with this work, the authors hope to contribute positively to improving people's quality of life, especially in the senior population with low economic resources who are still in a state of health that allows them to maintain their autonomy.

2. Related Work and Technologies

Several authors have addressed the issue of monitoring people and their health status remotely. In answer to this question, several low-cost technologies were selected that allow us to ensure optimal reference values in various metrics, namely reliability, quality of service (QoS) and total cost of ownership (TCO). To this end, this chapter provides to further analysis and a literature review on the subject, to provide a more comprehensive view of long-range communication technologies and their applications. The most promising areas in the use of remote monitoring technologies are, naturally, mobile communications—3G, 4G and, in the near future, 5G—but also the use of wireless communication technologies such as LoRa, SigFox, Weightless-N/Nwave, Long Term Evolution for Machines (LTE-M) and Narrowband IoT (NB-IoT), among others, these being the most relevant, as mentioned and analyzed in [21], and the TV whitespace (TVWS) analyzed in [22].

In view of the panoply of similar technologies with applications in remote monitoring, it is difficult from the outset to select one that best meets the requirements defined in order to answer our question. Thus, based on several studies published in scientific journals, the authors selected those that, in general, could answer our question. However, there are restrictions that may lead us to choose one technology over another due to a set of technical requirements at the outset, namely low acquisition cost, low energy consumption, ease of implementation, robustness and availability in the marketplace.

2.1. Long-Range (LoRa)

LoRa is a technology of wireless communication networks (radio frequency), which allows the communication of thousands of devices powered by batteries, over long distances and with a minimum consumption of energy. LoRa technology is part of a grouping of networks called Low Power Wide Area Networks (LPWANs), capable of communicating over long distances, even in adverse conditions, because of their simple way of organizing information [23]. LoRa's low energy consumption is essential for integration into devices that are intended to be installed over a long period of time and powered by a battery, while, for thousands of devices to communicate, the efficiency of the network and the use of a radio frequency spectrum is important to ensure that no information is lost. LoRa technology uses unlicensed Industrial, Scientific and Medical (ISM) bands, i.e., 868 MHz in Europe, 915 MHz in North America and 433 MHz in Asia. Bidirectional communication is provided by Chirp Spread Spectrum (CSS) modulation, which spreads a narrow band signal over a wider channel bandwidth [24].

There are several works and applications of LoRa networks in the context of remote monitoring, namely in [25], where the authors propose an advanced architecture combining edge computing, fog computing, LoRa and other technologies based on IoT. The proposed architecture can help to overcome the limitations of existing IoT-based health monitoring systems (for example, drop detection or IoT-based electrocardiogram (ECG) monitoring systems) and satisfy the requirements of high data rate applications and the regulation of the LoRa work cycle, demonstrating the functionality of the proposed architecture through the presentation of a case study involving fall detection.

The work presented in [26] shows the advantages and disadvantages of current communication systems and technologies, proposing new IoT architectures in the medical field, dedicated to home and hospital care services, based on LoRa technology.

In [27], the authors studied the internal performance of LPWAN LoRa technology, using measurements in the context of real life. Measurements were performed using commercially available equipment on the main campus of the University of Oulu to test the suitability of LoRa LPWAN technology for health and well-being monitoring. In the study, authors analyzed the performance

of LoRa communications used to monitor a person's well-being in the workplace during normal working days.

In the study presented in [28], the authors show an irrigation monitoring system with practical application in precision agriculture on a Czech Republic farm, using LoRa networks, while evidencing the potential of IoT, in the case using LoRaWAN, in helping farmers, namely in irrigation control.

2.2. TV Whitespace

TV whitespace (TVWS) refers to TV channels located between frequency bands not used for TV broadcasting in certain regions. TVWS are parts of the radio frequency spectrum not used by transmission, also called interleaved spectra [22]. In global, TVWS are also referred to as currently unoccupied portions of the spectrum in the terrestrial region in television frequency bands in the Very High Frequency (VHF) and Ultra High Frequency (UHF) TV spectra (either analogue or digital, especially in the UHF band). In a simpler way, the TVWS spectrum represents a large part of the UHF spectrum (300 MHz–3 GHz), that is, hundreds of MHz, which in some countries also include VHF, which is available in a specific geographic region and can be used in a shared way. This spectrum can be used by primary (licensed) users or by secondary users who, using non-licensed equipment, can share the spectrum with digital TV transmitters, among other types of users. The amount of terrestrial whitespace available depends on several factors, such as geographical characteristics, the level of potential interference in the incumbent TV broadcast service, TV coverage objectives and related planning and use of television channels [29].

In the work presented in [22], the authors refer to the importance of TVWS, focusing their study on the application component and key areas in the application field. Starting by proving that TVWS has excellent penetration in buildings and good propagation characteristics, which, in turn, makes a TV band an innovative platform with great potential in a wide range of important applications, whether used indoors or abroad, the authors show that it is of great interest to investigate not only the quantity of TVWS and characterize its main properties, but also to evaluate the real applications of TVWS in reality. The TVWS use cases discussed in the study are particularly focused on wireless broadband access in rural environments, future wireless home networks, WLAN wireless services and smart grid network/smart meter communication.

In [30], several pilot projects are presented, namely in Africa, Europe, Asia and North America, mostly in rural areas [31], showing great potential. However, after several tests, the question of the applicability of TVWS was left unanswered. It is not clear why the TVWS tests were defined years ago, but they did not result in any commercial applications. This is relevant when considering the power of the restricted market for telecommunications operators, implying that both governments and regulators are not interested in the implementation of this technology. For example, in 2013–14, there was a movement to implement Microsoft-funded TVWS in Bangladesh, yet no regulatory movement was expressed by the Bangladeshi government in regard to TVWS at that time. The reason believed to be behind this decision concerns all operators being busy with their 4G licensing during that period. Adding to this problem is the fact that the commercial deployment of TV blanks, especially 470–698 MHz, are not allowed to be used for research purposes in many countries, as they present a risk of security and interference with other sectors of commercial activity [30].

The works presented in [32,33] focus on the need for the existence of a geographic database of previous TVWS available in different countries, since the frequencies available are different from country to country and within countries (from region to region), each of which has their own policies and different regulatory regimes. It is therefore necessary to safeguard the spectrum of frequencies that are used by security forces, emergency and commercial entities, without any type of interference.

In [34], the authors describe external field measurements in TVWS carried out in Munich, Germany. Fixed and mobile measurements in rural, suburban and urban settings showed that the model presented is appropriate to describe the path loss over distances of up to a few kilometers and that they can be used in the process of filling a geolocation database. This work had the contribution of the European

project ICT-COGEU (COgnitive radio systems for efficient sharing of TV white spaces in EUropean context), whose website is currently offline.

In Portugal, the process of converting analog TV to digital TV started in 2012, and the Portuguese entities recently changed the frequencies of digital terrestrial television broadcasters to new frequencies in order to free up the space previously occupied for future 5G networks. This process is expected to be completed by the end of December 2020 [35,36].

The use of TVWS is of great interest to the scientific community, namely for communications on LPWAN and long-range networks, but there seems to be a lack of investment in this technology, namely by the current players in the telecommunications market. As an example, the most recent document on this subject published by the authority responsible for the regulation of communication policies in Portugal, ANACOM (Autoridade Nacional de Comunicações), is dated August 2016 [37].

2.3. SigFox

SigFox is an LPWAN network operator that offers a complete IoT connectivity solution based on its patented technologies. SigFox deploys its base stations with equipment previously configured with proprietary software and connects them to the back-end servers using an IP-based network. End devices are connected to these base stations using phase-shift keying (BPSK) modulation on an ultra-narrow band (100 Hz) sub-GHz ISM carrier. Like LoRa technology, SigFox uses unlicensed ISM bands, for example, 868 MHz in Europe, 915 MHz in North America and 433 MHz in Asia. By using an ultra-narrow band, SigFox uses frequency bandwidth efficiently and achieves very low noise levels, leading to very low power consumption, ensuring the high sensitivity of the receiver and low cost antenna design at the expense of a maximum transfer rate of just 100 bps. SigFox initially supported only uplink communication, but later evolved into bidirectional technology with significant link asymmetry. Downlink communication, that is, data derived from base stations to end devices, can only occur after an uplink communication. The number of messages per uplink is limited to 140 messages per day. The maximum payload length for each uplink message is 12 bytes, but the number of messages in the downlink is limited to four per day, meaning that confirmation of each uplink message is not supported [38].

In [39], the authors show the use of a system based on SigFox networks, with applications in agriculture, for monitoring environmental factors. In this article, they present SigFox technology, as well as how this type of communication would be integrated into precision agriculture, while referring to other technologies already in use in this field. The authors concluded that SigFox and LPWAN technologies represent the future of IoT. Regardless the domain in which it is used, the IoT finds its applicability, leading researchers and developers to find and implement new solutions in order to increase its performance, productivity, and market value.

2.4. Weightless-N/NWave

NWave technology uses advanced demodulation techniques to allow a network to coexist with other radio technologies without additional noise. This proprietary technology is particularly aimed at the smart parking sensor monitoring market, where it has found a considerable market niche [40].

In [40], a comparative study of the three LoRa technologies, Xbee Pro (XBee868) and NWave, with LoRa technology appearing to have a slight advantage, is also presented. In the context of this study, specialized hardware was created to incorporate the different technologies and provide quantitative and qualitative scientific information related to data rates, success rates, modes of energy transmission and energy consumption and communication ranges.

2.5. XBee868LP/ZigBee

ZigBee communication technology uses a low data communication rate, low power consumption, and operates with a wireless network protocol aimed at computer applications and remote control. It has a low power specification based on the IEEE 802.15.4—2003 Wireless Personal Area Networks

standard, whose distance does not exceed 150 m [41]. The XBee 868LP (Low Power) is designed to provide a long-range radio frequency connection with significant performance and low power consumption. The modules have 30 channels between the frequencies 863 MHz and 870 MHz in the “Listen Before Talk” mode, which frees them from a work cycle. In [42], the authors refer that the Xbee868LP module is the first Radio frequency (RF) module in the industry to use Listen Before Talk and Adaptive Frequency Agility (LBT + AFA) techniques. The module “listens” to the environment before communicating. If disturbed, it automatically changes channels in a matter of microseconds, which does not affect its overall performance. With Surface Mount (SMT) technology, the XBee 868LP is compatible with the XBee ecosystem. The configuration is also carried out with the free software XCTU, a platform common to all products in the XBee ranges. Point to point, point to multipoint and DigiMesh networks are supported. The XBee868LP module allows communications up to 4 Km [42].

In [41], the authors show how a network of sensors can be implemented to monitor the doors of a building using ZigBee.

2.6. LTE-M

LTE-M technology (also LTE-Machine Type Communication (MTC) and LTE Cat M) also operates as an LPWAN, which allows for the reuse of an installed base LTE (mobile network) with extended coverage. LTE-M, which stands for LTE-Machine Type Communication (MTC), is also an LPWAN technology developed by 3GPP to enable devices and services specifically for IoT applications. LTE-M offers a data rate of 1 Mbps for 3GPP Release 13, increasing to 4 Mbps for Release 14, leading to greater mobility and voice capacity on the network [43].

2.7. NB-IoT

Narrowband IoT (NB-IoT) technology is also a radio technology deployed in mobile networks that is especially suitable for indoor coverage, low cost and long battery life for a large number of devices. NB-IoT limits bandwidth to a single narrow band of 200 kHz, offering maximum downlink speeds of 26 kbs in version 13 of the 3GPP standard. Version 14 will see this increase to 127 kbps. Both LTE-M technology and NB-IoT operate over a mobile network, requiring coverage with a sufficient signal [43]. All Global System for Mobile communications (GSM) cells that work with LTE can also support NB-IoT, but this requires new protocol installation and licensee fees, so not all operators provide it by default. It is crucial to check if the local GSM operator offers NB-IoT. Moreover, the Subscriber Identity Module (SIM) card must have this protocol enabled. SIMs with LTE may or may not work with NB-IoT—this depends on the GSM operator [44].

In [45], the authors present a comparative study of the different technology applications in the health care area, namely SigFox, LoRaWAN and NB-IoT.

In [46], the authors present a study related to health care, particularly the remote development of rural regions and the application of IoT in these regions for remote health monitoring based on NB-IoT technology. They feature an intelligent IoT-based edge system for remote health monitoring, in which vital wearable sensors transmit data and alerts to an IoT system. The collected data and alerts are then sent to doctors based on a risk-stratified push/pull protocol using the best combination of cellular/mobile/NB-IoT networks. Clinical validation through implantation at the hospital where the system was tested and remote telemedicine location demonstrated that the NB-IoT-based system can be a low-cost, yet feature-rich alternative and that it adds value to devices for remote patient monitoring.

2.8. Analysis and Decision

Several authors present comparative studies of different technologies, namely [47,48], who contributed to the decision regarding the technology to be used in our work. Moreover, in [49], a technical comparison of LoRaWAN and NB-IoT can be found, explaining that LoRaWAN is an open LPWAN system architecture developed and standardized by LoRa Alliance, a non-profit association of more than 500 member companies that operates in the unlicensed spectrum, while, in opposition, the NB-IoT

operates in the licensed spectrum. While both technologies can compete on QoS, IoT applications that require more frequent communications are better served by NB-IoT, which has no duty cycle limitations operating in the licensed spectrum, at the expense of higher TCO relative to LoRaWAN.

We elaborate on our analysis of the options offered by the two main long-range technologies—with the use of mobile networks vs. without the use of mobile networks—in Table 1, which summarizes the main characteristics of the two best options. When there is no mobile network coverage, LoRa technologies were considered the best option due to the several advantages over other technologies, the wide use, robustness, low cost, great ease and availability of equipment and also because they allow total customization and the system can be built entirely from scratch, and integrated into The Things Network. Alternatively, when using mobile networks, it is understood that NB-IoT technology is the one that can best meet the requirements, considering that it can be operated on the future 5G network, when globally available. However, this is not our focus in the present research work, since the studied areas are remote, rural and either do not have mobile network coverage or have poor signals.

The need for a project based on LoRa networks of low consumption, low acquisition cost and long reach is precisely related to the absence of mobile communications networks in the targeted regions, excluding any solution that implies the use of mobile networks. The existence of mobile network coverage would make possible other solutions. The possibility of using TVWS seems to be a distant reality; nevertheless, the results obtained in [30–34] are promising, as long as guaranteed commitment from the agents involved and the regulatory entities can be provided.

Advocated in this analysis, as well as in the works presented in [47–49], the authors conclude that LoRa technology supported by the LoRaWAN architecture is the one that best meets the requirements.

Table 1. Technology summary comparison: Long-Range Wide Area Network (LoRaWAN) vs. Narrowband Internet of Things (NB-IoT) (source: [49]).

Technology Parameters	LoRaWAN	NB-IoT
Bandwidth	125 kHz	180 kHz
Coverage	165 dB	164 dB
Battery Life	15+ years	10+ years
Peak Current	32 mA	120 mA
Sleep Current	1 μ A	5 μ A
Throughput	50 Kbps	60 Kbps
Latency	Device Class Dependent	<10 s
Security	Advanced Encryption Standard (AES) 128 bit	3GPP (128 to 256 bit)
Geolocation	Yes (TDOA)	Yes (in 3GPP Rel 14)
Cost Efficiency (Device and Network)	High	Medium

3. Materials and Methods

The study and application of certain types of portable and easy-to-operate sensors have been growing considerably. Portable sensors, namely accelerometers, with small dimensions, low energy consumption and high precision have been used in many tests in individuals who have pathologies that can limit their mobility, allowing us to validate in real time if a given individual suffers an abrupt fall [50].

In the work presented in [20], several authors who have worked with these and other sensors, show the advantage of using these small devices for following and monitoring people. It is agreed that one of the main problems for the elderly is related to the occurrence of falls, which, in many cases, end up incapacitating people, namely due to fractures, and other disabling pathologies, namely those that are chronic, degenerative and naturally associated with aging (osteoarthritis, osteoporosis and chronic musculoskeletal pain (fibromyalgia), among others). People who suffer from disabling psychological and neurodegenerative diseases are naturally excluded.

Currently, mobile communication devices, commonly referred to as smartphones, have several sensors incorporated within them, including an accelerometer, gyroscope and GPS, etc., yet the elderly population often find them difficult to operate, not being accustomed to using this type of technology and, in most cases, having great physical limitations and barriers to the use of technologies, as mentioned in [51,52]. Thus, in the present work, we propose the real-time monitoring of the movement of elderly people, who are prone to eventual falls, as well as their state of health, both inside of their houses and in the surrounding area, while also monitoring their ability to move, their pulse and their fatigue resistance, using sensors incorporated in non-intrusive pervasive devices.

The system consists of an application set composed of software and hardware, namely an application developed for portable devices, based on the ESP32 microcontroller (MCU). This MCU incorporates technologies to support Wi-Fi and Bluetooth communications, except LoRa communication.

LoRa SX127x or RFM9x transceivers add the necessary support for LoRa communications and the LoRaWAN protocol that is required to establish communications with The Things Network (TTN) [53]. It should be noted that TTN is cloud server-based network communication infrastructure that connects LoRaWAN devices and gateways worldwide. Thus, every time someone connects a Gateway to TTN, coverage is expanded for all users and LoRaWAN devices, thus ensuring extended, free coverage of the LoRa network signal.

Equipment with different frequencies exists, depending on the target frequency band (433, 868 or 915 MHz). The frequencies used depend on the geographic region and the regulations of the local Industrial, Scientific, and Medical (ISM) band, being, in most countries in the European Union and, in particular, Portugal, the 868 MHz frequency band [54]. This can be integrated with a low-cost GPS sensor, for example GY-GPS6MV2 [55]. For personal use, another ESP32 device with LoRa support can be used, which already includes GPS [56]. The low-cost ADXL335 accelerometer sensor [57] is compatible with ESP32 and can be used for fall detection and system activation (by motion detection).

ESP32 is designed for mobile, wearable electronics and IoT applications. It has all the most recent features of low-power chips, including fine-grained clock gating, multiple power modes and dynamic power scaling. For example, in a low-power IoT sensor hub application scenario, the ESP32 is enabled periodically and only when a specified condition is detected. The low load cycle is used to minimize the amount of power the chip consumes. The output of the power amplifier is also adjustable, thus contributing to an optimized trade-off between communication range, data rate and energy consumption [58].

To control vital signs, we can connect the body temperature sensor [59], body humidity [60] and pulse rate [61] to the ESP32 microcontroller. The equipment is installed in a device suitable for each person (bracelet, waistcoat, etc.), in order to make it safe, concealed and comfortable, eliminating user interaction in most operations. Communication is carried out automatically through the communication of the main module with the LoRaWAN gateway, sending user monitoring data to the TTN at pre-defined intervals, which are stored in a database with real-time analysis by the entities and authorized in a network scheme similar to that shown in Figure 1. In this context, it is important to note some definitions [53]:

- End Device, Node, Mote: an object with an embedded low-power communication device.
- Gateway: devices that form the bridge between other devices and The Things Network. These devices use low-power networks like LoRaWAN to connect to the gateway, while the gateway uses high bandwidth networks like Wi-Fi, ethernet or cellular connections to connect to The Things Network.
- Network Server: servers that route messages from end devices to the right application, and back.
- Application: a piece of software running on a server.

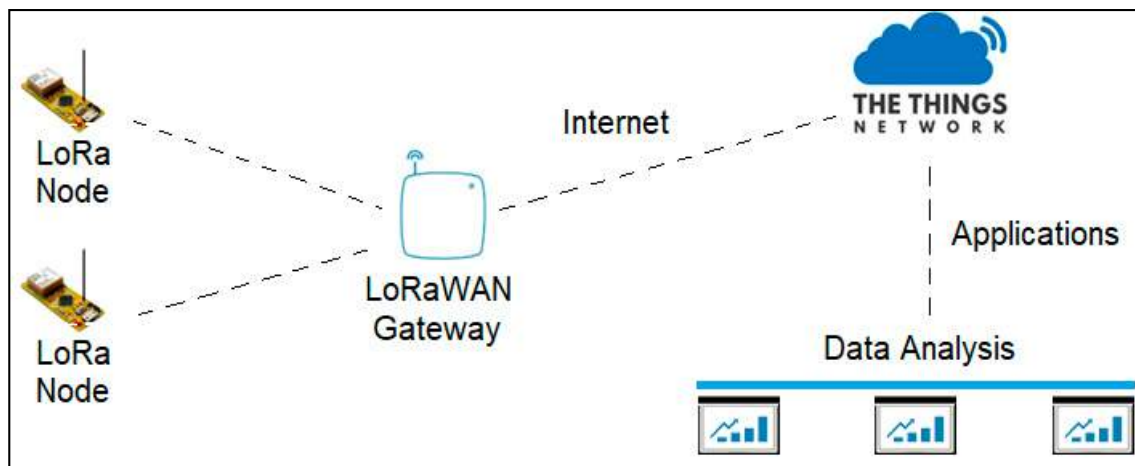


Figure 1. LoRaWAN architecture (adapted from [62]).

3.1. LoRaWAN Protocol

LoRa is a wireless modulation for long-range, low-power, and low-data rate applications developed by Semtech. LoRaWAN is a network protocol that belongs to the set of LPWANs specified in [13] by the LoRa Alliance, which uses LoRa modulation in its physical layer. In the new specification (version 1.1), a Join Server (JS) was added in order to make communications more reliable and secure, being responsible for storing several keys.

LoRa devices (nodes) are located around the different gateways. Gateways then connect to servers (to the network) using IP connections, bridging the devices and the network (backend).

The devices use different channels and binary rhythms depending the request. By LoRa modulation, the change in this binary rhythm is promoted through an Adaptive Data Rate (ADR) scheme specific to the LoRaWAN network.

A representation of the LoRaWAN stack can be seen in Figure 2.

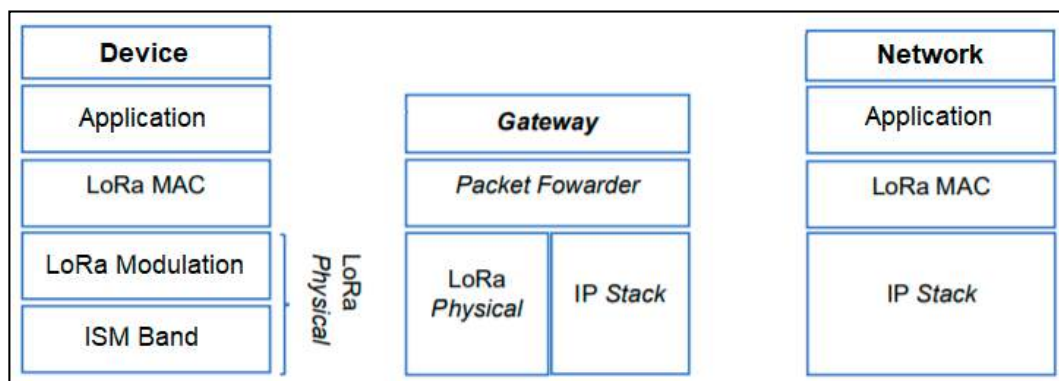


Figure 2. Stack LoRaWAN (adapted from [13]).

The LoRa application layer is composed of data from various actuators and sensors on the device.

The LoRa Medium Access Control (MAC) layer is responsible for managing the network. This management obeys the type of device class used. Medium Access Control (MAC) commands allow us to make changes or check the status from the web.

The LoRa Modulation layer concerns the type of modulation used, which is usually modulation LoRa. LoRaWAN also provides the use of frequency-shift keying (FSK) modulation.

The Industrial, Scientific and Medical (ISM) band concerns a set of specifications of the frequency band of a given region, namely the frequencies and bandwidth of the transmission channels and a set

of rules to be respected. Among these rules are the duty cycle allowed per channel and the timing of entry into sleep and active modes. EU863-870 MHz is an example of a European ISM band.

The gateway bridges the device and the network and translates LoRa messages from the physical layer to IP protocol messages.

Before an end device (LoRa Node) can communicate on the LoRaWAN network, it must be activated and the following information is required [13]:

- **Device Address (DevAddr):** This is a 32-bit identifier that is unique within the network, present in each data frame and shared between the end device, network server, and application server. This differentiates nodes within the network, allowing the network to use the correct encryption keys and properly interpret the data.
- **Network Session Key (NwkSKey):** This is a 128-bit AES encryption key that is unique per end device and is shared between the end device and network server. This provides message integrity for communications and provides security for end device to network server communication.
- **Application Session Key (AppSKey):** This is a 128-bit AES encryption key that is unique per end device and is shared between the end device and application server. This is used to encrypt/decrypt application data messages and to provide security for application payload.

The LoRaWAN protocol defines three classes of devices (A, B and C) with different functionalities. The LoRaWAN network must be prepared to handle devices of all classes.

- **Class A:** All devices on the LoRaWAN network need to implement the functions described by this class, even those of class B and class C. Class A devices send information at their discretion (ALOHA). ALOHA is a specific type of MAC that is characterized by sending packets through the terminals when there is information to send from higher layers. As such, collisions can occur when there are simultaneous transmissions, since the medium is shared and not dedicated. In the case of not receiving a message, this type of MAC waits a certain time, called backoff, in order to retransmit the packet.
- **Class B:** In addition to the capacity of class A devices, these devices are characterized by opening windows of extra time (ping slots) at defined time intervals. So, more data from the servers can be forwarded to devices in this class. Gateways send beacons so that the devices are synchronized and ready to open these extra windows. If a device wants to have class B functionality, it looks for the existence of these beacons. If these are not found, a BEACON_NOT_FOUND message is sent from the MAC layer to the application layer of the device. If a message from BEACON_LOCKED is found, it is sent to the application. The information that the device has passed class B is communicated to the network by sending a 1 bit message in the Fctrl field of the uplink messages.
- **Class C:** This type of device configuration often has active reception windows, which implies higher energy consumption, so its implementation in real systems is rare. The reception windows practically only close when the device is transmitting.

3.2. Functional Requirements

In terms of functional requirements, the following operations are mainly considered:

- Creation of a LoRaWAN gateway network in remote and isolated regions with redundant coverage, in which at least one gateway will be connected to the Internet (3G/4G);
- Secure connection service with user registration, authentication and validation;
- Data collection function of sensors attached to the device (GPS, accelerometer, temperature, pulse, body humidity);
- Housing sensor data collection function (temperature, gases, smoke, flood);
- User data sending function via LoRa communication;
- Portable device parameterization and configuration function;

- Real-time analysis of the data collected, to detect deviations beyond the permitted tolerances;
- Alerts when there is an abnormal occurrence in the user's device (sent to family, friends or entities and security forces), namely when the equipment signal is lost, or a fall occurs.

3.3. Non-Functional Requirements

Equally important are the non-functional requirements, which are responsible for ensuring functionality and operability in accordance with minimum quality standards, namely:

- Reliability—the system must be tested in order to improve its robustness, guaranteeing its operation in low-signal situations
- Security—the system must be safe from the user's point of view, namely through both the placement of sensors in places that do not compromise the user's mobility, and the use of sensors that are not so fragile that they deteriorate, making the data collection useless.
- Usability—the system should be easy to use. At this level, it is intended that the system has an easy-to-use interface, without requiring major user intervention.
- Effectiveness—the system must be effective, accurate and proven to be useful in response to social needs at critical moments, such as those currently experienced by society.

4. Conceptual Scheme

The set of applications supporting our monitoring and follow-up system for the elderly includes several modules, as shown in Figure 3, namely:

- A data collection module for personal use, consisting of an ESP32-LoRa microcontroller with sensors, as mentioned above;
- A housing status data collection module, consisting of an ESP32/LoRa microcontroller with environmental sensors (temperature, humidity, carbon monoxide, gas and smoke).

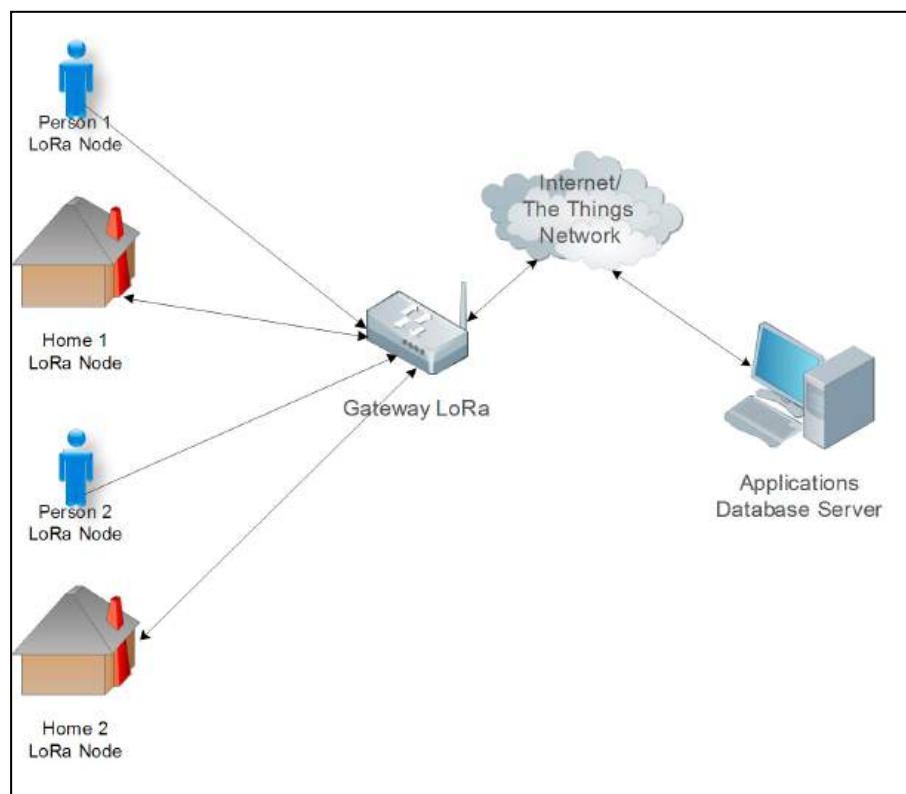


Figure 3. System conceptual scheme.

The LoRa gateway is connected to the Internet, receiving the data that are periodically sent from the LoRa nodes. As for LoRa nodes, these are divided into two distinct types—personal LoRa nodes and residential LoRa nodes.

4.1. Personal LoRa Node

The personal LoRa node is composed of an ESP32-based MCU with the various sensors coupled and placed in areas that do not interfere with the user's daily life, therefore being as unintrusive as possible. The ESP32 MCU allows a battery saving mode (Deep Sleep Mode) that is only activated during the scheduled period, collecting and sending the data at that moment.

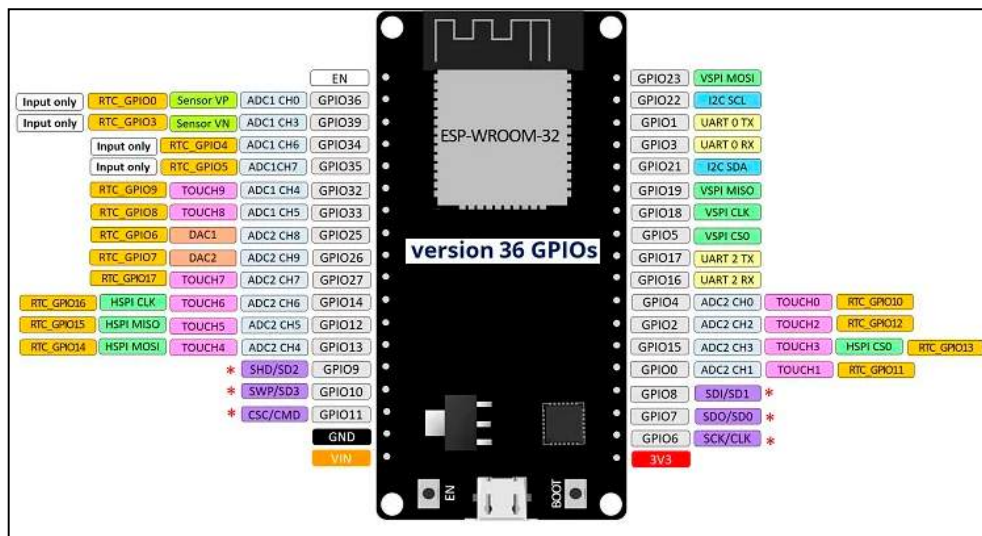
Considering that one of the built-in sensors is the gyroscope, whenever there is a sudden change, due to a fall, for example, it can automatically activate the MCU by programming a General Purpose Input/Output (GPIO) interruption of the Real Time Clock (RTC). The ESP32 MCU consists of several modules (Figure 4b) and can operate in the following modes, as it seen in Table 2 [58]:

- Active Mode: the chip radio is powered on. The chip can receive, transmit, or listen.
- Sleep Mode Modem: the CPU is operational and the clock is configurable. The Wi-Fi/Bluetooth baseband and radio are disabled
- Light Sleep Mode: the CPU is paused. The RTC memory and RTC peripherals, as well as the Ultra-Low Power (ULP) co-processor, are running. Any wake-up events (MAC, host, RTC timer, or external interrupts) will wake up the chip.
- Deep Sleep Mode: only the RTC memory and RTC peripherals are powered on. Wi-Fi and Bluetooth connection data are stored in the RTC memory. The ULP co-processor is functional.
- Hibernation Mode: the internal 8-MHz oscillator and ULP co-processor are disabled. The RTC recovery memory is powered down. Only one RTC timer on the slow clock and certain RTC GPIOs are active. The RTC timer or the RTC GPIOs can wake up the chip from Hibernation Mode.

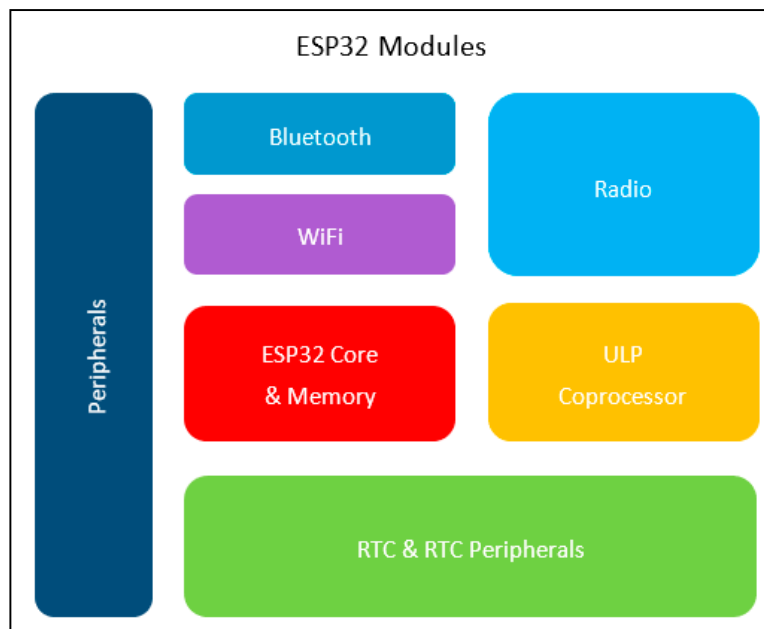
Table 2. Power consumption by power modes (source: [58]).

Power Mode		Description	Power Consumption
Active (RF working)		Wi-Fi Tx packet	160 mA~260 mA
		Wi-Fi/BT Tx packet	
		Wi-Fi/BT Rx and listening	
Modem-sleep	The CPU is powered on	240 MHz	Dual-core chip(s) Single-core chip(s)
			N/A
		160 MHz	Dual-core chip(s) Single-core chip(s)
			27 mA~44 mA 27 mA~34 mA
		Normal speed: 80 MHz	Dual-core chip(s) Single-core chip(s)
			20 mA~31 mA 20 mA~25 mA
Light-sleep		-	0.8 mA
Deep-sleep		The ULP co-processor is powered on	150 μ A
		ULP sensor-monitored pattern	100 μ A @ 1% duty
Hibernation		RTC timer + RTC memory	10 μ A
		RTC timer only	5 μ A
Power-off		CHIP_PU is set to low level; the chip is powered off	1 μ A

ESP32 has 34 GPIO pins that can be assigned several functions by programming the appropriate registers. There are several kinds of GPIOs: digital-only, analog-enabled, capacitive-touch-enabled, among others. Analog-enabled GPIOs and capacitive-touch-enabled GPIOs can be configured as digital GPIOs. The ESP32 Pin Layout is shown in Figure 4a. MCU ESP32 contains one or two low-power Xtensa 32-bit LX6 microprocessor(s) with several features, namely a seven-stage pipeline to support a clock frequency of up to 240 MHz (160 MHz for ESP32-S0WD, ESP32-D2WD, and ESP32-U4WDH) and a 16/24-bit instruction set that provides high code density, among others [58].



(a)



(b)

Figure 4. (a) ESP32 DevKit V1 GPIO Scheme (adapted from [54]); (b) ESP32 microcontroller (MCU) modules (adapted from [58]).

Therefore, via the ESP32 MCU, the Deep Sleep battery-saving mode can be activated, which will have an extremely low power consumption. In this mode, the CPUs, most RAM and all clocked digital peripherals are turned off. The only parts of the chip that can still be connected are the RTC controller, RTC peripherals (including the ULP coprocessor) and RTC memories. This device has several ways of activating ESP32 when in Deep Sleep mode, and wake-up sources can be set up at any time before entering Deep Sleep mode. It is possible to wake up ESP32 through the timer, external wakeup (ext0), external wakeup (ext1), ULP coprocessor wakeup and the touchpad (GPIO touch sensor), so in the present situation an external wakeup (ext0) can be used. The RTC IO module contains firmware to trigger the alarm clock when one of the RTC GPIOs enters a predefined logic level. RTC IO is part of the power domain of RTC peripherals; therefore, RTC peripherals will be kept on during Deep Sleep if this activation source is requested [58].

Only GPIOs with RTC functionality can be used, in this case pins 0, 2, 4, 12–15, 25–27 and 32–39.

4.2. Residential LoRa Node

As with the personal node, the residential LoRa node is composed of an ESP32-based microcontroller (MCU) with various sensors coupled and placed in areas that do not interfere with the use of the home. The data are sent periodically, in previously defined periods, and can also be sent immediately, whenever certain values read on the sensors exceed the previously established limits, considering that there will be a situation of alert or threat to the safety of residents and housing.

All the necessary data modeling is supported on a platform developed for this purpose and hosted on a dedicated server, which serves as a form of service infrastructure.

To take advantage of IoT technologies, namely LoRa communications, The Things Network (TTN), which is a collaborative communication infrastructure for Internet of Things, is used as a reference, and is accessible in [53].

5. System Prototype

For proof of concept and the demonstration of the potential of telecommunications by LoRa Technology, a LoRaWAN gateway (Single Channel) was configured with connection to the TTN network in the Viseu region and a LoRa node as a client that attaches a temperature and humidity sensor (DHT22). The equipment used has the following characteristics:

5.1. LoRa Gateway

The equipment used to build the LoRa gateway was as follows:

- TTGO ESP32 OLED SX1276 LoRa 868/915 MHz Bluetooth WI-FI Lora Internet Antenna Development Board;
- USB 3.3 V–5 V (power supply);
- Internet connection (Wi-Fi via ADSL/Fiber/3G/4G);
- Transparent PVC box;
- One-channel gateway, with server software adapted from [63].

In the prototype (Figure 5), the gateway is configured with software available on the GitHub page mentioned above [63], with appropriate adaptations both to the characteristics of the local internet connection network, and to the registration and access properties of TTN.

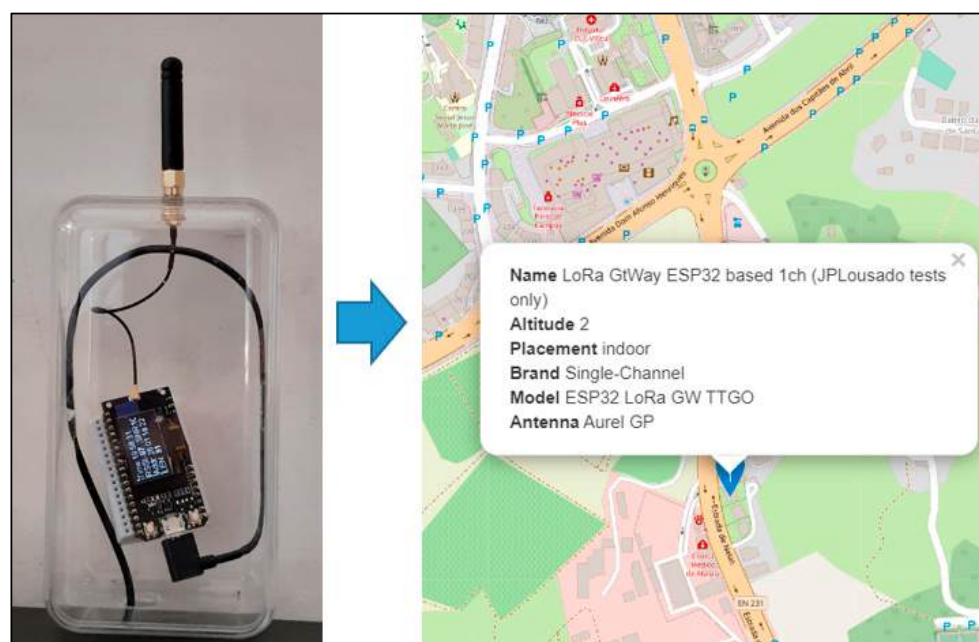


Figure 5. LoRa gateway registered with The Things Network (TTN).

It is important to know the address of the TTN routing server in advance, so that a correct connection can be established and to create the gateway service on the TTN network. After establishing the Internet connection, accessing the server is possible via the IP address and by having access to its configuration, where it is also possible to make changes to the configuration parameters, as well as to gain access to the statistics of packets sent and received. In this administration interface (Figure 6), it is possible to change some of the parameters and have access to the history as well as the general state of the connection.

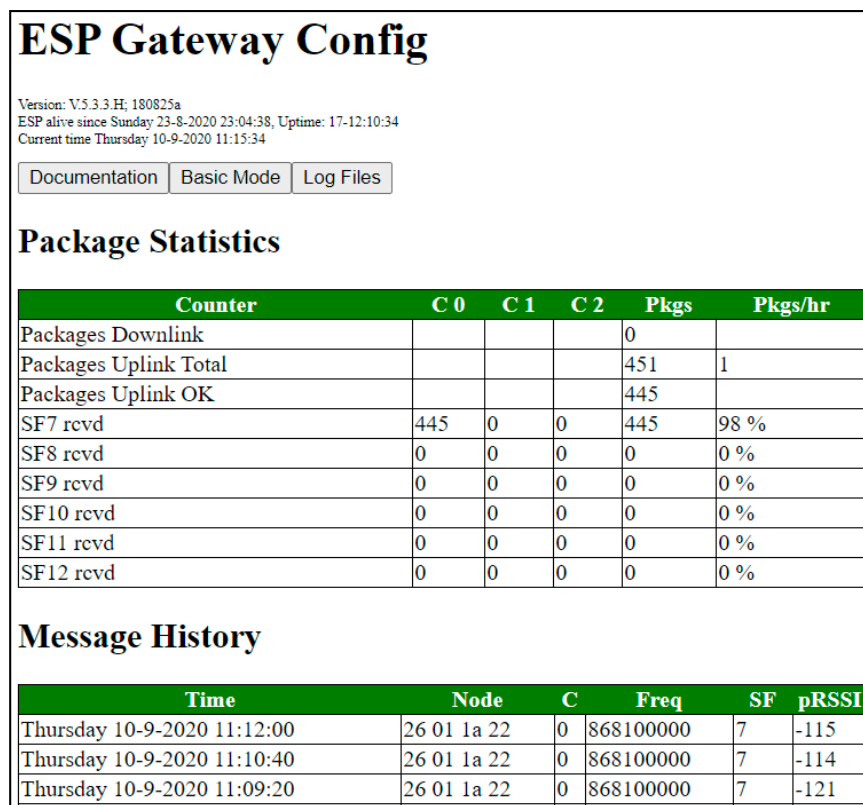


Figure 6. LoRaWAN gateway based on ESP32 MCU in operation.

5.2. LoRa Node

The equipment used to build the prototype node was as follows:

- TTGO ESP32 OLED SX1276 LoRa 868/915 MHz Bluetooth WI-FI Lora Internet Antenna
- DHT22 sensor (temperature and humidity);
- Protoboard;
- Resistance of 10K Ω ;
- Connection cables;
- USB 3.3 V–5 V (Power Bank 5000 mAh SoundLogic Solar Powered);
- Node software based on Cayenne LPP (secure up to 51 bytes of data), available in [64].

Figure 7 shows the prototype assembled and in operation.

After the LoRa node is operational, an application has to be created on the TTN registration system console.

Through this application, a set of operations is understood, with which the devices communicate on the Internet via TTN. This can be as simple as a small web application, or a visual flow using Node-RED to customize code on a server, as described in [65]. Before communication with devices, it is necessary to add the application to TTN and register the device.

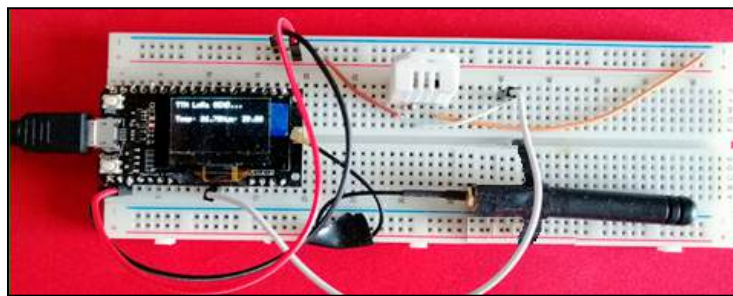
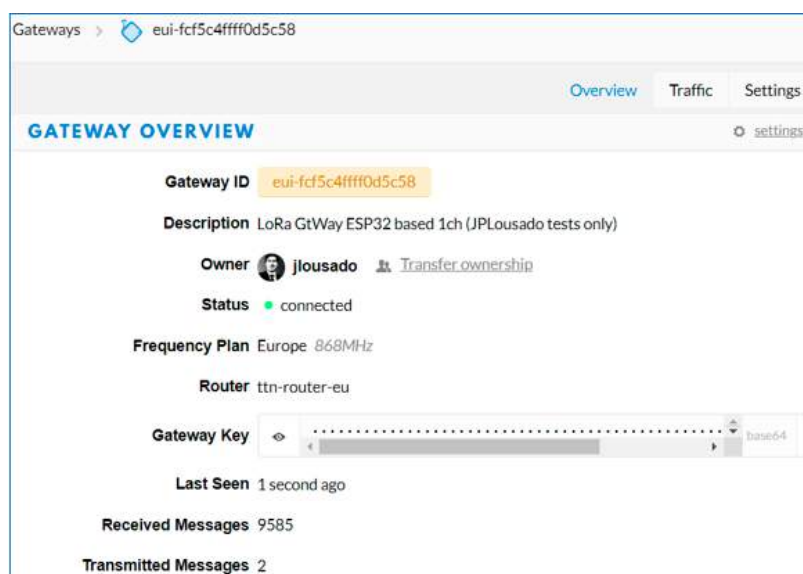


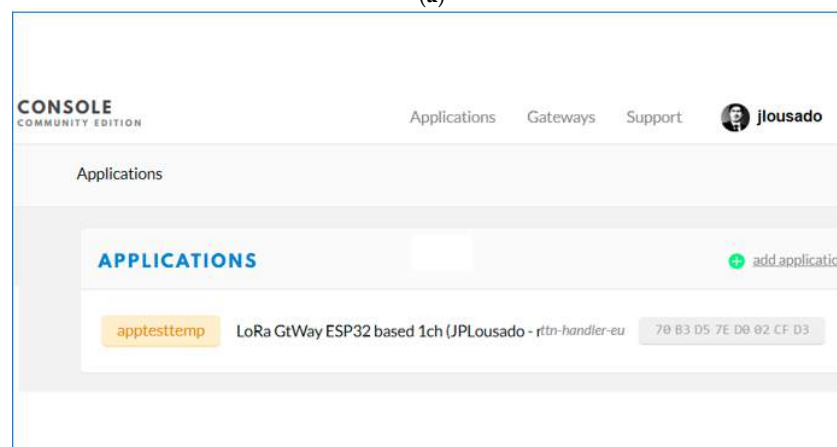
Figure 7. LoRa node in operation with reading data from the DHT22 sensor.

5.3. TTN Application Creation

Once the gateway is configured and connected to TTN, the application that will collect the data can be added. For this, it is also necessary to register the device (LoRa node). The node in the present case only collects temperature and humidity data by sending the data to the server every minute, via the LoRa gateway. In order for the application to be able to collect the device data, it is necessary to proceed with the configuration of the device with the data of the access keys to the application, otherwise the added device will not be visible in the application. In this way, TTN ensures that packets sent by the device are effectively collected by the correct application (Figure 8).



(a)



(b)

Figure 8. Cont.

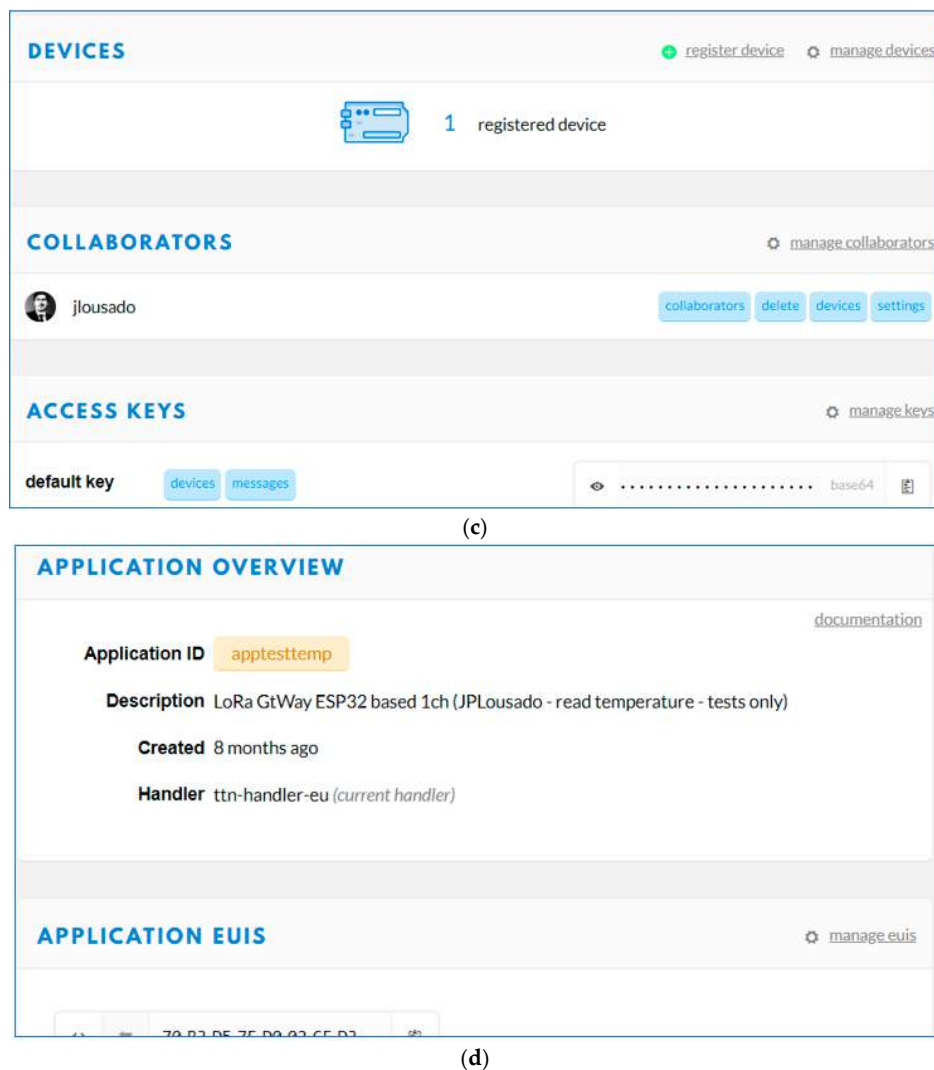


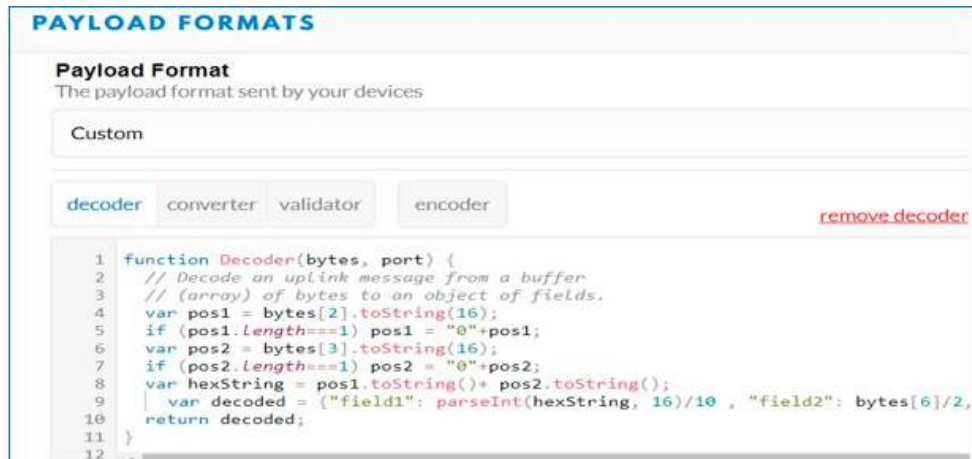
Figure 8. Overview of the LoRa node and gateway devices and the application in the TTN: (a) gateway registration status information; (b) registration of the “apptesttemp” application on the TTN; (c) registration of the LoRa node device in the application; (d) application registration information.

The more expanded the LoRa gateway network, is the better coverage it will have, so once the application is created and the LoRa node device is configured, data packets can be received by more than one gateway. Since multiple gateways can receive the same LoRa RF data packets from a single end device, LNS (LoRa Network Server) eliminates duplicate data and removes all copies. Based on the Received Signal Strength Indication (RSSI) levels of identical messages (data packets), the network server typically selects the gateway that received the best RSSI message when transmitting a downlink message because, from the outset, that gateway it is the closest to the device that sent the message, ensuring a better quality of service [65].

5.4. Connection with ThingSpeak

Once the prototype is working, it is important to select the payload format, which represents the way data are received and displayed on the network. By default, Cayenne LPP (low-power payload) will be selected; however, in this case, this has been changed to a custom format in order to program the decoder function so that the data are presented in the correct format, compatible with the platform we intend to use for data visualization, the ThingSpeak platform [66].

In the received packet, we need to decode the parameter “bytes” that comes in the Cayenne LPP format and present the fields in JavaScript Object Notation (JSON) format [64], with the positions bytes(2) and bytes(3) representing the temperature times 10, which is necessary to proceed with the correction. The bytes(6) position represents the humidity as a double value, so it is also necessary to correct this value. To this end, we implemented the JavaScript function shown in Figure 9a, because the ThingSpeak platform works with predefined composites (field1, field2, etc.).



PAYLOAD FORMATS

Payload Format
The payload format sent by your devices

Custom

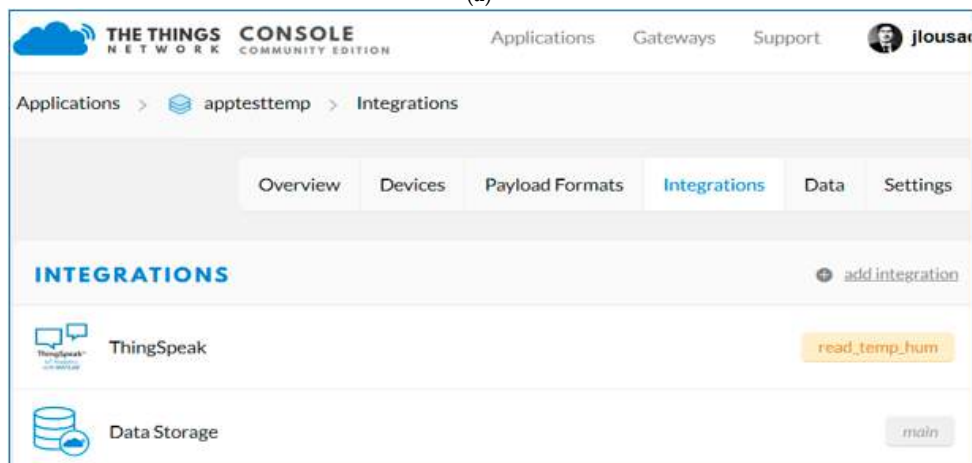
decoder converter validator encoder remove decoder

```

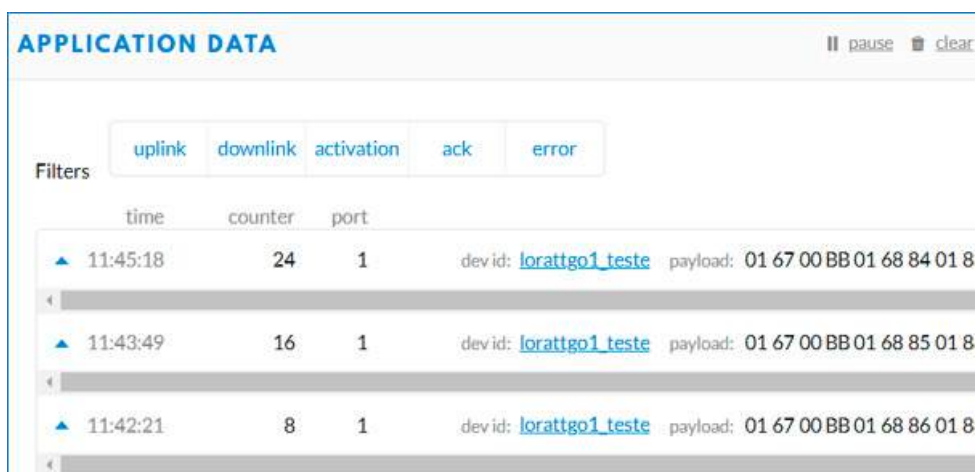
1 function Decoder(bytes, port) {
2   // Decode an uplink message from a buffer
3   // (array) of bytes to an object of fields.
4   var pos1 = bytes[2].toString(16);
5   if (pos1.length===1) pos1 = "0"+pos1;
6   var pos2 = bytes[3].toString(16);
7   if (pos2.length===1) pos2 = "0"+pos2;
8   var hexString = pos1.toString()+ pos2.toString();
9   var decoded = {"field1": parseInt(hexString, 16)/10 , "field2": bytes[6]/2,
10  return decoded;
11 }
12


```

(a)



(b)



APPLICATION DATA || pause  clear

Filters uplink downlink activation ack error

	time	counter	port	
▲	11:45:18	24	1	dev id: lorattgo1_teste payload: 01 67 00 BB 01 68 84 01 8
◀				
▲	11:43:49	16	1	dev id: lorattgo1_teste payload: 01 67 00 BB 01 68 85 01 8
◀				
▲	11:42:21	8	1	dev id: lorattgo1_teste payload: 01 67 00 BB 01 68 86 01 8
◀				

(c)

Figure 9. (a) Decoder function compatible with ThingSpeak; (b) received message (payload); (c) TTN integration with ThingSpeak infrastructure.

For example, when temperature and humidity are detected and sent on the LoRa network to the LoRaWAN gateway with payload 01 67 00 FB 01 68 72, as shown in Figure 9b, these are encoded as follows [67]:

1. Device with temperature sensor: (Hex)—01 67 00 FB. The data channel is one (01), the type is temperature (67) and the value is 00FB \Rightarrow 251 \Rightarrow 25.1 °C.
2. Device with humidity sensor: (Hex)—01 68 72. The data channel is one (01), the type is humidity (68) and the value is 72 \Rightarrow 114 \Rightarrow 57%.

To register a data analysis application on the ThingSpeak platform, a registration is required, which is free in its basic version. After creating the channel, we selected the fields that we wanted to display and defined the metadata, field1—temperature and field2—humidity, according to what was defined in the decoder function (bytes, port). It is possible to have up to eight fields in a channel and GPS coordinates. The channel ID and channel write Application Program Interface (API) key are required to register the channel in the TTN (Figure 9c) and allow data communication. The channel also allows for the configuration of other parameters, as well as exporting the data in XML and JSON format (Figure 10).

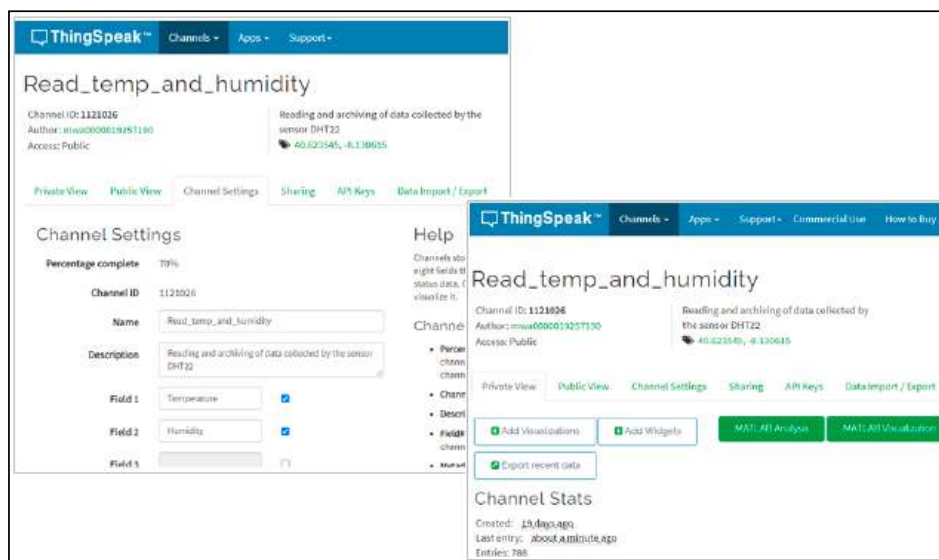


Figure 10. Parameterization of the ThingSpeak channel.

5.5. ThingSpeak Dashboard

After ensuring that the channel is properly configured and communicating with TTN, it is possible to gain access to the graphical display of the data, as well as the geographic location of the device (Figure 11).

The ThingSpeak platform also provides the necessary APIs so that data can be collected by Representational State Transfer (REST) web services, to be incorporated into an Android, iOS or Windows application, thus allowing real-time monitoring in another system developed for this purpose.

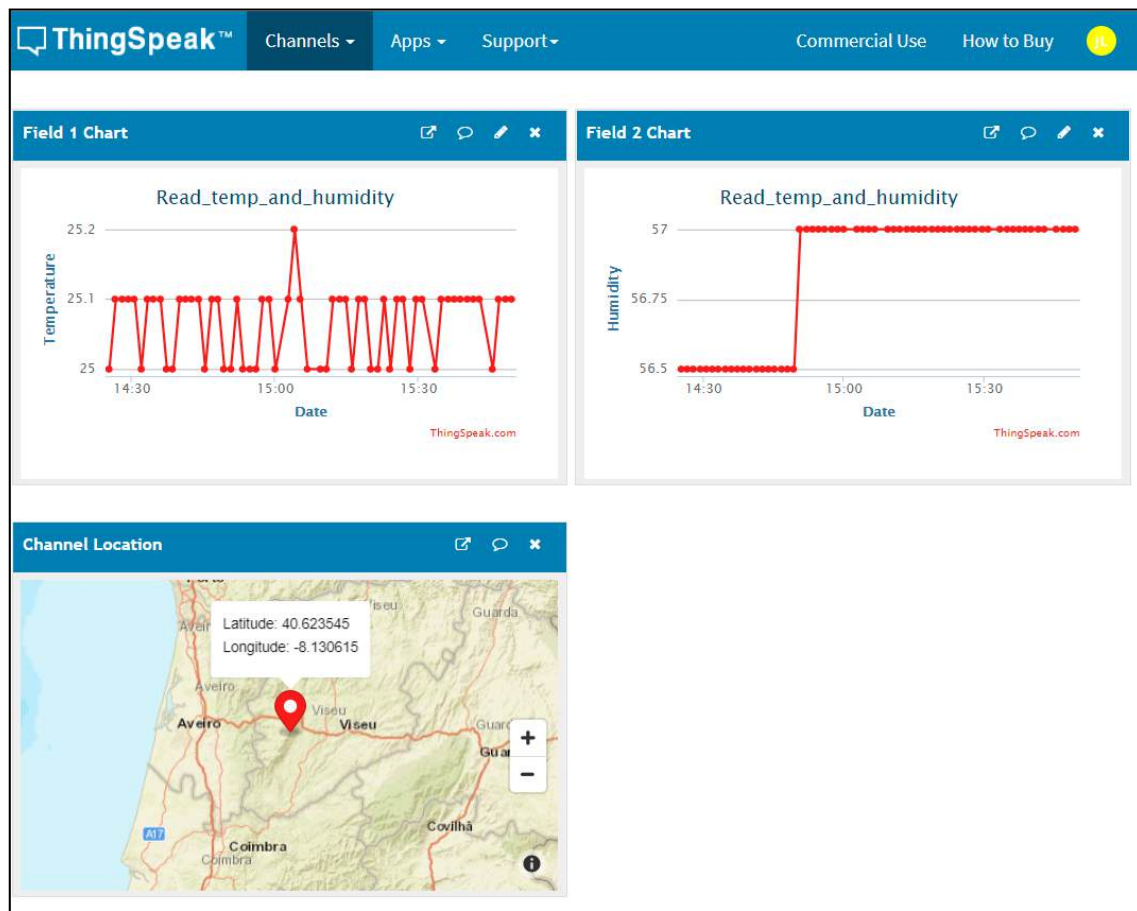


Figure 11. Real-time display of data sent by the DHT22 sensor.

6. Results

In this article, we show that it is possible to implement low-cost and low-energy consumption systems, even for domestic consumption, based on LoRa networks and an ESP32 microprocessor. Nevertheless, there are some considerations that concern us, and these need to be solved so that the system's effectiveness can be measured, namely:

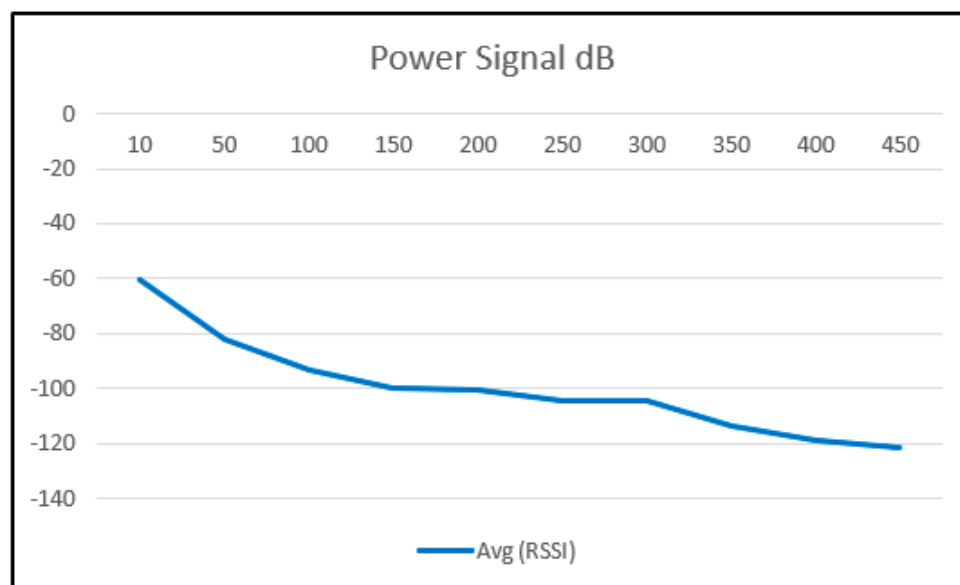
- Statistical analysis of RSSI in remote areas, rural areas, and rugged terrains.
- Shadow zones with several urban and natural obstacles, such as buildings and hills, among others, that cause disturbances in communication, reducing coverage.
- The poor network of gateways installed in Portugal, which does not allow for minimum acceptable coverage for the implementation of a generalized system.
- The tests carried out, despite being very limited and taking into account the fact that the gateway is installed indoors, did not allow communications beyond 1.2 km, which, although not negligible, is below the expected value.
- The one-channel experimental gateway does not allow for application stress tests, nor system overload and robustness, so it only served as a proof of concept; however, the recent installation in the multi-channel gateway region can easily extend the range of the test and its robustness.

To analyze the performance in terms of the received signal power, several locations were previously selected at 10 distances (in meters), as presented in the following table (Table 3), with average values obtained from 100 measurements of RSSI power (LoRa gateway):

Table 3. List of distances from gateway and average Received Signal Strength Indication (RSSI) value.

Distance (m)	Avg (RSSI)
10	−60.4
50	−82.3
100	−93.0
150	−99.8
200	−100.6
250	−104.2
300	−104.8
350	−113.6
400	−119.1
450	−121.5

Figure 12 shows a graph of the average data obtained, as presented in the previous table.

**Figure 12.** Graphical representation of average RSSI data and respective distances.

Higher RSSI values represent greater signal quality, while lower values represent poorer signal quality. According to [36], the RSSI values for LoRa networks are:

- −30 dBm -> excellent quality;
- −120 dBm -> very poor quality.

The results obtained in previously defined hybrid rural and urban areas were different from what was expected, and it was found that any obstacle, wall, or housing could interfere with the signal. It was also found that, with the used equipment, it is not possible to communicate beyond 450 m. Nevertheless, several restrictions must be considered, namely the fact that the test gateway is only one channel, with an indoor antenna housed at the bottom, as shown in Figure 5. During data collection, the device operator remained in the same place for some time in order to collect 10 samples for each distance, moving only the LoRa node (random movement inside one circle with a bias of no more than 5 m) to check if there were failures, which was confirmed.

Linear Regression Model

In order to obtain a detailed statistical analysis and validation of our system that allowed for the measurement of the service quality and the influence of environmental factors on the signal quality,

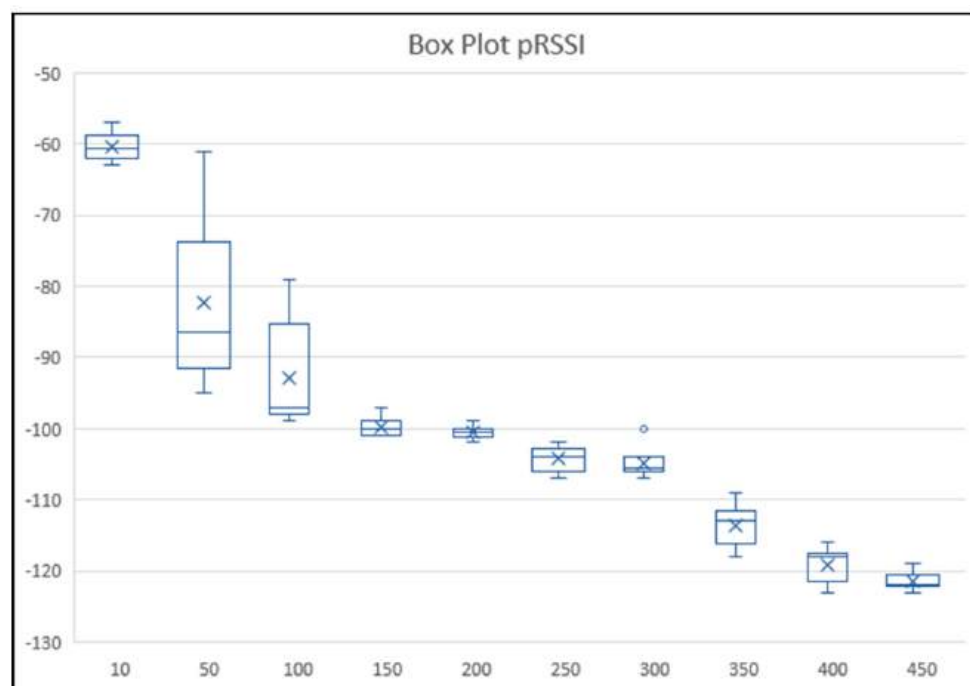
we applied a linear regression model to the data obtained, with the dependent variable being power of RSSI (pRSSI) and the independent variable being the distance to the gateway. The results in Table 4 allow us to observe that, with a correlation factor, Multiple R is equal to 0.9016, showing that there is a strong correlation between the two variables. However, the value obtained for R Square, 0.813, shows us that only 81.3% of the observed cases fit the obtained model, when the desirable value would be 95%. Several factors may have contributed to this result, namely environmental factors such as obstacles, trees, walls, and the geography of the terrain.

Table 4. Linear regression summary output.

Regression Statistics	
Multiple R	0.901694435
R Square	0.813052854
Adjusted R Square	0.81114523
Standard Error	7.784507372
Observations	100

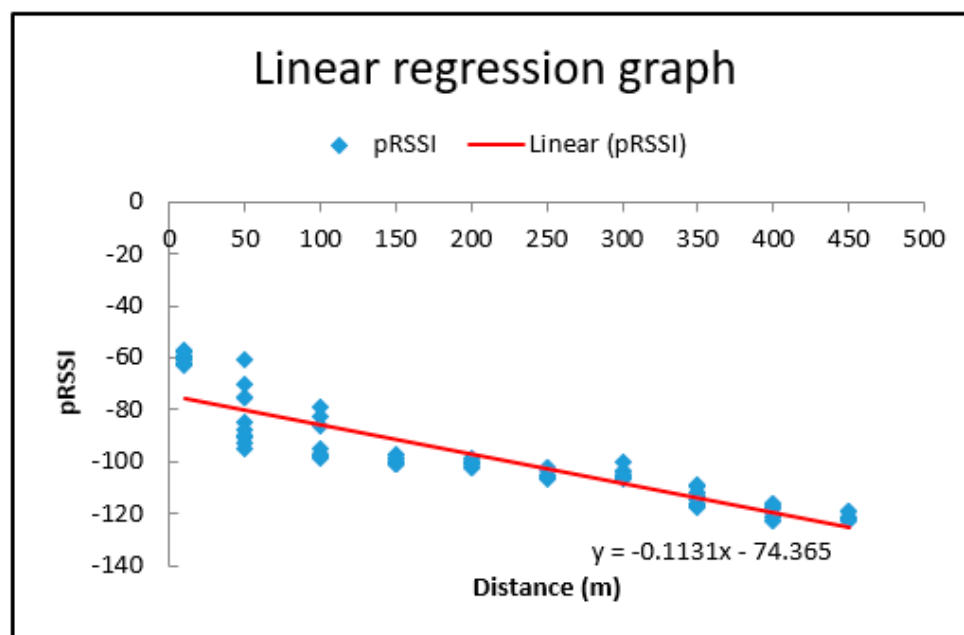
From the analysis, a p -value much lower than 0.05 (tends to zero) was obtained, so it is clear that there is a strong dependence of the variable pRSSI on distance, as expected.

Figure 13a presents a boxplot of the data obtained, with some deviations that help to explain the bias observed in R Square. In Figure 13b, we present a scatter plot with a trend line, which shows the adjustment of the line to the point cloud obtained from the pRSSI readings.



(a)

Figure 13. Cont.



(b)

Figure 13. (a) Boxplot graph for pRSSI; (b) Linear regression scatter plot graph with trend line.

To carry out accurate measurements with an error of less than one meter, we used the Google Maps© tool, specifically the “Measure distance” option. Figure 14 shows the process of obtaining the sites for measuring RSSI power.



Figure 14. Distance calculation process for RSSI tests.

We are convinced that the use of strategically located gateways that comply with the LoRa and LoRaWAN specifications, namely in terms of the power and gain of the external antennas, will have a considerable impact on the coverage and power of the received signal.

7. Discussion

Relevant facts related to the COVID-19 pandemic obliges us to think about new approaches of fast application regarding the protection and monitoring of the elderly, while promoting physical distancing and keeping them in their comfort zone, with the current proposal serving as a catalyst for a fast implementation of systems that can save human lives. In the case of housing, actuators may be

incorporated that will trigger certain actions, such as cutting the gas, water supply or electrical power, as well as triggering the discharge of fire-retardant chemicals.

Using the system proposed, isolated inhabitants, mostly the elderly, can move freely through the outside spaces of their homes without feeling confined in terms of their freedom and privacy, and in case of suffering some type of accident, fall or change in vital signs, a distress mechanism can be triggered by entities, family or friends, acquiring access to the GPS coordinates of their most recent location.

Taking into account both the current situation of the COVID-19 pandemic in Portugal and across the world, elderly people, who are naturally more vulnerable, and their families can benefit from this system, essentially due to the fact that family members, firefighters and security forces will have access to users' information and will be able to trigger a support action whenever any critical value in a given sensor is reached.

By including monitoring alongside georeferencing, the event of a fall or immobilization outside the residence will also enable the triggering of rescue means.

One of the most frequent causes of death in Portuguese rural regions has to do with carbon monoxide poisoning related to the use of braziers. This is another situation in which there can be a considerable benefit—whenever the sensors detect too high values of carbon monoxide, support teams, security forces or family members can provide support immediately.

It is also important to refer to the data obtained in the RSSI power readings, which raise some doubts in terms of coverage, as analyzed in the results section. The global solution to solving these coverage failures must use gateways with redundant and multichannel coverage, so that there is no blocking of devices when a gateway is sending data. When a device (LoRa node) sends data, it can be received by several gateways, though, depending on the quality of the received RSSI signal, only one gateway sends data, with the other data being discarded.

Another issue to be considered in this field is related to the placement of the antenna of each gateway, ensuring that they are properly located, at strategic points, in order to maximize the gain.

8. Conclusions

This article explores a very relevant application area for society, considering the potential underlying Long-Range (LoRa) telecommunications equipment and devices that are currently available on the market at low cost, but with high potential. The massification of IoT, directly related to the use of these miniaturized devices in the field of ubiquitous and pervasive computing, provides an excellent opportunity for their use in the follow-up and monitoring of elderly people and in the monitoring of their homes, namely with sensors that can detect floods, gas leaks, excess carbon monoxide, fires and other data.

The use of LoRa and TTN networks is specifically targeted at agricultural production and farming as well as the monitoring of environmental conditions in cities. Our approach, by introducing aspects related to the monitoring of people who are in a particularly vulnerable situation, especially the elderly, derived from COVID-19, is a challenge for us. We believe that we have demonstrated that it is possible to monitor people and their homes, offering them more security as well as a low-cost system, while ensuring their privacy.

Nevertheless, there are some barriers to the mass use of systems based on this technology, first of all due to the weak network of gateways available in the region, making it necessary to carry out studies on the implementation of equipment that reduces the shadow zones and allows redundant coverage.

Another relevant factor is related to the frequency of sending data and the volume of data produced, since LoRa networks are not designed to send large volumes of data, nor to be permanently connected in order to send data at a high frequency, for example, every 5 s. They are usually designed to send data in intervals of several minutes (10 to 30 min or more), which, to guarantee assistance to individuals in danger or who have experienced an accident, may be too long and may save lives or minimize risks. For this reason, the use of Artificial Intelligence with machine learning algorithms

can make an important contribution, foreseeing and anticipating risk situations and minimizing the probability of risky events occurring.

As in all data collection systems with continuous processes, having production databases in Online Analytical Process (OLAP) mode would be ideal; however, given the restrictions mentioned above, it is not currently possible to have systems that are capable of continuous data analysis processes. Thus, the inclusion of data mining, machine learning and Artificial Intelligence algorithms will have to operate on previously prepared databases rather than on production databases, using the concept of data warehousing, with pre- and post-processing. However, this does not invalidate the fact that, through Node-RED, a continuous connection for data analysis can be established, for example, by Message Queue Telemetry Transport (MQTT), which is the standard for IoT messaging.

Finally, it should be noted that, in order to guarantee all ethical and data protection principles in future work, the National Data Protection Commission will be informed of the objectives of our system and the way in which it functions. Anonymous data collection authorization will be requested, for statistical purposes only, such as for academic and scientific research.

9. Future Work

In terms of future work, the implementation of a robust system is foreseen, with technology that is more suitable for the common user, such as miniaturized sensors, and with a second-generation prototype, where the various sensors proposed here are all operational and coupled.

Another feature to be developed includes the use of Artificial Intelligence with machine learning algorithms, so that the data acquired by the system can be used in predictive and data mining methods and algorithms. With this functionality, it becomes possible to predict risk situations for the elderly, anticipating situations that could be harmful. As an example, when some parameters of the user's body position are successively exceeded, perhaps meaning a situation of imminent risk of falling, a possible preventive action can be triggered.

This approach will be based on studies carried out on subjects, namely the study presented in [68], where the authors characterize different types of sensors and their applications to prevent and predict both fall situations and the possible factors contributing to falls, namely physiological and biological factors. Real-time monitoring of the elderly can benefit from the use of data mining algorithms, namely Support Vector Machine (SVM), Gaussian Distribution of Clustered Knowledge, Multilayer Perceptron, Naive Bayes, Decision Trees, ZeroR, and OneR to gain insights into the data in order to detect and even predict future falls, as referred to in [69]. The integration of the system with Cloud platforms, namely the commercial platforms of Azure ML [70] or MathLab ML [71], may be an option, but they have high maintenance costs, so opensource platforms are the most favorable option, namely Node-RED [72].

The implementation of machine learning flows in Node-RED can be performed on a small low-cost computer, "Raspberry PI 3", by simply installing the necessary software and its dependencies, "[node-red-contrib-machine-learning]". This library adds the necessary functionalities to Node-RED to implement flows and to test and evaluate the predictive methods of machine learning incorporated in the tool. Figure 15 shows a set of sample workflows for a predictive method, in this case a decision tree classifier.

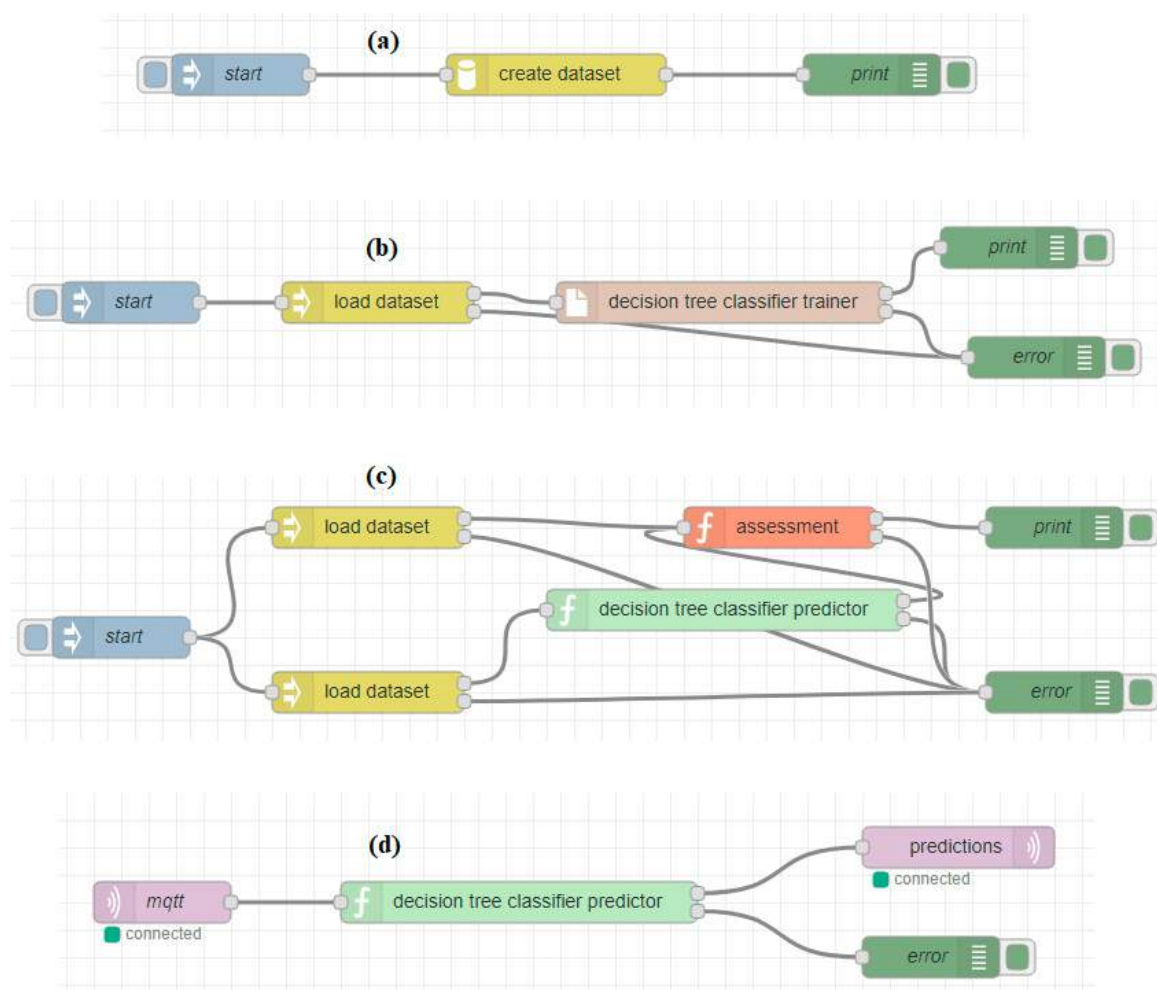


Figure 15. Example of flows for machine learning (source [72]): (a) this flow loads a csv file, shuffles it and creates a training and a test partition; (b) this flow loads a training partition and trains a ‘decision tree classifier’, saving the model locally; (c) this flow loads a test partition and evaluates a previously trained model; (d) this flow shows how to use a trained model during deployment. Data are received via Message Queue Telemetry Transport (MQTT), predictions are made and then sent back.

Author Contributions: Conceptualization, methodology, software and hardware assembly, J.P.L.; validation, S.A.; writing—review and editing, J.P.L., S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by National Funds through the Foundation for Science and Technology (FCT), I.P., within the scope of the project Ref^a UIDB/05583/2020.

Acknowledgments: This work was funded by National Funds through the Foundation for Science and Technology (FCT), I.P., within the scope of the project Ref^a UIDB/05583/2020. Furthermore, we would like to thank the Research Centre in Digital Services (CISeD) and the Polytechnic of Viseu for their support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Walls, A.C.; Park, Y.J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281–292.e6. [CrossRef] [PubMed]
2. Borges Guimarães, R.; Marques da Costa, N.; Nuno Nossa, P. Saúde urbana e território: Dos desafios pré e durante a pandemia às respostas pós-pandemia Territorial and urban health: From pre-pandemic and pandemic challenges to post-pandemic responses Correspondência. *Saúde Soc.* **2020**, *29*. [CrossRef]

3. Hamim, M.; Paul, S.; Hoque, S.I.; Rahman, M.N.; Baqee, I.-A. IoT Based Remote Health Monitoring System for Patients and Elderly People. In Proceedings of the 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 10–12 January 2019; pp. 533–538.
4. Khoi, N.M.; Saguna, S.; Mitra, K.; Ahlund, C. IReHMo: An efficient IoT-based remote health monitoring system for smart regions. In Proceedings of the 2015 17th International Conference on E-health Networking, Application & Services (HealthCom), Boston, MA, USA, 14–17 October 2015; pp. 563–568.
5. AlSharqi, K.; Abdelbari, A.; Elnour, A.A.; Tarique, M. Zigbee Based Wearable Remote Healthcare Monitoring System for Elderly Patients. *Int. J. Wirel. Mob. Netw.* **2014**, *6*, 53–67. [CrossRef]
6. Farhan, F.; Peifer, J. Remote Wellness Monitoring System with Universally Accessible Interface. U.S. Patent US7772965B2, 10 August 2010.
7. Zhai, Y.; Cheng, X. Design of smart home remote monitoring system based on embedded system. In Proceedings of the 2011 IEEE 2nd International Conference on Computing, Control and Industrial Engineering, Wuhan, China, 20–21 August 2011; pp. 41–44.
8. Maki, H.; Ogawa, H.; Matsuoka, S.; Yonezawa, Y.; Caldwell, W.M. A daily living activity remote monitoring system for solitary elderly people. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August–3 September 2011; pp. 5608–5611.
9. Atzori, L.; Iera, A.; Morabito, G.; Nitti, M. The social internet of things (SIoT)—When social networks meet the internet of things: Concept, architecture and network characterization. *Comput. Netw.* **2012**, *56*, 3594–3608. [CrossRef]
10. Wickramasinghe, N. Pervasive Computing and Healthcare. In *Pervasive Health Knowledge Management*; Springer: New York, NY, USA, 2013; pp. 7–13.
11. Gambi, E.; Montanini, L.; Pignini, D.; Ciattaglia, G.; Spinsante, S. A home automation architecture based on LoRa technology and Message Queue Telemetry Transfer protocol. *Int. J. Distrib. Sens. Netw.* **2018**, *14*. [CrossRef]
12. Sandoval, R.M.; Garcia-Sanchez, A.J.; Garcia-Haro, J. Performance optimization of LoRa nodes for the future smart city/industry. *Eurasip J. Wirel. Commun. Netw.* **2019**, *2019*, 1–13. [CrossRef]
13. Lora Alliance. *LoRaWANTM 1.1 Specification*; Lora Alliance: Beaverton, OR, USA, 2017.
14. Skouby, K.E.; Lynggaard, P. Smart home and smart city solutions enabled by 5G, IoT, AAI and CoT services. In Proceedings of the 2014 International Conference on Contemporary Computing and Informatics (IC3I), Mysore, India, 27–29 November 2014; pp. 874–878.
15. Haghmohammadi, H.F.; Neculescu, D.S.; Vahidi, M. Remote measurement of body temperature for an indoor moving crowd. In Proceedings of the 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 24–26 May 2018; pp. 1–6.
16. Ministério do Trabalho, Solidariedade e Segurança Social; Gabinete de Estratégia e Planeamento. Relatório de Portugal—Terceiro ciclo de revisão e avaliação da estratégia de implementação regional (RIS) do plano internacional de ação de madrid sobre o envelhecimento (MIPAA). 2017. Available online: https://www.unece.org/fileadmin/DAM/pau/age/country_rpts/2017/POR_report_POR.pdf (accessed on 15 January 2020).
17. Abdullah, S.; Choudhury, T. Sensing Technologies for Monitoring Serious Mental Illnesses. *IEEE Multimed.* **2018**, *25*, 61–75. [CrossRef]
18. Maresova, P.; Tomsone, S.; Lameski, P.; Madureira, J.; Mendes, A.; Zdravevski, E.; Chorbev, I.; Trajkovic, V.; Ellen, M.; Rodile, K. Technological Solutions for Older People with Alzheimer’s Disease: Review. *Curr. Alzheimer Res.* **2018**, *15*, 975–983. [CrossRef]
19. O’Neil, A.; Nicholls, S.J.; Redfern, J.; Brown, A.; Hare, D.L. Mental Health and Psychosocial Challenges in the COVID-19 Pandemic: Food for Thought for Cardiovascular Health Care Professionals. *Heart Lung Circ.* **2020**, *29*, 960–963. [CrossRef]
20. Lousado, J.P.; Antunes, S. e-Health Monitoring System for Senior Citizens based on LoRa Technology. In Proceedings of the 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 23–26 September 2020; IEEE: Split, Croatia, 2020; pp. 1–5.
21. Foubert, B.; Mitton, N. Long-Range Wireless Radio Technologies: A Survey. *Future Internet* **2020**, *12*, 13. [CrossRef]
22. Zhang, W.; Yang, J.; Zhang, G.; Yang, L.; Kiat Yeo, C. TV white space and its applications in future wireless networks and communications: A survey. *IET Commun.* **2018**, *12*, 2521–2532. [CrossRef]
23. Technical Marketing Workgroup. A Technical Overview of LoRa® and LoRaWAN™ What is it? 2015. Available online: https://www.tuv.com/media/corporate/products_1/electronic_components_and_lasers/TUeV_Rheinland_Overview_LoRa_and_LoRaWANtmp.pdf (accessed on 18 November 2020).

24. De Carvalho Silva, J.; Rodrigues, J.J.P.C.; Alberti, A.M.; Solic, P.; Aquino, A.L.L. LoRaWAN—A low power WAN protocol for Internet of Things: A review and opportunities. In Proceedings of the 2017 2nd International Multidisciplinary Conference on Computer and Energy Science (SpliTech), Split, Croatia, 12–14 July 2017; pp. 1–6.
25. Queralta, J.P.; Gia, T.N.; Tenhunen, H.; Westerlund, T. Edge-AI in LoRa-based Health Monitoring: Fall Detection System with Fog Computing and LSTM Recurrent Neural Networks. In Proceedings of the 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 1–3 July 2019; pp. 601–604.
26. Drăgulescu, A.M.C.; Manea, A.F.; Fratu, O.; Drăgulescu, A. LoRa-Based Medical IoT System Architecture and Testbed. *Wirel. Pers. Commun.* **2020**. [CrossRef]
27. Petajajarvi, J.; Mikhaylov, K.; Hamalainen, M.; Iinatti, J. Evaluation of LoRa LPWAN technology for remote health and wellbeing monitoring. In Proceedings of the 2016 10th International Symposium on Medical Information and Communication Technology (ISMICT), Worcester, MA, USA, 20–23 March 2016; pp. 1–5.
28. Stočes, M.; Vaněk, J.; Masner, J.; Pavlík, J. Internet of Things (IoT) in Agriculture—Selected Aspects. *AGRIS On-Line Pap. Econ. Inform.* **2016**, *8*, 83–88.
29. Anabi, K.H.; Nordin, R.; Abdullah, N.F. Database-Assisted Television White Space Technology: Challenges, Trends and Future Research Directions. *IEEE Access* **2016**, *4*, 8162–8183. [CrossRef]
30. Oliver, M.; Majumder, S. Motivation for TV white space: An explorative study on Africa for achieving the rural broadband gap. In Proceedings of the 2nd Europe—Middle East—North African Regional Conference of the International Telecommunications Society (ITS): “Leveraging Technologies For Growth”; International Telecommunications Society (ITS): Aswan, Egypt, 2019.
31. DSA Worldwide Commercial Deployments, Pilots, and Trials. *Dyn. Spectr. Alliance* **2015**, 8736143, 1–23.
32. Mueck, M.; Noguet, D. TV White Space standardization and regulation in Europe. In Proceedings of the 2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), Chennai, India, 28 February–3 March 2011; pp. 1–5.
33. Andhini, N.F. Opportunities and Challenges of Using TV White Spaces—A Comparative Analysis of Approaches among U.S.A., U.K., and S. Korea. *J. Chem. Inf. Model.* **2017**, *53*, 1689–1699.
34. Dionísio, R.; Marques, P.; Rodriguez, J. Experimental Assessment of a Propagation Model for TV White Spaces. In *Wireless Internet: WICON 2014*; Mumtaz, S., Rodriguez, J., Katz, M., Wang, C., Nascimento, A., Eds.; Springer: Lisbon, Portugal, 2015; pp. 284–290. ISBN 978-3-319-18802-7.
35. ANACOM. ANACOM Aprova Nova Adenda ao Roteiro Nacional da Faixa dos 700 MHz. Available online: <https://www.anacom.pt/render.jsp?contentId=1563517> (accessed on 9 November 2020).
36. ANACOM. Calendário de Migração dos Emissores. Available online: <https://www.anacom.pt/render.jsp?categoryId=414983> (accessed on 9 November 2020).
37. ANACOM. Técnicas Inovadoras de Partilha do Espectro. Available online: <https://www.anacom.pt/render.jsp?categoryId=387636> (accessed on 10 September 2020).
38. Mekki, K.; Bajic, E.; Chaxel, F.; Meyer, F. A comparative study of LPWAN technologies for large-scale IoT deployment. *ICT Express* **2019**, *5*, 1–7. [CrossRef]
39. Pitu, F.; Gaitan, N.C. Surveillance of SigFox technology integrated with environmental monitoring. In Proceedings of the 2020 International Conference on Development and Application Systems (DAS), Suceava, Romania, 21–23 May 2020; pp. 69–72.
40. Kartakis, S.; Choudhary, B.D.; Gluhak, A.D.; Lambrinos, L.; McCann, J.A. Demystifying low-power wide-area communications for city IoT applications. In Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization—WiNTECH’16; ACM Press: New York, NY, USA, 2016; pp. 2–8.
41. Allahham, A.A.; Rahman, M.A. a Smart Monitoring System for Campus Using Zigbee Wireless Sensor Networks. *Int. J. Softw. Eng. Comput. Syst.* **2018**, *4*, 1–14. [CrossRef]
42. MatLog XBee 868 LP—Low Power Consumption. Available online: <https://www.matlog.fr/products/modules-xbee-serie-8-868mhz-low-power?lang=en> (accessed on 11 November 2020).
43. Telenor Connexion a Guide To Mobile Iot: How To Choose Between Lte-M and Nb-Iot for Global Deployments. 2020. Available online: <https://www.telenorconnexion.com/iot-insights/lte-m-vs-nb-iot-guide-differences/> (accessed on 11 November 2020).

44. Olimex NB-IOT Development Board with BC-66 Module. Available online: <https://www.olimex.com/Products/IoT/NB-IoT/NB-IoT-BC66/> (accessed on 11 November 2020).
45. Baker, S.B.; Xiang, W.; Atkinson, I. Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities. *IEEE Access* **2017**, *5*, 26521–26544. [CrossRef]
46. Pathinarupothi, R.K.; Durga, P.; Rangan, E.S. IoT-Based Smart Edge for Global Health: Remote Monitoring With Severity Detection and Alerts Transmission. *IEEE Internet Things J.* **2019**, *6*, 2449–2462. [CrossRef]
47. Cilfone, A.; Davoli, L.; Belli, L.; Ferrari, G. Wireless Mesh Networking: An IoT-Oriented Perspective Survey on Relevant Technologies. *Futur. Internet* **2019**, *11*, 99. [CrossRef]
48. Lavric, A.; Popa, V. Internet of Things and LoRa™ Low-Power Wide-Area Networks: A survey. In Proceedings of the 2017 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 13–14 July 2017; pp. 1–5.
49. ABIresearch for Visionaries LORAWAN AND NB-IOT: Competitors or Complementary. 2019. Available online: https://lora-alliance.org/sites/default/files/2019-06/cr-lora-102_lorawanr_and_nb-iot.pdf (accessed on 11 November 2020).
50. Ermes, M.; Pärkkä, J.; Mäntyjärvi, J.; Korhonen, I. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Trans. Inf. Technol. Biomed.* **2008**, *12*, 20–26. [CrossRef] [PubMed]
51. Vaportzis, E.; Clausen, M.G.; Gow, A.J. Older adults perceptions of technology and barriers to interacting with tablet computers: A focus group study. *Front. Psychol.* **2017**, *8*, 1–11. [CrossRef] [PubMed]
52. Parker, S.J.; Jessel, S.; Richardson, J.E.; Reid, M.C. Older adults are mobile too! Identifying the barriers and facilitators to older adults' use of mHealth for pain management. *BMC Geriatr.* **2013**, *13*, 1. [CrossRef] [PubMed]
53. The Things Industries. The Things Network. Available online: <https://www.thethingsnetwork.org/> (accessed on 1 February 2020).
54. Shenzhen Xin Yuan Electronic Technology Co., Ltd. TTGO LORA V1.3 868Mhz ESP32 Chip SX1276 Module 0.96 Inch OLED Screen WIFI and Bluetooth Development Board. Available online: http://www.lilygo.cn/prod_view.aspx?TypeId=50003&Id=1253&FId=t3:50003:3 (accessed on 15 January 2020).
55. Ruchir Sharma NEO-6M GPS Module. Available online: <https://create.arduino.cc/projecthub/ruchir1674/how-to-interface-gps-module-neo-6m-with-arduino-8f90ad> (accessed on 15 January 2020).
56. Shenzhen Xin Yuan Electronic Technology Co., Ltd. TTGO T-Beam V0.7 ESP32 868/915Mhz WiFi Wireless Bluetooth Module GPS NEO-6M SMA LORA 32 18650 Battery Holder. Available online: http://www.lilygo.cn/prod_view.aspx?TypeId=50033&Id=1237&FId=t3:50033:3 (accessed on 15 January 2020).
57. Analog Devices. ADXL335. Available online: <https://www.analog.com/en/products/adxl335.html> (accessed on 20 February 2020).
58. Espressif Systems. This document provides the specifications for the ESP32-WROOM-32D and ESP32-WROOM-32U Modules. 2019. Available online: https://www.espressif.com/sites/default/files/documentation/esp32-wroom-32d_esp32-wroom-32u_datasheet_en.pdf (accessed on 20 February 2020).
59. Texas Instruments Temperature Sensor LM35. Available online: <http://www.ti.com/product/lm35?qgpn=lm35> (accessed on 2 February 2020).
60. Ali, A.S.; Zanzinger, Z.; Debose, D.; Stephens, B. Open Source Building Science Sensors (OSBSS): A low-cost Arduino-based platform for long-term indoor environmental data collection. *Build. Environ.* **2016**, *100*, 114–126. [CrossRef]
61. Murphy, J.; Gitman, Y. Pulse Sensor Amped. Available online: <https://pulsesensor.com/products/pulse-sensor-amped> (accessed on 20 February 2020).
62. Design, N. LoRaWAN Architecture. Available online: <https://www.hackster.io/nootropicdesign/using-lorawan-end-devices-on-the-things-network-206a86> (accessed on 20 February 2020).
63. Westenberg, M. Single Channel LoRaWAN Gateway. Available online: <https://github.com/things4u/ESP-1ch-Gateway> (accessed on 25 March 2020).
64. Cayenne Cayenne LPP. Available online: <https://community.mydevices.com/t/cayenne-lpp-2-0/7510> (accessed on 6 October 2020).
65. Lekić, M.; Gardašević, G. IoT sensor integration to Node-RED platform. In Proceedings of the 2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia-Herzegovina, 21–23 March 2018; pp. 1–5.
66. The MathWorks, Inc. ThingSpeak for IoT Projects. Available online: <https://thingspeak.com> (accessed on 4 June 2020).

67. Semtech LoRa and LoRaWAN. Available online: <https://lora-developers.semtech.com/library/tech-papers-and-guides/lora-and-lorawan/> (accessed on 29 September 2020).
68. Chaccour, K.; Darazi, R.; Hassani, A.H.; Andrès, E. From Fall Detection to Fall Prevention: A Generic Classification of Fall-Related Systems. *IEEE Sens. J.* **2017**, *17*, 812–822. [CrossRef]
69. Yacchirema, D.; de Puga, J.S.; Palau, C.; Esteve, M. Fall detection system for elderly people using IoT and ensemble machine learning algorithm. *Pers. Ubiquitous Comput.* **2019**, *23*, 801–817. [CrossRef]
70. Microsoft Azure Machine Learning. Available online: <https://azure.microsoft.com/en-us/services/machine-learning/> (accessed on 9 October 2020).
71. The MathWorks, Inc. MathLab Machine Learning. Available online: <https://www.mathworks.com/company/newsletters/articles/developing-an-iot-analytics-system-with-matlab-machine-learning-and-thingspeak.html> (accessed on 9 October 2020).
72. OpenJS Foundation Node-RED Machine Learning. Available online: <https://flows.nodered.org/node/node-red-contrib-machine-learning> (accessed on 9 October 2020).

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Geospatial Assessment of the Territorial Road Network by Fractal Method

Mikolaj Karpinski ^{1,*}, Svitlana Kuznichenko ², Nadiia Kazakova ²,
Oleksii Frazee-Frazenko ² and Daniel Jancarczyk ^{1,*}

¹ Department of Computer Science and Automatics, University of Bielsko-Biala, 43-300 Bielsko-Biala, Poland

² Department of Information Technologies, Odessa State Environmental University, 65016 Odessa, Ukraine; skuznichenko@gmail.com (S.K.); kaz2003@ukr.net (N.K.); frazenko@gmail.com (O.F.-F.)

* Correspondence: mkarpinski@ath.bielsko.pl (M.K.); djancarczyk@ath.bielsko.pl (D.J.)

Received: 25 October 2020; Accepted: 13 November 2020; Published: 17 November 2020

Abstract: This paper proposes an approach to the geospatial assessment of a territorial road network based on the fractals theory. This approach allows us to obtain quantitative values of spatial complexity for any transport network and, in contrast to the classical indicators of the transport provisions of a territory (Botcher, Henkel, Engel, Goltz, Uspensky, etc.), consider only the complexity level of the network itself, regardless of the area of the territory. The degree of complexity is measured by a fractal dimension. A method for calculating the fractal dimension based on a combination of box counting and GIS analysis is proposed. We created a geoprocessing script tool for the GIS software system ESRI ArcGIS 10.7, and a study of the spatial pattern of the transport network of the Ukraine territory, and other countries of the world, was made. The results of the study will help to better understand the different aspects of the development of transport networks, their changes over time and the impact on the socioeconomic indicators of urban development.

Keywords: geoinformation technology; fractal dimension; territorial road network; box-counting framework; script Python; ArcGIS

1. Introduction

The development of an efficient transport infrastructure is one of the most pressing problems both for the whole territory of Ukraine and for other countries. As is known, a transport system has fairly high dynamics of development, and the effectiveness of its function depends on the quality of its organization and management. The presence of a large number of diverse properties and characteristics makes it impossible and inefficient to manually process large flows of input information. This increases the relevance of the development and implementation of automated approaches and analysis tools, as well as appropriate tools for working with geodata. The best modern tool for the analysis of spatial information is geographic information technology, which combines the functionality of traditional cartography and intelligent data processing in geographic information systems (GISs) [1].

An important aspect of the application of GISs is solving environmental problems, including terrain analysis, hydrological modelling, land use analysis and modelling, ecological modelling, and ecosystem service valuation [2]. GIS techniques and procedures have an important role to play in analyzing the multicriteria decision problems of planning and management. A variety of theoretical and methodological perspectives on multicriteria decision analysis (MCDA) in GISs have been suggested over the last 20 years [3]. Examples of spatial problems that are successfully addressed by integrating MCDA and GISs are suitability multicriteria analysis and site selection analysis [4]. GIS technology can also be useful in planning the development of engineering infrastructure facilities and the construction of environmentally hazardous facilities [5]. In [6], practical examples of the use of GISs in sustainable

urban planning are shown. GIS technologies allow one to observe and register changes in urban areas, manage the complex process of urban growth, and also help assess the impact of various multicriteria decision-making procedures for urban planning. Besides that, a GIS enables planners to develop and analyze urban transport development models and solve various transport-related problems.

One of the important indicators characterizing the transport system of any country is the transport provision of the territory. The transport provision level of the territory is traditionally estimated by the transport network density, the calculation of which involves using the coefficients of Botcher, Henkel, Engel, Goltz, and Uspensky [7]. A territory transport provision analysis example, based on the calculation of the Engel and Uspensky coefficients for assessing the impact of the transport system on economic security, is presented in [8]. The main drawback of the given coefficients is their use in the calculation formulas of the entire area of the territory instead of the inhabited area, which does not always adequately reflect the real picture. In order to take into account only the level of complexity of the transport network itself, and to not tie the indicator of transport provision to the area, it is proposed to calculate this indicator based on the fractals theory.

A fractal is a geometric figure that has the property of self-similarity; that is, it is composed of an infinite number of parts, each of which is similar to the whole figure [9]. The basic property of all fractal structures is their dimension. Although there is no exact definition of fractals, Mandelbrot B., the scientist who was the first to introduce the concept of fractals into science [10], gave his definition, stating that “A fractal is by definition a set for which the Hausdorff–Besicovitch dimension strictly exceeds the topological dimension” [11]. Unlike Euclidean geometry, in which dimensions are expressed in integers, the dimensions of fractal geometry can be expressed by fractional numbers between one and two in a two-dimensional space [12]. The bigger the non-integer value of the Hausdorff–Besicovitch dimension, the more irregular and complex the shape of the object.

The fractal geometry theoretical foundations development has contributed to the widespread use of fractals to describe various spatial phenomena in urban geography, urban morphology, landscape structure, and transport networks [13,14]. In [15], it was shown that the fractal geometry brings very effective apparatus to measure an object’s dimension and shape metrics in order to supply, or even substitute, other measurable characteristics of the object. Based on the fractal geographical interpretation, scientists are exploring the relationship between various aspects of urban space and the fractal dimension of cities and its changes over time [16,17]. In [18], it was shown how information on the fractal dimensions of the urban boundary and urban area can be used as a parameter for decision-making in the spatial development field, such as in the case of new residential area planning. The calculation of the fractal geometry of urban land use, performed in [19], made it possible to study the dynamics of urbanization and city expansion over recent years. Some aspects of the interpretation of the results of fractal analysis, as well as the analysis of scientific publications on the use of fractal models for urban analysis and planning, are presented in [20].

The calculation of fractal characteristics and research of the fractal pattern in the spatial structure of urban road networks provide extremely useful information for urban planning [21,22]. In particular, [23] used a modified box-counting method to describe the fractal properties of urban transportation networks and investigated the relationship between the mass size of cities and the complexity of their road systems. In [24], a box-counting method was applied to obtain a simple statistical model for determining the efficiency of filling the space of the transport system and identifying the variation in the level of fractality within the city itself and between parts of the city.

In this study, we propose a model for the geospatial assessment of the transport development of any land area based on the fractals theory. By transport development, we mean the provision of a territory with transport routes. This model will solve the problem that arises with the reliability of previous indicators that assess the level of transport development, based on the ratio of the transport network length and the territory area. Such areal indicators may give unreliable results when comparing the transport development of different countries such as, For example, the territory of Bolivia, which has an uneven population density associated with the presence of the Andean mountains, and the Amazonian

jungle, which contributes to the lack of transport networks in that part of the country. Such features should be taken into account, and when calculating transport development, only the area of inhabited areas should be taken into account.

The use of the territory transport development indicator on the basis of the roads fractal dimension will allow for excluding the use of the territory area value and take into account transport network structure peculiarities by itself, while also getting a geospatial estimate of the road network complexity.

Thus, the purpose of this study is to create a model for the quantitative geospatial assessment of the territorial road network, based on the fractal dimension of roads and its implementation in the form of GIS-oriented software (scripted geoprocessing tool) for ArcMap 10.7.

2. Statement of the Problem

2.1. Classical Indicators of the Transport Provision of Territories

The concept of transport development of territories is associated with such concepts as transport provision, which reflects the quality level of transport services for facilities and the population. Obviously, the more developed the network of communication lines in a particular region, the higher these indicators are. To assess transport provision, quantitative indicators are usually used that express the ratio of the length of tracks to a unit area of a territory, or to a certain number of residents, production volumes, or other factors. For example, there are the coefficients of Engel (1), Goltz (2), and Uspensky (3) [8]:

$$k_E = \frac{L}{\sqrt{SH}} \quad (1)$$

$$k_G = \frac{L}{\sqrt{SN}} \quad (2)$$

$$k_U = \frac{L}{\sqrt[3]{SHt}} \quad (3)$$

where L is the length of the transport network in km, S is the area of the developed territory in thousands of km², H is the total population in thousands of people, N is the number of settlements, and t is the total weight of the cargo sent to the territory.

When exploring the transport provision of a territory, it is not valid to compare the length of the transport networks with the area of the territory. It is assumed that moving away from the areal component and analyzing the level of complexity of the transport network itself may become more rational. A good example is a study of the calculation of the transport provision of a territory based on the theory of fractals, which is displayed in [25]. In contrast to the indicators in Equations (1)–(3), an indicator based on the fractal dimension excludes the area of the territory and takes into account the structural features and complexity of the road network itself, where the fractal dimension of each cell of the territory reflects a certain level of its density of the road network.

2.2. Calculation of the Transport Provision of Territories Based on the Fractals Theory

It is known that the Hausdorff–Besicovitch dimension is a natural way to determine the dimension of a subset in metric space. In three-dimensional Euclidean space, the Hausdorff–Besicovitch dimension of a finite set is zero, the dimension of a smooth curve is one, the dimension of a smooth surface is two, and the dimension of a set of nonzero volume is three. For more complex (fractal) sets, the Hausdorff–Besicovitch dimension may not be an integer [26].

The determination of the Hausdorff–Besicovitch dimension can be considered by measuring the dimension of a curve (Figure 1), which is covered by a fixed grid of squares with a side $\varepsilon > 0$. Each point

of a linear object belongs to one of the squares. Squares in which there are no points are not taken into account. The sum over all the squares covering the object (Hausdorff measure) has the following form:

$$m_p = \sum \varepsilon^p \quad (4)$$

where p is an arbitrary parameter.

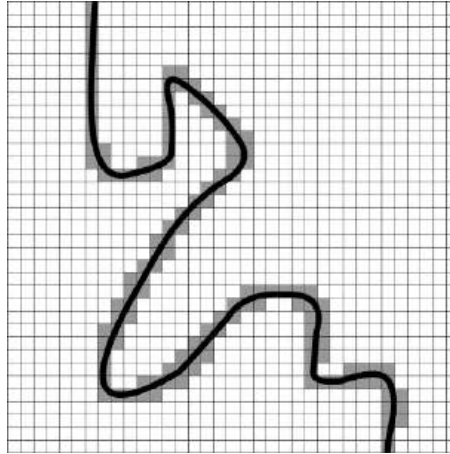


Figure 1. Determination of the Hausdorff–Besicovitch dimension using the ε coating.

There is a critical value of p_0 such that $\lim_{\varepsilon \rightarrow \infty} m_p = \infty$ for all $p < p_0$ and $\lim_{\varepsilon \rightarrow \infty} m_p = 0$ for all $p > p_0$. This value $p_0 = D_H$ is the value of the Hausdorff–Besicovitch dimension.

For example, for a square Q having a unit size with $\varepsilon = 1/10$, the number of square boxes covering Q equals $N(\varepsilon) = \varepsilon^{-2} = (1/10)^{-2} = 100$. The Hausdorff measure is $m_p(Q) = N(\varepsilon)\varepsilon^p = \varepsilon^{p-2}$. Let us say $\varepsilon \rightarrow 0$. Then, $m_p(Q) \rightarrow \infty$ for all $p < 2$, and $m_p(Q) \rightarrow 0$ for all $p > 2$. Thus, $D_H(Q) = \text{Dim}Q = 2$.

The dimension in general is determined by the law of similarity:

$$N(\varepsilon) \approx \frac{1}{\varepsilon^D} \quad (5)$$

By taking the logarithm of the right and left sides of Equation (5), we obtain

$$\ln N(\varepsilon) = -D \ln \varepsilon \quad (6)$$

$$D = \lim_{\varepsilon \rightarrow \infty} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)} \quad (7)$$

where $N(\varepsilon)$ is minimal number of the square boxes covering the object and ε is the square box size.

Let us give a definition of the transport provision of the territory based on the fractals theory [20].

We will consider the geospace (territory) as a two-dimensional space, and the maximum transport provision of the territory is the possibility of getting from each point of this territory to any other point in the shortest distance.

By destination points, we mean areal objects (points) whose dimensions (area) in this scale of research are negligible. Then, any territory can be represented as a finite number of such areal objects on a certain scale. A hit at any point of the area of the object is equivalent to falling into its center.

Since it is necessary to cover with points the entire investigated territory, it is advisable to choose the corresponding figure, a hexagon, as an area object. Thus, the transport development of the territory will be at a maximum when all the centers of the hexagons are interconnected by faces (a linear object) (Figure 2). These line features are actually roads on a territory map.

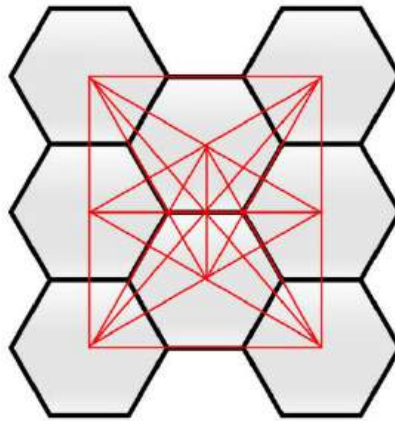


Figure 2. An example of maximum transport provision of territories.

Since any line to a certain scale is a fractal, the transport provision of the territory can be understood as the desire of the roads to occupy the entire area on which they are located. Therefore, the level of transport provision of the territory (TP) can be represented as the ratio of the fractal dimension of the studied road to the dimension of the area (equal to 2) or, taking into account Equation (7) [25], this can be expressed as

$$TP = \frac{1}{2} \lim_{\varepsilon \rightarrow 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)} \quad (8)$$

3. The Main Research

3.1. Research Methodology

The research methodology provided for the development of a model for the geospatial assessment of a territory's transport development is based on the fractals theory, in accordance with Equations (7) and (8).

In this study, the box-counting method was used to calculate the fractal dimension [9,27]. The essence of the method is as follows. The original dataset was split into a square box of size ε as a fixed grid (Figure 1). Next, the minimum number of square boxes $N(\varepsilon)$ that cover the original object was calculated. Calculations of $N(\varepsilon)$ were performed for various sizes of ε . For a small ε value, the square box number should behave like $\sim \varepsilon^{-D}$, and in this case, $\log N(\varepsilon) = D \cdot \log 1/\varepsilon$. According to the data obtained, we constructed a dependence of the following form:

$$\ln N(\varepsilon) = f(\ln(1/\varepsilon)) \quad (9)$$

Then, we calculated its slope d , defined as

$$D = -\lim_{\varepsilon \rightarrow 0} \frac{\ln N(\varepsilon)}{\ln(\varepsilon)} \quad (10)$$

To find the slope of Equation (8) in logarithmic scale, we had to build the general linear regression, expressed in the following form:

$$\ln N(\varepsilon) = \alpha + \beta \ln(1/\varepsilon) + \Delta \quad (11)$$

where Δ represents the error of a linear approximation.

The transport development of the territory was calculated as the ratio of the studied road dimension to the area dimension (i.e., a dimension equal to 2), based on Equation (8).

To implement the model, was created a Python script that calculated the fractal dimension for a polyline shapefile with a road network in the ArcMap program of the ArcGIS platform. In accordance

with the proposed model, the study area was covered with a network of hexagons of a given size. Then, using the box-counting method, the road network fractal dimension and the territory transport development within each hexagon were calculated, as in Equation (8).

The script execution result was a polygonal shapefile, the attribute table of which had a numerical field added with the territory transport development's calculated value for each polygonal object (hexagon).

As they were required for the script to work, base maps of the administrative area (boundary) and the road network were imported from the OpenStreetMap dataset in .shp format [28].

Modeling was performed for the following countries: Ukraine, Germany, and Bolivia. To compare their transport development level values, the same hexagon size was chosen, equal to 1000 km², so that most settlements had a single transport development and did not introduce additional errors into the study on the selected scale of countries. In the attributive tables of polyline layers of the road networks of countries, using an SQL query, only major trunk roads of international and regional importance were selected. These polyline features are tagged with highway = (motorway; trunk; primary; secondary).

Testing of the script was also carried out for a large-scale map, representing the territory of the city of Odessa (Ukraine). All road types were considered, including streets and roads within residential areas. The hexagon area was chosen as 0.25 km². The results obtained made it possible to compare the network fractal dimension indicators within the same settlement.

3.2. Results

For geospatial assessment of the transport provision of territories, a geoprocessing tool was created: an autonomous Python script that allowed one to calculate the fractal dimension for vector geodata with a linear geometry type.

ArcGIS contains a large library of geoprocessing tools for spatial modeling and the analysis of geographic data. The tools are grouped into toolboxes by the type of actions they perform (e.g., 3D Analyst, Spatial Analysis, and Cartography) [29,30]. To provide access to all standard ArcGIS geoprocessing tools via Python code, the ArcPy library was imported. In addition, the NumPy and SciPy libraries were imported into the script, allowing the use of high-level mathematical functions with data arrays, including data linear approximation [31,32].

Two inputs to the script were entered: a polygon shapefile with the administrative boundary of the study area and a polyline shapefile with the road network (Figure 3). In addition, it was necessary to indicate the area scale for fractal analysis (i.e., the area of one hexagon on the ground).

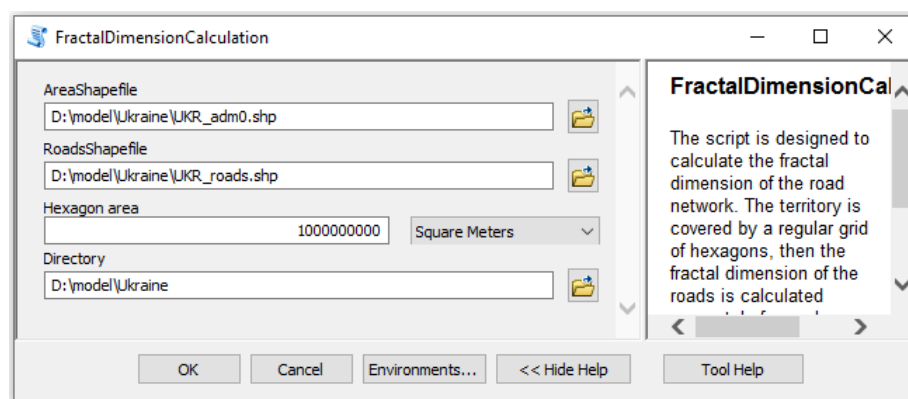


Figure 3. Python script interface.

When running the script using the GenerateTessellation() tool, a polygon shapefile (Hexagons.shp) was created with regular hexagons of a given area (Figure 4), which covered the study area and was then used to calculate the fractal dimension of the roads. The attribute table of this layer contained an ID column with a unique code for each hexagon object.

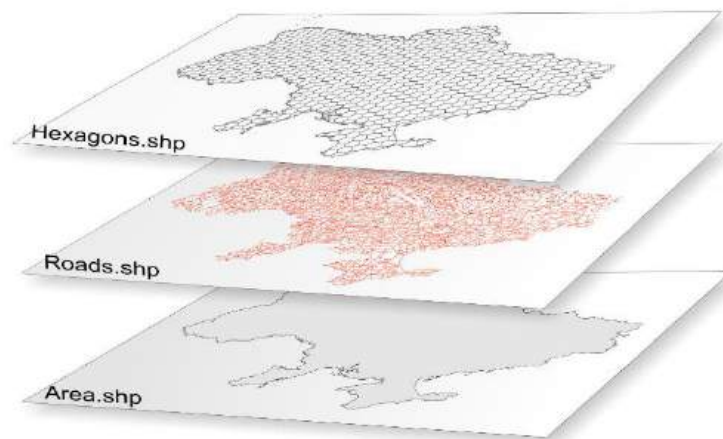


Figure 4. Hexagon polygon shapefile covering the study area.

Using the Intersect() tool, the intersection of linear objects of the Roads.shp road network with the Hexagons.shp hexagon layer was performed. According to these results, each section of the road was assigned the ID of the hexagon in which it was located. The combination of road sections belonging to one hexagon into a single object that had an ID that matched the hexagon ID was performed using the Dissolve() tool. The resulting polyline shapefile would then be used in a script called HexagonsDiss.shp.

The calculation of the fractal dimension was carried out in a cycle for each polygonal object from the Hexagons.shp attribute table. In each iteration of the cycle, using the CreateFishnet() tool, a grid was constructed with the cell size ϵ , and the number of squares $N(\epsilon)$ that covered all linear objects (roads) inside the current hexagon was calculated.

Calculations of $N(\epsilon)$ were carried out for various values of ϵ . There were five steps. Each hexagon was covered by a grid with 1, 4, 16, 64, and 256 squares, and the size of the square decreased by 1, 2, 4, 8, and 16 times, respectively (Figure 5).

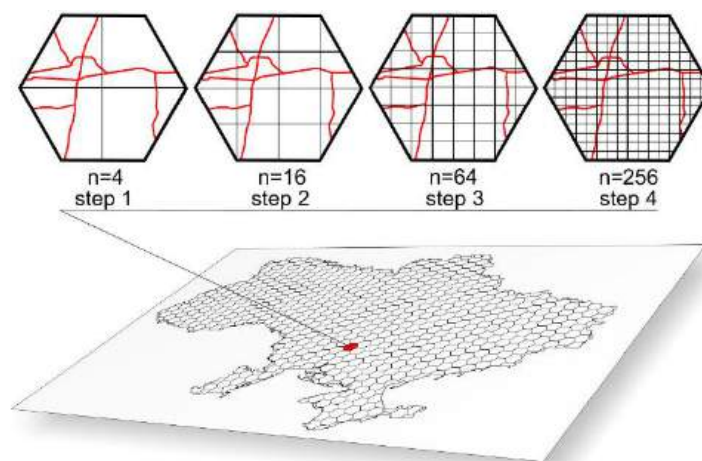


Figure 5. Steps for executing the box-counting method.

The calculation of the fractal dimension was carried out using linear approximation by the method of least squares. As an estimate of the fractal dimension, the slope value of the straight line was used. The following is a fragment of the program code for a script that calculates the coefficients α and β of the linear function in Equation (10) using the least squares method:

```
#Target function
fitfunc = lambda p, x: (p[0] + p[1] * x)
```

#Distance to the target function

errfunc = lambda p, x, y: (fitfunc(p, x) - y)

#Minimize the sum of squares of a set of equations; p1 = [α , β]

p1, success = optimize.leastsq(errfunc, [0,0], args = (logx, logy))

For the example shown in Figure 5, changes to $N(\epsilon)$ as a function of ϵ are given in Table 1. The regression line to estimate the fractal dimension is shown in Figure 6.

Table 1. The change of $N(\epsilon)$ versus different values of ϵ for the example in Figure 5.

Steps	Step 0	Step 1	Step 2	Step 3	Step 4
$N(\epsilon)$	1	4	11	28	65
ϵ	1	0.5	0.25	0.125	0.0625

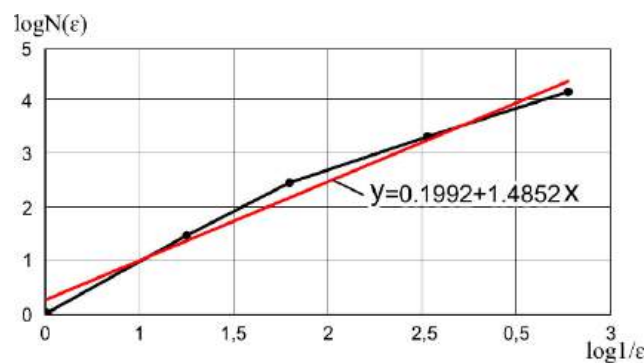


Figure 6. The regression line to estimate the fractal dimension.

After performing linear regression, an equation of the following form was obtained: $\ln(N(\epsilon)) = 0.1992 + 1.4852\ln(1/\epsilon) + \Delta$. The slope of this curve is equal to the box-counting dimension $d = 1.4852$. Accordingly, the level of transport provision, in accordance with Equation (11), is equal to $TP = 1.4852/2 = 0.7426$.

Figure 7 shows examples of roads of various fractal geometries and the values of the indicator of transport provision of the territory calculated for them.

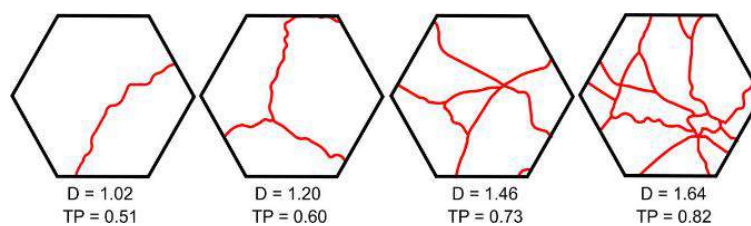


Figure 7. Fractal dimensions and transport utilization rates for different road patterns.

The script operation algorithm is shown in Figure 8. The result of the script was a vector layer of hexagons, the attribute table of which contained calculated values of the level of transport provision for each hexagon in the TP field. The resulting value fell in a range from zero to one.

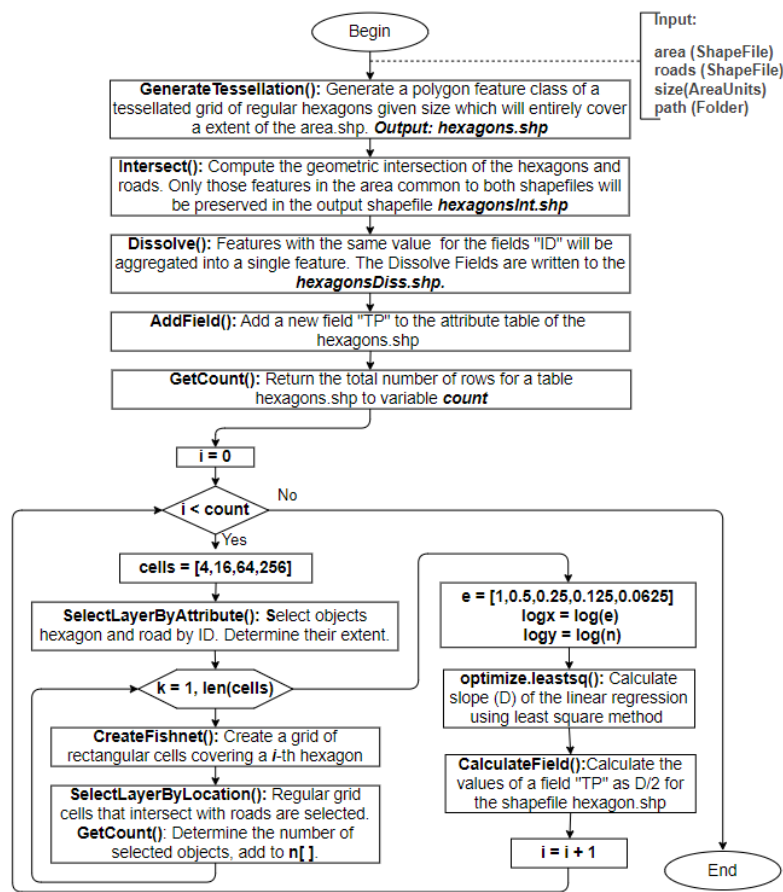


Figure 8. Python script algorithm.

4. Discussion

Using the created script, as an example of use, a map of the transport provision of Ukraine was constructed, shown in Figure 9.

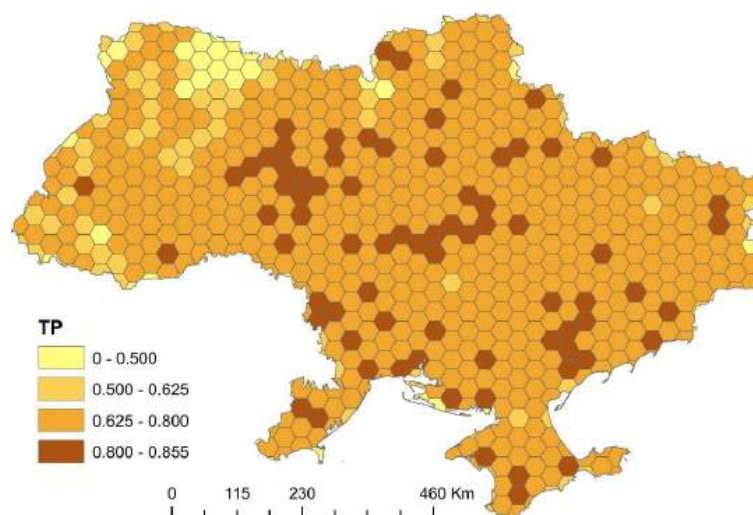


Figure 9. Ukraine road network transport provision level.

The area of the hexagonal cell of the fixed grid was chosen, equal to 1000 km^2 . The resulting vector map, consisting of 710 polygon features, was classified by the TP field, which contained the calculated values of the transport provision indicator in accordance with Equation (11).

The script execution time was 44 m 54 s. For simulation, a PC with modest technical characteristics was used: an Intel Pentium Processor G4400 (3 M Cache, 3.30 GHz) with an Intel HD Graphics 510 integrated graphics processor, 4.00 Gb DDR4 RAM, and an Asus H110M-K motherboard. When using a PC with better performance, one should expect a reduction in script execution time.

In the territory of Ukraine, low values of fractal dimensions are observed in the absence of settlements and an increase in its values in the vicinity of cities. At the same time, the relationship between the size of the population of the city and the growth of the area with a high fractal dimension index around it is clearly traced.

In percentage terms, in 11% of the Ukraine territory, transport provision is less than 0.5 (very low value with a sparse road network). In 14%, the level of transport provision is from 0.5 to 0.625 (low value with a sparse road network; that is, there are single primary roads crossing the territory). In 63%, the value of the indicator is from 0.625 to 0.8 (average value; that is, there is a network of primary roads between settlements). Lastly, in 12%, a high level of transport provision appears in a range from 0.8 to 0.855.

If we compare the transport provision of different countries around the world, we can see the relationship between the population density and the growth of the territory with a high fractal dimension index. In Figure 10, (a) shows the result of calculating the transport provision of Bolivia, located on the continent of South America, which has a low population density, and (b) shows Germany, a country in Europe with a high population density. The area of the hexagonal cell of the fixed grid was also chosen to be 1000 km². The simulation results are presented in Table 2. In 61% of Bolivia's territory, the indicator of transport provision is less than 0.5, including 47% of the territory that lacks a road network (i.e., the indicator is zero). This is due to the presence of large areas of Amazonian rainforests in this part of the country. The Bolivia territory is also crossed by the Andean mountains, which contributes to the scarcity of transportation lines. In the remaining 39% of the territory, various values of transport provision are observed, with a predominance of values in the range from 0.5 to 0.79. As for Germany, about 79% of its territory has a transport provision index above 0.625, including 34% above 0.75.

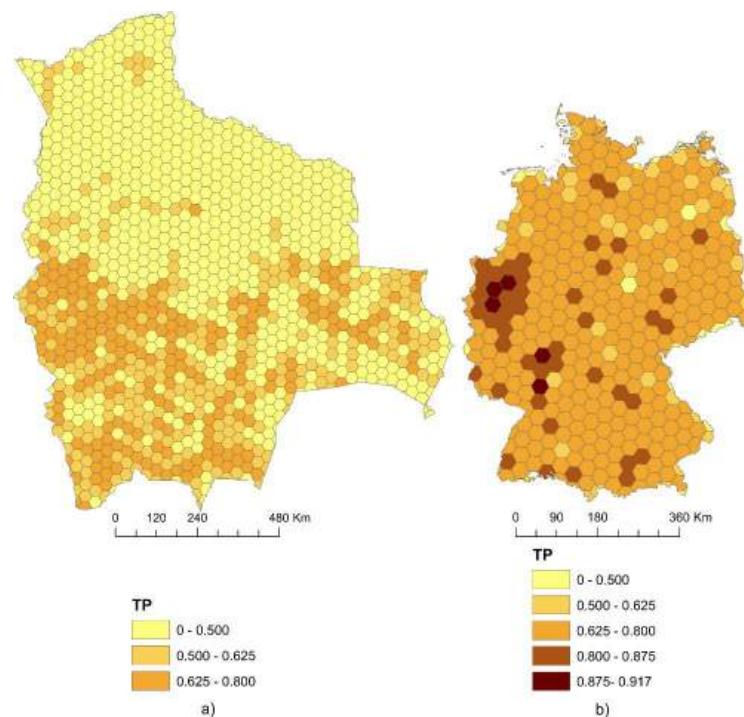


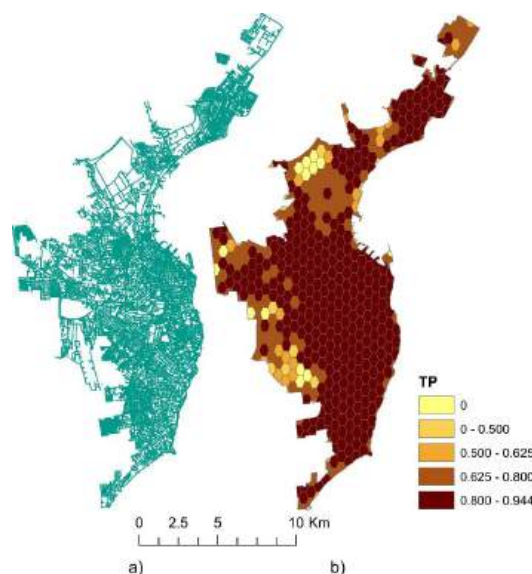
Figure 10. The level of transport provision of the territory (a) for Bolivia and (b) for Germany.

Table 2. Comparative characteristics of the calculation of transport development of a territory.

Specifications	Ukraine	Germany	Bolivia
Area of the country (km ²)	603,628	357,386	1,099,000
Length of paved roads (km)	78,660	45,395	31,216
Population (thousand people)	49,980	83,020	11,350
Hexagon area (km ²)	1000	1000	1000
Number of hexagons	710	429	1190
Script execution time (s)	2694	1652	2435
TP < 0.5 (very low) (%)	11	11	60
TP = (0.5 ÷ 0.625) (low) (%)	14	10	23
TP = (0.625 ÷ 0.8) (average) (%)	63	67	17
TP = (0.8 ÷ 0.875) (high) (%)	12	10	–
TP = (0.875 ÷ 1) (very high) (%)	–	2	–
Density of roads (km/km ²)	0.130	0.127	0.028
Coefficients of Engel (K _E)	14.3	8.3	8.8

Table 2 shows the comparative characteristics of the calculation of the transport provision of a territory for different countries. We also presented these countries' transport network density values, calculated as the highway length ratio to the territory area, and the values of Engel's coefficient, in accordance with Equation (1). Although it is not possible to compare the indicators of transport development calculated using Engel's coefficient (generalized indicator) and the fractal method (geospatial indicator), the main differences are still visible. The transport development of the territory of Germany, based on the value of Engel's coefficient (8.3), is almost 1.7 times less than the transport development of the territory of Ukraine (14.3) and is practically equal to the transport development of Bolivia (8.8). However, as can be seen in Figures 9 and 10, the values of transport development in Germany in most of the territory, calculated by the fractal method, have medium and high values, reaching 0.917 in individual hexagons, while for Ukraine, the maximum transport development value does not exceed 0.855.

The transport network fractal dimension calculation, in accordance with the proposed algorithm, can be performed for any land area, such as cities and towns. In this case, when using larger scale maps, one should choose a smaller hexagon size. Figure 11 shows the result of modeling the fractal dimension for the city of Odessa (Ukraine), the area of which is 163 km². The data source is the OpenStreetMap map service. All road types were considered, including streets and roads within residential areas. The hexagon area was 0.25 km².

**Figure 11.** Modeling results for the territory of the city of Odessa, showing (a) the transport network and (b) the transport development level.

The obtained modeling results (Figure 11b) allowed us to identify differences in the level of fractality within the city itself and compare the transport development indicator between parts of the city. A high or very high level of transport development (0.8–0.94) was found for the central and coastal regions, the most inhabited areas of the city of Odessa, while the suburban areas had a low level of transport development (0.5–0.625).

5. Conclusions

This paper proposes a geospatial approach to the study of transport provision based on the fractal dimension of roads, which allows one to obtain quantitative values of provision (level of spatial complexity) for a road network of any land territory. An algorithm for calculating the fractal dimension of roads based on the box-counting method was developed, and a scripted geoprocessing tool for ESRI ArcGIS 10.5 was created. The Python code for the FractalDimensionCalculation script has been uploaded to the GitHub repository, available to download for free (<https://github.com/kuznichenko-s/FractalDimension>).

Using the developed script, a study of the density of the road networks of the territory of Ukraine and other countries of the world was carried out. Additionally, the script was used to study urban and suburban areas (for example, the city of Odessa). The resulting map of the geospatial assessment of the transport development indicator made it possible to identify differences in the level of fractality within the city itself and compare the indicator of transport development between parts of the city.

The proposed quantitative model and script can be useful for scientists studying urban transport networks, in order to analyze the dynamics of their change over time, as well as to compare the level of complexity of transport networks in individual urban areas and their impact on socioeconomic indicators of urban development.

Given the high computational complexity of the algorithm, the development of parallel algorithms for calculating the fractal dimension (for example, on GPUs) can be a vector of subsequent research in this direction.

Author Contributions: Conceptualization, M.K., S.K., N.K., O.F.-F. and D.J.; Formal analysis, M.K. and D.J.; Methodology, M.K., S.K., N.K., O.F.-F. and D.J.; Software, S.K., N.K. and O.F.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hanks, R.R. *Encyclopedia of Geography Terms, Themes, and Concepts*; ABC-CLIO: Santa Barbara, CA, USA, 2011; p. 405.
2. Zhu, X. *GIS for Environmental Applications a Practical Approach*; Swales & Willis Ltd.: Exeter, Devon, UK, 2016; p. 471.
3. Maliene, V.; Grigonis, V.; Palevičius, V.; Griffiths, S. Geographic information system: Old principles with new capabilities. *Urban Des. Int.* **2011**, *16*, 1–6. [CrossRef]
4. Malczewski, J.; Rinner, C. *Multicriteria Decision Analysis in Geographic Information Science*; Springer: New York, NY, USA, 2015; p. 331. [CrossRef]
5. Kuznichenko, S.; Buchynska, I.; Kovalenko, L.; Gunchenko, Y. Suitable site selection using two-stage GIS-based fuzzy multi-criteria decision analysis. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2020.
6. Kuznichenko, S.; Kovalenko, L.; Buchynska, I.; Gunchenko, Y. Development of a multi-criteria model for making decisions on the location of solid waste landfills. *East. Eur. J. Enterp. Technol.* **2018**, *2*, 21–31. [CrossRef]
7. Kozhevnikov, S.A. Spatial and territorial development of the European North: Trends and priorities of transformation. *Econ. Soc. Chang. Facts Trends* **2019**, *12*, 91–109. [CrossRef]

8. Plotnikov, V.; Makarov, I.; Shamrina, I.; Shirokova, O. Transport development as a factor in the economic security of regions and cities. In Proceedings of the TPACEE-2018. E3S Web of Conferences, Moscow, Russia, 3–5 December 2018; Volume 91, p. 05032. [CrossRef]
9. Falconer, K. *Fractal Geometry: Mathematical Foundations and Applications*, 3rd ed.; Wiley: Chichester, UK, 2014; p. 400.
10. Mandelbrot, B. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* **1967**, *156*, 636–638. [CrossRef] [PubMed]
11. Mandelbrot, B. *The Fractal Geometry of Nature*; W.H. Freeman and Company: New York, NY, USA, 1982; p. 15.
12. Barnsley, M.F.; Rising, H. *Fractals Everywhere*; Academic Press Professional: Boston, MA, USA, 1993.
13. Frankhauser, P.; Pumain, D. Fractals and Geography. In *Models in Spatial Analysis*; ISTE Ltd.: London, UK, 2007.
14. Jiang, B.; Brandt, S. A fractal perspective on scale in geography. *Isprs Int. J. Geoinf.* **2016**, *5*, 95. [CrossRef]
15. Pastzo, V.; Marek, L.; Tucek, P.; Janoska, Z. Perspectives of fractal geometry in GIS analysis. *GIS Ostrav.* **2011**, *1*, 232–236.
16. Batty, M.; Longley, P.A. *Fractal Cities: A Geometry of Form and Function*; Academic Press Inc.: San Diego, CA, USA, 1994.
17. Shen, G. Fractal dimension and fractal growth of urbanized areas. *Int. J. Geogr. Inf. Sci.* **2002**, *16*, 419–437. [CrossRef]
18. Jevric, M.; Romanovich, M. Fractal Dimensions of Urban Border as a Criterion for Space Management. *Procedia Eng.* **2016**, *165*, 1478–1482. [CrossRef]
19. Chen, Y. Defining urban and rural regions by multifractal spectrums of urbanization. *Fractals* **2016**, *24*, 1650004. [CrossRef]
20. Purevtseren, M.; Tsegmid, B.; Indra, M.; Sugar, M. The fractal geometry of urban land use: The case of Ulaanbaatar city, Mongolia. *Land* **2018**, *7*, 67. [CrossRef]
21. Wang, H.; Luo, S.; Luo, T. Fractal characteristics of urban surface transit and road networks: Case study of Strasbourg, France. *Adv. Mech. Eng.* **2017**, *9*. [CrossRef]
22. Sun, Z.; Zheng, J.; Hu, H. Fractal pattern in spatial structure of urban road networks. *Int. J. Mod. Phys.* **2012**, *26*. [CrossRef]
23. Lu, Y.; Tang, J. Fractal dimension of a transportation network and its relationship with urban growth: A study of the Dallas—Fort Worth area. *Environ. Plan. B Plan. Des.* **2004**, *31*, 895–911. [CrossRef]
24. Sreelekha, M.G.; Krishnamurthy, K.; Anjaneyulu, M.V.L.R. Fractal Assessment of Road Transport System. *Eur. Transp.* **2017**, *5*, 65.
25. Korolev, A.; Yablokov, V. *The Model of Transport Development of the Territory Based on the Theory of Fractals*; Regional Studies; Publishing House of the Smolensk Humanitarian University: Smolensk, Russia, 2014; pp. 29–34.
26. Feder, J. *Fractals*; Plenum Press: New York, NY, USA, 1989.
27. Robinson, J.C. *Dimensions, Embeddings, and Attractors*; Cambridge University Press: New York, NY, USA, 2011; p. 205.
28. Arsanjani, J.J.; Zipf, A.; Mooney, P.; Helbich, M. *An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Applications*; Springer International Publishing: Cham, Switzerland, 2015; p. 324. [CrossRef]
29. Tateosian, L. *Python for ArcGIS*; Springer: Cham, Switzerland; Heidelberg, Germany; New York, NY, USA; Dordrecht, The Netherlands; London, UK, 2015; p. 544.
30. Lawhead, J. *Learning Geospatial Analysis with Python*, 3rd ed.; Packt Publishing Ltd.: Birmingham, UK, 2019; p. 447.
31. Bressert, E. *SciPy and NumPy: An Overview for Developers*; O'Reilly Media: Sebastopol, CA, USA, 2012; p. 68. ISBN 1449305466.
32. SciPy Community. SciPy Reference Guide Release 1.0.0. 2017. Available online: <https://docs.scipy.org/doc/scipy-1.0.0/scipy-ref-1.0.0.pdf> (accessed on 15 November 2020).

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

An Analysis of the Supply of Open Government Data

Alan Ponce ^{1,*}  and Raul Alberto Ponce Rodriguez ² 

¹ Institute of Engineering and Technology, Autonomous University of Cd Juárez (UACJ), Cd Juárez 32315, Mexico

² Institute of Social Sciences and Administration, Autonomous University of Cd Juárez (UACJ), Cd Juárez 32315, Mexico; rponce@uacj.mx

* Correspondence: alan.ponce@uacj.mx

Received: 17 September 2020; Accepted: 26 October 2020; Published: 29 October 2020

Abstract: An index of the release of open government data, published in 2016 by the Open Knowledge Foundation, shows that there is significant variability in the country's supply of this public good. What explains these cross-country differences? Adopting an interdisciplinary approach based on data science and economic theory, we developed the following research workflow. First, we gather, clean, and merge different datasets released by institutions such as the Open Knowledge Foundation, World Bank, United Nations, World Economic Forum, Transparency International, Economist Intelligence Unit, and International Telecommunication Union. Then, we conduct feature extraction and variable selection founded on economic domain knowledge. Next, we perform several linear regression models, testing whether cross-country differences in the supply of open government data can be explained by differences in the country's economic, social, and institutional structures. Our analysis provides evidence that the country's civil liberties, government transparency, quality of democracy, efficiency of government intervention, economies of scale in the provision of public goods, and the size of the economy are statistically significant to explain the cross-country differences in the supply of open government data. Our analysis also suggests that political participation, sociodemographic characteristics, and demographic and global income distribution dummies do not help to explain the country's supply of open government data. In summary, we show that cross-country differences in governance, social institutions, and the size of the economy can explain the global distribution of open government data.

Keywords: data science; open government data; governance and social institutions; economic determinants of open data

1. Introduction

Open data (OD) refers to information that has been generated by public or private entities and then published under a license that allows its use, reuse, and distribution freely [1]. Information collected and released from the public sector (e.g., transportation, pollution, agriculture, education, health, and census, among others) is referred to as Open Government Data (OGD) [2]. The public sector is considered one of the main contributors to the open data movement, due to the vast amount of information it generates [3]. According to [4], during recent years, there has been an increase in the number of countries that are adopting open data policies as part of their governmental agenda. Authors also argue that this trend is related to the potential benefits that OGD offers as a shared value (social and economic). From the social perspective, OGD is considered as a trigger of transparency, accountability, fighting against corruption, and the empowerment of citizens. The economic aspect of OGD is related to fostering innovation, enterprise opportunities, and job creation, because OGD is considered a production asset in the digital economy.

Additional evidence of global interest in the open data topic is the recent creation of different portals in which governments consolidate their data from different public entities (e.g., education, health, transportation) on a single website in order to release their data for free and collective use. Some examples of these portals developed by governments are the US (<https://www.data.gov/open-gov/>), Canada (<https://open.canada.ca/en/open-data>), Brazil (<http://www.dados.gov.br/>), Mexico (<https://datos.gob.mx/>), or the European Data Portal (<https://www.europeandataportal.eu/en>) funded by the European Commission. Other aspects related to open data interests are the initiatives constituted in conjunction with citizens, academics, and non-governmental organizations that are creating indexes, such as the Global Open Data Index (GODI) (<https://index.okfn.org/>), Open Data Barometer (ODB) (https://opendatabarometer.org/?_year=2017&indicator=ODB), Open Data Watch (ODW) (<https://opendatawatch.com/>), and Open Data Impact (ODI) (<https://odimpact.org/>) which are measuring the amount of data published by different governments around the world, as well as potential benefits and challenges (technical, legal, economic, social) that these public datasets (e.g., education, health, transportation) are generating in society.

Although there has been an increased interest in the phenomenon of open government data, most research has been conducted by applying qualitative methodologies through surveys, case studies, and desk research focusing on diverse topics, such as challenges and barriers in adopting and implementing open government data initiatives, and other qualitative studies have been focused on the release, provision, or value of these public datasets [5–7]. However, there is a gap in the literature for analyzing and measuring the determinants of the supply of open government data adopting a quantitative approach. This work pretends to fill this gap and contribute to the state of the art of open government data, providing a statistical analysis explaining countries' variabilities of the release of open government data through economic, social, and institutional factors. According to the Global Open Data Index published in 2016 by the Open Knowledge Foundation (OKF) (<https://okfn.org/>), there are significant differences across countries in the supply of open government data. In particular, Australia, the United Kingdom, and France obtain the highest GODI scores reported by the Open Knowledge Foundation (meaning that these countries contribute the most to the supply of open government data), while countries such as Myanmar, Barbados, Malawi, and Botswana obtained the lowest records on the GODI score (meaning that, in a global comparison, these countries contribute the least to the supply of open government data). This leads us to the following question: What explains the high heterogeneity in the global supply of open government data?

The objective of this paper is to provide an answer to this question by extending a single academic perspective, due to this research being based on an interdisciplinary approach aligned by the fields of data science and economics. The intersection point of these disciplines involves analyzing and estimating the determinants of the heterogeneity in the supply of global open government data by means of gathering information from different sources, featuring extraction and variable selection, modelling through the implementation of statistical methods, and explaining the effect and relationship of this heterogeneity. On the one hand, the data science approach is implemented in order to systematically create a data pipeline collected from different portals. This task is executed following a process of obtaining, scrubbing, exploring, modelling, and interpreting (OSEMN) the information collected from several sources. Then, we apply feature engineering in order to extract and analyze the data by way of a regression model that seeks to analyze the statistical association between some political and economic determinants of open government data. To estimate our model of regression analysis, we develop a sample with country cross-section data with data of the Global Open Data Index (GODI) for the year 2016. In this process, we solve empirical issues that arise in the regression analysis, such as multicollinearity, heteroscedasticity, missing data, outliers, and high dimensionality with our target variable (open government data).

On the other hand, economic theory is adopted to develop an empirical analysis (using our data pipeline) for the analysis of variables and their justification based on domain knowledge. Open government data is considered as a pure public good [8]; that is to say, we consider open

government data as satisfying two important properties: it is a non-excludable (once open government data is provided, then any person who seeks access can have access to that good) and it is a non-rival good (the consumption of open government data by some agent does not preclude the consumption of the same good by everyone else). Applying this theoretical framework, we test if political and social institutions such as civil rights, transparency, quality of democracy, and political participation, as well as economic and sociodemographic characteristics at the country level (such as the size of the economy, the efficiency of the government, the demand for Internet services, the median age of the population of a country, and the size of the population), can explain the global variability in the supply of open government data.

Using a cross-country regression model, our analysis provides evidence that cross-country differences in governance and social institutions such as civil liberties, government transparency, and the quality of democracy are statistically significant predictors of cross-country differences in the supply of open government data. Our estimates suggest that the government's transparency and civil liberties have a marginal positive and statistically significant effect on the supply of open government data. In our model, our variable that captures changes in the demand of web resources, that being the penetration of users (the proportion of Internet users over the country's population), is also positively and statistically significant in all of our estimated models.

In addition, our indicators of the efficiency of government intervention and economies of scale in the provision of public goods (analyzed through the variable population in each country) are also statistically significant predictors of cross-country differences in open government data. Our models also provide weak support to the hypothesis that open government data is a normal good; that is to say, countries with higher incomes are associated with higher levels of supply of open government data. Finally, our estimates suggest that political participation, the sociodemographic characteristics of citizens, demographic dummy variables, and dummy variables capturing the global distribution of income do not help to explain cross-country differences of the supply of open government data. In summary, we find evidence that cross-country differences in the supply of open government data are associated with the heterogeneity of social and political institutions, and economic factors are also correlated with the supply of open government data.

It is relevant to mention that the main limitation of our analysis is that we use cross-section data for our regression analysis, which limits the generality of our results. We decided to use data from the GODI for the year 2016 because this is the most up-to-date data on the GODI. Even if there is data for the Global Open Data Index for other years, the Open Knowledge Foundation has clearly stated that changes in methodology in the calculation of the GODI make unsuitable the comparison of data between 2016 and other years. This limits the study of what factors could explain the changes of the GODI over time. However, this limitation could be eased, as long as more data sets become available in the future that allow other forms of regression analysis, such as regression with panel data, that might improve the properties of estimation and hypothesis testing, as well as the generality of the results.

The structure of the paper is as follows. Section 2 includes a brief literature review, postulating the technical, social, economic, and political determinants of global open government data. Section 3 describes the data collection, the preparation process for our analysis, and the identification of the linear regression model. Section 4 contains the results of our analysis. Section 5 concludes the paper.

2. Literature Review

The adoption and implementation of open data is a socio-technological phenomenon that has been studied by different disciplines, trying to understand and estimate its dimensions and barriers [9–11]. For instance, the technical outlook is associated with the relevance of improving the data interoperability, quality, accessibility, usability, accuracy, platforms, and infrastructure needed in order to release open data [12–17]. The social stance refers to the empowerment that data offers to society such as, for example, the potential benefits that the information released by governments could produce through transparency and the accountability of citizens [18–21]. The economic point of view is related to possible impacts on

the economy that open data could offer through the creation of new businesses, products, and services, as well as employment [22–24]. This perspective also includes the crucial role that innovation plays as a driver of economic growth in the private and public sectors using open data [25–30]. The political perspective covers the strategies, policies, and impacts of the data released by the state [31–34]. The data published and freely accessible by public entities is referred to as Open Government Data (OGD). This particular kind of data plays an important role in the open data movement because it is considered as one of the main supporters through legislations such as the Open Data Directive (<https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information>) or global political initiatives like the Open Government Partnership (OGP) (<https://www.opengovpartnership.org/>). These political actions aim at increasing efficiency, promoting transparency, empowering citizens, and driving a knowledge-based economy through the release of data generated by public sectors. The study performed in [35] claims that the creation of open data policies is essential for defining the financial and technological infrastructure required, publication process definition, legal framework certainty, and political sustainability of open government data (OGD). The author also argues that open data policies should disseminate the economic and social value of OGD in order to stimulate the use and reuse of it in society. It is argued in [36] that the release of OGD is relevant because there are datasets collected by different sectors and for specific purposes (e.g., transportation, pollution, agriculture, education, health, and census). The authors also claim that OGD is a driver for innovation and business opportunities for society. Finally, they argue that the infrastructure of these data sets is paid by taxpayers; therefore, this information is considered a public good.

2.1. Determinants of the Supply of Open Government Data

In this section, we develop an analysis of the determinants of the supply of open government data. Hence, we explain the incentives of government policy makers to provide goods and services. As we mentioned before, in this paper, we consider open government data as a pure public good which has two important properties: it is a non-excludable (once a pure public good is provided, then any person who seeks to have access can have access to that good) and non-rival good (the consumption of the good by some agent does not preclude the consumption of the same good by everyone else). In our analysis, we consider that households and firms demand open government data because they find value in it; that is to say, households and firms might find open government data valuable because this information might help them to make rational and informed decisions. This information can also be used to foster their objectives, such as engaging in civic activities, political debates, and other activities regarded as desirable. For the case of households and in the case of firms, open government data might help them to make more efficient decisions (see [22,37,38]).

The literature on public economics has made important contributions to the study of the provision of this type of good, and this literature can be classified in two distinctive lines of research. The first line is the normative theory, and the second is the positive theory of the provision of public goods. The normative literature has emphasized that the preferences of households for private and public goods, the technology of production, and the costs of taxation that finance these goods are the main determinants of the provision of public goods (for a comprehensive review of the normative literature on public goods, see [39] and, more recently, [40]).

In contrast, the positive literature on public goods has emphasized that, in addition to household preferences and the technology of production, governments are suppliers of public goods, and candidates to public office are elected through a democratic process. It should be pointed out that the supply of open data has a production cost and, therefore, governments need to allocate public budgets for the collection and administration of data. This means that the allocation of budgets that allow the supply of open government data is subjected to a regular process of political negotiation in Congress and executive powers. Therefore, the provision of public goods could be explained by electoral incentives; political candidates seek to win public office, and they compete for votes in an election to form the government and make decisions over public policy (see [41,42]). For this reason,

politicians have incentives to provide public goods that benefit a significant proportion of voters in the electorate, with the hope of attracting votes in the election and maintaining political support while politicians hold office.

To be more specific about how electoral competition creates incentives for politicians to provide different levels of goods and services, we describe in detail the quid pro quo of models of electoral competition (see [41,42]). In a democracy, voters with different socio-demographic characteristics such as age, gender, and income might demand certain goods and services from the government because they benefit from these goods and services. Hence, candidates might consider that the distribution of demands of voters for goods and services from the government might be characterized by Figure 1. For the purpose of exposition, we assume g is the size of the provision of the public good. Hence, Figure 1 shows that there might be voters who would like the lowest size of the public good, equal to g_{min} (maybe because he or she does not benefit from the provision of this good), while g_{max} is the size of the provision of other voters who want the highest level of g in the distribution (maybe because the personal characteristics of these voters make them benefit a great deal from this good). Every point in the line shown in Figure 1 represents the ideal policy demanded by a certain voter, and the position g_{MV} is the ideal policy demanded by the median voter; that is to say, this position is the voter who is in the middle of the distribution of policies demanded by all voters participating in the election.

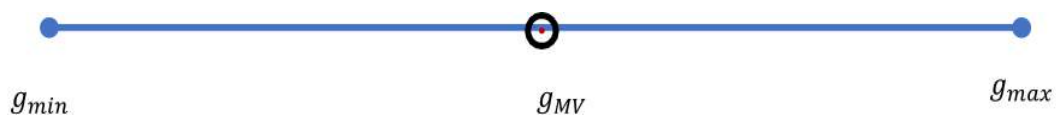


Figure 1. Distribution of policies demanded by voters.

Models of electoral competition consider that a voter will vote for the party that provides the policy that is closer to the voter's own preferences over policies. To see this, assume two parties, say parties 1 and 2, are competing for the vote of a particular voter, with the ideal policy given by g_h (shown in Figure 2). This voter will vote for party 1 because the policy position of this party is closest to the voter's own preferences for this public good or service.

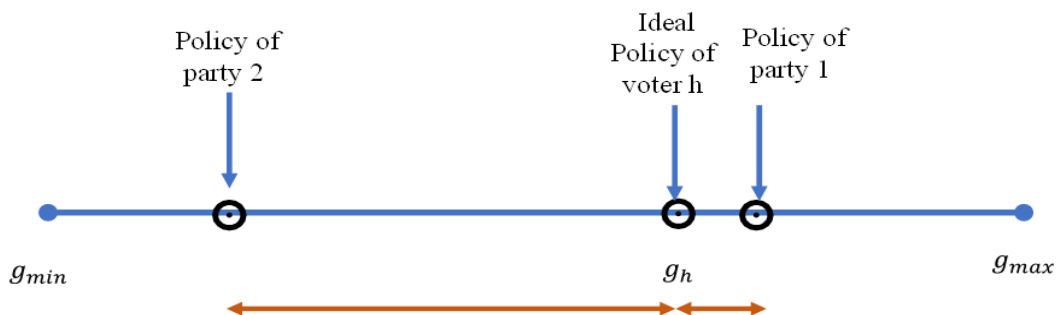


Figure 2. Policy positions of parties and the choice of the vote.

Hence, models of electoral competition predict that parties who want to maximize the expected votes to be received in the election should decide to offer the provision of the public good or service demanded (or desired) by the median voter (see Figure 3). That is, the government should select a level of its policy equal to $g = g_{MV}$. By doing so, the expected proportion of the vote for each party is 50% of the vote. If any party deviates from providing the median voter policy, then the party expects to receive a proportion of the vote lower than 50% of participating voters and will lose the election. Hence, models of electoral competition make a strong prediction that can be tested empirically: parties who want to maximize the electoral support from voters in the election should estimate the demand for goods and services of the median voter.

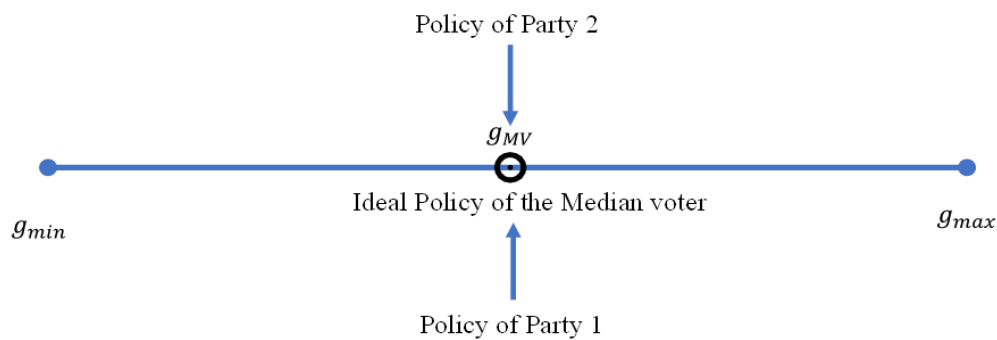


Figure 3. Prediction of models of electoral competition.

A further example of how this mechanism works is the following. Assume the demand of the median voter for goods and services from the government increases (perhaps because the median voter has more income and desires more government services). Then, policy makers in the government should increase the supply of government services to satisfy the demand of the median voter. It is relevant to mention that there is a great deal of evidence suggesting that public goods, such as education, health, and infrastructure (roads and bridges), which are provided by national and subnational governments, are correlated with the incentives of elections and political competition. For global empirical evidence of such a relationship covering 118 countries for three decades, see [43].

The positive literature on public goods has emphasized that, if elections matter and if there is perfect electoral competition (as a similar concept to the idea of perfect economic competition), then parties select the ideal provision of public goods of the median voter [42]. In this case, there is electoral accountability, meaning that elections create incentives for the provision of public goods that satisfy at least a majority of voters. However, if there is imperfect electoral competition and parties have preferences for public goods (that is to say, individuals controlling parties desire certain types and levels of public goods), then parties might select the ideal policy of activists inside parties or the ideal policy over public goods of a minority group of voters in the electorate (In an election, there could be imperfect electoral competition when a party does not have strong incentives to use in its policy positions to attract a majority of the votes in the election. This could be the case if voters do not vote for parties based on their policy positions, but instead on party identification (whether a voter self-identifies with a party), or the choice of the vote could be strongly determined by other non-policy issues, such as the personal characteristics of candidates (e.g., age, gender). For instance, if a significant proportion of voters (for purposes of exposition, let us say 30% of voters) vote for some party based on party identification, then this party does not necessarily select the policy of the median voter, because this party only needs another 21% of the vote to win the election. In this case, the policy positions of parties might be heavily influenced by the preferences, rather than the policies, of candidates or influential groups of voters inside of the party. Hence, there might be low electoral accountability and, if there is an increase in the demand of voters for open government data, the government might not respond by changing the supply of open government data. In this case, the demand of voters for that public good might not be satisfied) (see [44]).

In this latter case, there might be little electoral accountability, and the provision of public goods might be different relative to the ideal policy of the median voter in the economy (which might be considered as the ideal public policy for the society as a whole). Hence, the quality of the democratic process matters to determine the degree of electoral accountability and the size of the provision of public goods. Hence, for democracies with electoral accountability, if there is an increase in the demand of voters for open government data, well-functioning governments might increase the supply of open government data to satisfy the demand of voters for that public good (see [45–48]).

The demand for public goods can also be related with political participation. Citizens express their demand for public goods and services through voting in elections (see [40,42,49]). The political

participation of citizens can be observed by different political channels, such as voting in elections, attending meetings to express their demands for specific goods and services to elected representatives in congress and executive powers, or by contributing to political campaigns. Hence, we expect that more political participation leads to more accountability and better governance in democracies. Therefore, in democracies in which there is electoral accountability, higher political participation should lead to a better match between the public goods and services demanded by citizens and the supply of such services by the government.

A well-functioning democracy is also related with the civil liberties of citizens and the provision of public goods (for analysis along these lines, see [50]). Civil liberties are associated with the access of free printed and electronic media which provides relevant information to all citizens. Civil liberties can also be related with freedom of association and protest and, more relevant to our analysis, with political institutions that foster the free access to the Internet. Hence, we expect that more civil liberties are positively associated with less political restrictions to access the Internet and, therefore, more demand of content freely available on the Internet. In this line of thinking, the supply of open government data should also be positively related to transparency from the government. Transparency might help well-informed voters and economic agents to make rational decisions about the functioning of the government. Hence, voters might demand that their government provides useful information about the decisions of public policy by their elected officials. Therefore, in countries in which citizens demand more transparency, we could expect that their government satisfies this demand by providing more open government data.

The literature has also recognized that the sociodemographic characteristics of individuals, such as age, gender, and marital status, might be important determinants of the demand of public goods and services. (For a classical analysis on this issue see [51] and, for a literature review of the impact of socio-demographic characteristics on the size of government spending on public goods and other type of goods and services, see [50]) Hence, changes in the sociodemographic characteristics of households are related with changes in the demand for public goods and services (for instance, a change in the average age of individuals might lead to a change in the demand of certain services such as public education and public welfare). Therefore, changes in the sociodemographic characteristics of voters in a democracy might lead to changes in their demand for public goods, and governments have incentives to change their supply of public goods accordingly.

Most theoretical models that seek to explain the demand of private and pure public goods consider whether public goods are normal, neutral or inferior (see [52,53]). If a good is normal, then an increase in the income of households leads to an increase in the demand for such a good. If a good is inferior, then an increase in household income leads to a fall in the demand of such a good. When income increases, households might substitute the demand of low-quality goods for high-quality goods, which might explain why the demand of certain goods might fall as household income increases. A neutral good does not respond to changes in household income (see [54]). Hence, we could expect that an increase in the country's income might lead to an increase (or fall) in the demand of open government data if this good is normal or inferior, and governments might respond by increasing (or reducing) the supply of open government data.

In addition, most theoretical models that study the provision of pure public goods consider the size of the population as an important determinant of the provision of pure public goods (see [39,52,53]). As we mentioned before, a pure public good is non-excludable (once a pure public good is provided, then any person can have access to that good) and non-rival (the consumption of the good by some agent does not preclude the consumption of the same good by everyone else). Under these circumstances, the non-excludable property of a pure public good means that there could be economies of scale in the costs of providing a pure public good (see [55]). This means that the per capita costs of providing a pure public good decrease as the cost of public goods are shared among more people. In addition, an increase in the size of the population might also be associated with an increase in the size of the tax base that finances the provision of a pure public good (see [52,53]). This, in turn, leads to a fall

in the per capita cost of providing public goods, which increases the demand for this type of goods. Therefore, we could expect that an increase in the size of the population of a country might lead to an increase of the demand of open government data, and governments might respond by increasing the corresponding supply of such goods.

In summary, to guide the empirical analysis to be conducted in the following sections, we have relied on formal economic theory to characterize empirically verifiable tests of probabilistic determinants on the provision of open government data. In our analysis, we consider open government data as a pure public good because it satisfies two properties identified in the economic literature: the non-excludable and non-rival good properties. Based on the contributions of economic theory, we state several hypotheses about a probabilistic relationship between the supply of open government data by a country and the quality of democracy, the country's political participation, civil liberties and transparency, the sociodemographic characteristics of the country, the size of the economy, the size of the country's population, and the demand of content freely available on the Internet.

3. Material and Methods

3.1. Data Collection and Preprocessing

A common task in the economics and data science fields is the collection of datasets from different sources in order to discover knowledge, patterns, and trends. This activity represents some challenges, such as the diversity of the data structure, formats, and time consistency, among others. In this research, we adopted the OSMEN workflow methodology (shown in Figure 4), which is the acronym of obtain, scrub, explore, model, interpret, proposed by [56], to deal with these challenges. This methodology was proposed in the data science research community in order to systematically collect data and provide research transparency and result reproducibility.

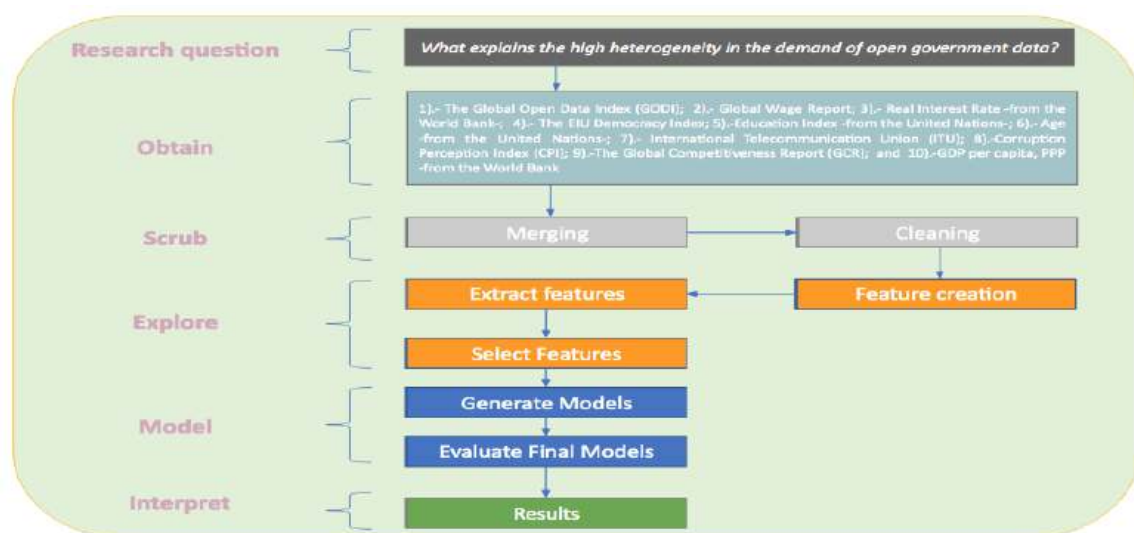


Figure 4. Illustration of the data science pipeline developed for our research.

Following this workflow to solve our research question, the first step was to select and gather the data from several sources, such as (1) the Global Open Data Index (GODI) from the Open Knowledge Foundation, (2) the global wage report from the International Labor Organization, (3) real interest rates from the World Bank, (4) the democracy index from the Economist Intelligence Unit, (5) the education index from the United Nations, (6) age data from the United Nations, (7) Internet use from the International Telecommunication Union, (8) the corruption perception index (CPI) from Transparency International, (9) the global competitiveness report (GCR) from the World Economic Forum, and (10) the gross domestic product (GDP) per capita and purchasing power parity (PPP) from

the World Bank. In this research, we defined the GODI indicator as our dependent variable (also called the response or target variable), and the other information extracted from these datasets became our independent variables (also called label variables).

Once the phase of data acquisition was completed, the next tasks were data integration and cleaning. The former involved analyzing and merging heterogeneous datasets. For example, some information was published in a long format and other datasets as wide formats containing different periods of time. The latter is associated with keeping consistency among datasets, such as homologating the names of countries because some datasets have different name labels (e.g., in some datasets, the country name was denoted as the United States of America or Venezuela, RB, and in other datasets, the country names appeared as the United States or Venezuela, respectively). Another aspect related to the cleaning process was to identify elements such as missing values, outliers, or other noise elements that could affect the quality of our model [57].

The next step was related to feature engineering, particularly feature creation, extraction, and selection. Some of our collected variables were categorical data. Therefore, we needed to create new variables in order to perform our models. This process is known as one-hot encoding in machine learning or dummy variables in econometrics. Then, we performed feature extraction, implementing principal component analysis (PCA), which is the process of dimensionality reduction from a large number of attributes without losing meaningful information by removing redundant data. This reduction process helped to identify features that could be more conducive to our analysis [58]. After performing PCA, our analysis indicated that there was multicollinearity in our independent variables. The topic of multicollinearity is an ongoing research area in feature engineering due to its implications when using diverse datasets [59–61]. For this reason, we include in the next section a variable selection process and a robustness check, based on an economic domain knowledge approach [62] and justified on the theoretical background introduced in the literature [63]. In order to complete a full sample with the desired variables for our empirical analysis, our final cross-sectional dataset was constituted by 18 variables and 49 observations during the year 2016. In the next section, we describe our model generation, interpretation, and results.

3.2. Empirical Analysis

In this section, we test if political and social institutions such as civil rights, transparency, quality of democracy, and political participation, as well as economic and sociodemographic characteristics at the country level (such as the size of the economy, the efficiency of the government, the demand for Internet services, the median age of the population of a country, and the size of the population), can explain the global variability in the supply of open government data. To test our hypotheses, we used one of the most popular tools in data science and economics: a linear regression analysis. This allowed us to estimate the marginal effect of how changes in independent variables (such as the size of the economy of a country, the sociodemographic characteristics of individuals in a country, civil liberties, and transparency) affect the supply of open government data. In the next model (see Equation (1)), we postulate that cross-country differences in the supply of open government data are associated with political and economic factors that affect the quality of democracy and the incentives of governments to provide open government data:

$$Od_i = \alpha + \beta'X + \varepsilon_i \quad (1)$$

In Equation (1), the differences in the supply of open government data across countries Od_i for $i = 1, \dots, I$, where the sub-index distinguishes the different countries in our sample, is explained by a set of k independent variables contained in the vector X (such as political participation, the size of the economy, socio-demographic characteristics of households, and indicators of demand for Internet services). The vector $\beta' = [\beta_1, \beta_2, \dots, \beta_k]$ represents the marginal effect of exogenous changes in X_i in our indicator of the supply of open government data; that is to say, $\frac{\partial Od_i}{\partial X_i} = \beta_i$. Our model also allowed us to test whether the marginal effect of X_i on open government data was statistically significant or not.

Finally, ε_i is a random error term from our model. To be more specific, the model we tested in our analysis is specified as follows:

$$Od_i = \alpha + \beta_1 GdpPPP_i + \beta_2 Dem_i + \beta_3 PolPar_i + \beta_4 Liberty_i + \beta_5 Trans_i + \beta_6 Pop_i + \beta_7 Age_i + \beta_8 GovEfficiency_i + \beta_9 IntPen_i + \varepsilon_i \quad (2)$$

Hence, in Equation (2), our model tests the postulated determinants of the supply of open government data discussed in Section 2.1 of this paper. Therefore, we were interested in testing whether the supply of open government data is associated with changes in the economic size of the country (see $GdpPPP_i$, which is the purchasing power parity of the gross domestic product of a country and affects the demand of open government data and its corresponding supply). The effect of the quality of democracy of a country is defined as Dem_i , which is a metric that measures the function, state, and trust of political freedoms and civil liberties through pillars such as political participation of citizens (defined as $PolPar_i$), civil liberties (defined as $Liberty_i$), and the transparency of a country's government (see $Trans_i$).

Other determinants in our model are the size of the population (defined as Pop_i) and the sociodemographic characteristics in a country (characterized by the median age of the population of a country, Age_i). Besides that, there is the efficiency of the government (labelled as $GovEfficiency_i$), which is an index that measures and compares per country the burden of government regulation, legal framework performance, and transparency of public policies, and our indicator of demand for Internet services (defined by $IntPen_i$, or Internet penetration), which is the number of Internet users as a proportion of the population in each country. To estimate our model of regression analysis, we developed a sample with country cross-section data. Our variable for open government data is the global open government data index (defined as Od_i), published by the Open Knowledge Foundation (OKF), which provides cross-country differences on the supply of open government data.

To estimate the model in Equation (2), we used a cross-section analysis with data on the GODI for the year 2016. A well-identified regression analysis needs to consider the possibility of endogeneity, which might bias the estimates of the model. Endogeneity might arise when changes in the independent variables X might be correlated with changes in the dependent variable Y , and changes in Y might also lead to changes in the X variables. In this case, the marginal effects in the regression model in (2) would not be properly identified. A standard way to solve this issue is to use the independent variables X , but lagged for one period (for technical analysis on this issue, see [64]). Hence, to avoid the possibility of endogeneity in our estimates, we used data for the year 2015 for our control variables; that is, we used the lagged observation for the control variables, such as the size of the country, the quality of democracy, political participation, civil liberties, transparency, the size of the population, the sociodemographic characteristics in a country, and efficiency of the government. In this case, changes in X could be correlated with changes in the dependent variable Y , but not the opposite case.

In summary, our assessment criteria for our empirical analysis was constituted as follows. First, we used theoretical analysis from the literature on economics on the main determinants of public goods and services provided by governments to identify control variables of the regression analysis (as a way to determine the structure of the X variables in the regression analysis, seen in Section 2.1). The theoretical analysis provided a rationale for a probabilistic link between the independent variable (GODI) and the explanatory variables of the model in Equation (2). Second, we used standard techniques of regression analysis to determine the best way to obtain unbiased and efficient estimators of the marginal effects of the independent variables over the dependent variable GODI. Third, we conducted a robustness test of our estimates and hypothesis testing by estimating several models (see Models I, II, III, IV and V in Table 1 in the following section) to test whether our results were sensitive to specific forms of linear regression analysis.

Table 1. Results of OLS estimators with heteroscedasticity-consistent standard errors.

Variable	Supply of Open Government Data GODI Score (I)	Supply of Open Government Data GODI Score (II)	Supply of Open Government Data GODI Score (III)	Supply of Open Government Data GODI Score (IV)	Supply of Open Government Data GODI Score (V)
C	11.04	50.17	55.5180	53.3683	55.9864
Gdppp	0.0002 *	0.0002	0.0002	0.0002	0.0002
	(1.6998)	(1.5876)	(1.6380)	(1.3166)	(1.4044)
Liberty	0.4111 **	0.4974 ***	0.5252 **	0.4306 *	0.4679 *
	2.1968	2.5259	2.4246	1.7019	1.6761
Dem	−0.4726	−1.2298 *	−1.3926 *	−1.4452 *	−1.4923 *
	−1.1856	−1.8287	−1.6874	−1.8960	−1.6831
Polpar	0.1581	0.1925	0.1978	0.2730	0.2410
	0.7461	0.8889	0.8105	1.1275	0.8072
Age	−0.3381	−0.3287	−0.2821	0.0522	0.0645
	−0.9542	−0.9782	−0.7732	0.1132	0.1178
Transparency	15.8476 *	17.48105 **	16.1369 *	13.9994	14.4925
	1.7610	1.9572	1.7938	1.5910	1.6067
Efficiency	−19.6248 **	−22.12 **	−20.8876 **	−18.6995 *	−19.6564 *
	−2.0313	−2.2669	−2.1968	−1.8295	−1.94
Population	2.5681 ***	3.1520 ***	3.3581 ***	3.3492 ***	3.3951 ***
	3.4713	3.50	3.3188	3.0274	2.8543
Internet-Penetration	0.5345 ***	−0.0142	−0.1968	−0.2213	−0.2840
	2.9659	−0.0474	−0.4547	−0.4952	−0.5341
Dem*Internet-Penetration		0.0088 *	0.0114 *	0.0116 *	0.0129 *
		1.8122	1.8913	1.8565	1.8966
High income			−0.015		−2.8034

Table 1. Cont.

Variable	Supply of Open Government Data GODI Score (I)	Supply of Open Government Data GODI Score (II)	Supply of Open Government Data GODI Score (III)	Supply of Open Government Data GODI Score (IV)	Supply of Open Government Data GODI Score (V)
			−0.0009		−0.1298
Upper Middle Income			4.8105		1.6383
			0.3667		0.1074
Lower Middle Income			0.6255		−0.0153
			0.0842		−0.0016
East Asia Pacific				3.9374	2.9713
				0.3864	0.2726
Europe Central Asia				−1.2709	−2.5741
				−0.1179	−0.2231
Latin America Caribbean				8.0636	5.0072
				0.6578	0.3605
Middle East North Africa				0.8853	2.6890
				0.0724	0.2168
North America				2.0636	0.7462
				0.1751	0.0595
Adjusted R-squared	0.6387	0.6605	0.6689	0.6792	0.6824
F-statistic	7.66 ***	7.39 ***	5.43 ***	4.65 ***	3.58 ***
Sample	49	49	49	49	49

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. All tests are two-tailed, and t -tests are below the corresponding estimates (numbers in parenthesis correspond to the t -test).

4. Results

We estimated our model with a set of different independent variables for a robustness check of our analysis. Our estimation technique used ordinary least squares with heteroscedasticity-consistent standard errors [65–68]. This technique allows having credible estimates of the probability distribution functions of the marginal effects β_i associated with the independent variable X_i and, therefore, credible hypothesis testing. Table 1 shows our empirical results, with each column considering different econometric specifications. Model I used our basic set of explanatory variables described in Equation (2). Model II incorporated an interaction term between democracy and Internet penetration, which allowed us to test if the quality of democracy leads to a differentiated response of countries to an increase in the demand of Internet services. Model III expanded Model II by incorporating geographical dummies. Model IV incorporated dummies related to the world's distribution of income, and Model V incorporated geographical and global income distribution dummies.

Our estimates showed that cross-country differences in governance and social institutions, such as civil liberties, the quality of democracy, and the degree of government transparency, are statistically significant predictors of cross-country differences in the supply of open government data. To see this (as shown in Table 1), note that all models show that the marginal effect of liberty on the supply of open government data is positive and statistically significant (When we refer to the marginal effects of a change of one variable and the variable of the GODI, this should be interpreted as a probabilistic marginal effect that the increase of one variable is correlated with increases or reductions, depending on the sign of the coefficient, the variable of the GODI. Hence, our analysis does not show causality, but a probabilistic correlation) (at different levels of value p). In addition, the government's transparency also has a marginal positive and statistically significant effect on the supply of open government data in Models I, II, and III, while the coefficient of the quality of democracy is statistically significant in Models II, III, IV, and V (see Table 1), and the interaction term between democracy and the penetration of users of the Internet is positive and statistically significant in all models in which we considered this interaction term.

In addition, our variable that captured changes in the demand of web resources—that is, the variable of penetration of users in our model, or the proportion of Internet users over the country's population—is positively and statistically significant in all of our estimated models. In Models II, III, IV, and V, we included an interaction term to test whether differences in the quality of democracy lead to different responses by governments to changes in the demand of use of the Internet, termed Dem*Penetration. The marginal effect of the interaction term is positive and statistically significant in all of our models in which we used this variable. That is to say, all countries in the sample have a marginal positive response in the supply of open government data when they observe increases in the demand of Internet services, but countries with higher qualities of democracy supply more open government data than countries with weaker democracies. This result confirms that governance is an important determinant of the cross-country differences in the supply of open government data.

However, in our models, political participation and the sociodemographic characteristic of citizens (in our models, the average age of citizens) were not statistically significant in any of the estimated models. The marginal effect of political participation had a positive effect (as expected) while the average age of citizens had a negative effect (as intuition might suggest) in Models I, II, and III, but a positive effect in Models IV and V. In addition, our models provided, at best, weak support to the hypothesis that open government data is a normal good; that is to say, countries with higher incomes are associated with higher levels of supply of open government data, since the positive income effect on open government data is significant only in Model I. Once we included demographic dummy variables and dummy variables of the global distribution of income, the marginal effect of the size of the economy on the supply of open government data was positive, but not statistically significant (see Models II, III, IV, and V).

In our analysis, two variables were associated with the efficiency of government intervention (this being the variable efficiency in Table 1) and economies of scale in the provision of public goods,

analyzed through the variable population. Our estimates showed that the efficiency of the government of a country has a negative and statistically significant marginal effect on the supply of open government data in all of our models. One possible explanation of this outcome is that an increase in the efficiency of government intervention might lead to more resources available to be spent by governments. However, those available resources are spent on other governmental programs that might have a high electoral impact relative to the choice of supplying more open government data. Therefore, a more efficient government increases spending in some programs (with high electoral impacts) and reduces spending in other programs (with relatively low electoral impacts). In addition, the size of the population, which is considered as a variable associated with the economies of scale in the provision of public goods, had the expected sign; that is, there was a positive and statistically significant marginal effect of the population on the supply of open government data in all of our estimated models. As we mentioned before, the non-excludable property of open government data means that there could be economies of scale in the costs of providing this pure public good. This means that the per capita costs of providing a pure public good are decreasing as the cost of public goods are shared among more people (i.e., as the size of the population in a country increases). This, in turn, leads to a fall in the per capita cost of providing public goods, which increases the demand for this type of goods. Hence, governments respond by increasing the corresponding supply of such goods.

Our models, through the individual t-test (the t-tests are displayed in Table 1 in parenthesis), also suggest that demographic dummy variables and dummy variables capturing the global distribution of income are not statistically significant to explain the cross-country differences in the supply of open government data in our sample (see Models III, IV, and V). However, the F statistic shows that, jointly, all independent variables considered in Models I–V are statistically significant to explain the cross-country differences in the supply of open government data in our sample. To see this, we tested if $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \dots \beta_k = 0$; that is, we tested if the joint effect of our independent variables helped to explain cross-country differences in the supply of open government data, and the corresponding F statistic showed that we rejected the null hypothesis shown above. Therefore, Models I–V as a whole are statistically significant (see Table 1).

5. Conclusions

We developed a cross-section analysis to provide tests for the institutional, political, and economic determinants of cross-country differences in the supply of open government data. We consider open government data as a pure public good and, therefore, it satisfies two properties: open government data is a non-excludable good (once open government data is provided, then any person who seeks to have access can have access to that good), and it is a non-rival good (the consumption of the good by some agent does not preclude the consumption of the same good by everyone else). We used an index of the supply of open government data and estimated a cross-section regression model to analyze the cross-country differences in civil rights, transparency, quality of government, the size of the economy, the size of the population, political participation, and sociodemographic characteristics. These can explain cross-country differences in the supply of open government data. We also conducted a robustness check of our analysis by estimating five different models of regression analysis that also include dummy variables associated with geographic heterogeneity and the global distribution of income.

Our analysis provided evidence that cross-country differences in governance and social institutions such as civil liberties, government transparency, and the quality of democracy are statistically significant predictors of cross-country differences in the supply of open government data. Our estimates suggest that civil rights and the transparency of government in each country have a marginal positive and statistically significant effect on the supply of open government data. In addition, our variable that captured changes in the demand of web resources—that is, penetration of users, or the proportion of Internet users over the country's population—was both positive and statistically significant in all of our estimated models. Our analysis also suggests a differentiated response of governments to changes with

the demand of web resources. In particular, if there is an increase in the demand of Internet services, countries with higher qualities of democracy supply more open government data than countries with weaker democracies. This result also shows that the level of governance in each country is an important determinant of the cross-country differences in the supply of open government data.

In our analysis, we included two variables that were associated with the efficiency of government intervention and with economies of scale in the provision of public goods. In this paper, we found evidence that the efficiency of government intervention and the economies of scale can also explain the cross-country differences in open government data. The efficiency of government intervention has a negative marginal effect on open government data. One possible explanation of this outcome is that an increase in the efficiency of government intervention might lead to more resources available to be spent by governments. However, those available resources might be spent on other governmental programs that might have a higher electoral impact relative to the choice of supplying more open government data, thus explaining the negative marginal effect of efficiency on the supply of open government data. In addition, the size of the population in a country was considered as a variable associated with economies of scale in the provision of public goods. Our analysis showed that there is a positive and statistically significant marginal effect of the population on the supply of open government data in all of our estimated models. As we mentioned before, the non-excludable property of open government data means that there could be economies of scale in the cost of providing open government data. This means that the per capita costs of providing a pure public good decrease as the cost of public goods is shared among more people. This, in turn, leads to a fall in the per capita cost of providing public goods, which increases the demand for this type of goods. Hence, governments respond by increasing the corresponding supply of open government data.

In addition, our models provide weak support to the hypothesis that open government data is a normal good; that is to say, countries with higher incomes are associated with higher levels of supply of open government data. However, in our models, political participation, the sociodemographic characteristics of citizens, demographic dummy variables, and dummy variables capturing the global distribution of income did not help to explain cross-country differences of the supply of open government data.

It is relevant to mention that the main limitation of our analysis is that we used cross-section data for our regression analysis, which limited the generality of our results. We decided to use data from the GODI for the year 2016 because this is the most up-to-date data on the GODI. Even if there is data for the Global Open Data Index for other years, the Open Knowledge Foundation has clearly stated that changes in methodology in the calculation of the GODI make unsuitable the comparison of data between 2016 and other years. This limits the study of what factors could explain the changes of the GODI over time. However, this limitation could be eased as long as more data sets become available in the future that allow other forms of regression analysis, such as regression with panel data, that might improve the properties of estimation and hypothesis testing, as well as the generality of the results.

Author Contributions: Conceptualization, A.P. and R.A.P.R.; methodology, A.P. and R.A.P.R.; validation, A.P. and R.A.P.R.; formal analysis, A.P. and R.A.P.R.; investigation, A.P. and R.A.P.R.; data curation, A.P. and R.A.P.R.; writing—original draft preparation, A.P. and R.A.P.R.; writing—review and editing, A.P. and R.A.P.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Both authors would like to thank Universidad Autónoma de Ciudad Juárez for the institutional support that made this research possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wainwright, T.; Huber, F.; Rentocchini, F. Open Wide? Business Opportunities and Risks in using Open Data eprints.soton.ac.uk. 2014. Available online: <https://eprints.soton.ac.uk/366901/> (accessed on 27 October 2020).
2. Attard, J.; Orlandi, F.; Scerri, S.; Auer, S. A systematic review of open government data initiatives. *Gov. Inf. Q.* **2015**, *32*, 399–418. [CrossRef]
3. Hardy, K.; Maurushat, A. Opening up government data for Big Data analysis and public benefit. *Comput. Law Secur. Rev.* **2017**, *33*, 30–37. [CrossRef]
4. Manolea, B.; Cretu, V. The influence of the Open Government Partnership (OGP) on the Open Data discussions. European Public Sector Information Platform. 2013. Available online: https://www.europeandataportal.eu/sites/default/files/2013_the_influence_of_the_ogp_on_the_open_data_discussions.pdf (accessed on 27 October 2020).
5. Zuiderwijk, A.; Helbig, N.; Gil-García, J.R.; Janssen, M. Special issue on innovation through open data-A review of the state-of-the-art and an emerging research agenda: Guest editors' introduction. *J. Theor. Appl. Electron. Commer. Res.* **2014**, *9*, I–XIII. [CrossRef]
6. Hossain, M.A.; Dwivedi, Y.K.; Rana, N.P. State-of-the-art in open data research: Insights from existing literature and a research agenda. *J. Organ. Comput. Electron. Commer.* **2016**, *26*, 14–40. [CrossRef]
7. Safarov, I.; Meijer, A.; Grimmelikhuijsen, S. Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Integr. Psychiatry* **2017**, *22*, 1–24. [CrossRef]
8. Sa, C.; Grieco, J. Open Data for Science, Policy, and the Public Good. *Rev. Policy Res.* **2016**, *33*, 526–543. [CrossRef]
9. Zuiderwijk, A.; Janssen, M.; Choenni, S.; Meijer, R.; Sheikh, A.R. Socio-Technical Impediments of Open Data. *Res. Gate* **2012**, *10*, 156–172.
10. Zuiderwijk, A.; Janssen, M. Open Data Policies, Their Implementation and Impact: A Framework for Comparison. *Government Information Q.* 2014. Available online: <https://www.sciencedirect.com/science/article/pii/S0740624X13001202> (accessed on 27 October 2020).
11. Janssen, M.; Charalabidis, Y.; Zuiderwijk, A. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Inf. Syst. Manag.* **2012**, *29*, 258–268. [CrossRef]
12. Vetrò, A.; Canova, L.; Torchiano, M.; Minotas, C.O.; Iemma, R.; Morando, F. Open data quality measurement framework: Definition and application to Open Government Data. *Gov. Inf. Q.* **2016**, *33*, 325–337. [CrossRef]
13. Kučera, J.; Chlapek, D.; Nečaský, M. Open Government Data Catalogs: Current Approaches and Quality Perspective. In *Technology-Enabled Innovation for Democracy, Government and Governance*; Kő, A., Leitner, C., Leitold, H., Prosser, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 152–166.
14. O'Hara, K. Data Quality, Government Data and the Open Data Infosphere. Available online: <http://eprints.soton.ac.uk/340045/> (accessed on 3 July 2012).
15. Sadiq, S.; Indulska, M. Open data: Quality over quantity. *Int. J. Inf. Manag.* **2017**, *37*, 150–154. [CrossRef]
16. Faichney, J.; Stantic, B. A Novel Framework to Describe Technical Accessibility of Open Data. ALLDATA 2015: The First International Conference on Big Data, Small Data, Linked Data and Open Data. (IARIA) XPS for Publishing. 2015. Available online: <http://www.iaria.org/conferences2015/ComALLDATA15.html> (accessed on 27 October 2020).
17. Kapoor, K.; Weerakkody, V.; Sivarajah, U. Open Data Platforms and Their Usability: Proposing a Framework for Evaluating Citizen Intentions. In *Open and Big Data Management and Innovation*; Janssen, M., Mantymaki, M., Hidders, J., Klievink, B., Lamersdorf, W., VanLoenen, B., Zuiderwijk, A., Eds.; Springer: Cham, Switzerland, 2015; Chapter 6330; pp. 261–271.
18. Böhm, C.; Freitag, M.; Heise, A.; Lehmann, C.; Mascher, A.; Naumann, F.; Ercegovac, V.; Hernandez, M.; Haase, P.; Schmidt, M. GovWILD: Integrating Open Government Data for transparency. In *Proceedings of the WWW'12—21st Annual Conference on World Wide Web Companion*, Lyon, France, 16–20 April 2012; pp. 321–324.
19. Peled, A. When Transparency and Collaboration Collide: The USA Open Data Program. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 2085–2094. [CrossRef]
20. Lourenço, R.P. An analysis of open government portals: A perspective of transparency for accountability. *Gov. Inf. Q.* **2015**, *32*, 323–332. [CrossRef]

21. Gurin, J. Open Governments, Open Data: A New Lever for Transparency, Citizen Engagement, and Economic Growth. *SAIS Rev. Int. Aff.* **2014**, *34*, 71–82. [CrossRef]
22. Ahmadi Zeleti, F.; Ojo, A.; Curry, E. Exploring the economic value of open government data. *Gov. Inf. Q.* **2016**, *33*, 535–551. [CrossRef]
23. Alt, R.; Franczyk, B. Business Models in the Data Economy: A Case Study from the Business Partner Data Domain. In *Proceedings of the 11th International Conference on Wirtschaftsinformatik (WI2013)*; Alt, R., Franczyk, B., Eds.; Universität Leipzig: Leipzig, Germany, 2013; p. 15.
24. Hansen, H.S.; Hvingel, L.; Schröder, L. Open Government Data—A Key Element in the Digital Society. In *Technology-Enabled Innovation for Democracy, Government and Governance*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 167–180.
25. Susha, I.; Grönlund, A.; Janssen, M. Driving factors of service innovation using open government data: An exploratory study of entrepreneurs in two countries. *Inf. Polity* **2015**, *20*, 19–34. [CrossRef]
26. Jetzek, T.; Avital, M.; Bjorn-Andersen, N. Data-Driven Innovation through Open Government Data. 2014. Available online: https://www.researchgate.net/publication/260929913_Data-Driven_Innovation_through_Open_Government_Data (accessed on 27 October 2020).
27. Lin, Y. Open data and co-production of public value of BBC Backstage. *Int. J. Digit. Telev.* **2015**, *6*, 145–162. [CrossRef]
28. Juell-Skielse, G.; Hjalmarsson, A.; Johannesson, P.; Rudmark, D. Is the Public Motivated to Engage in Open Data Innovation? *Lect. Notes Comput. Sci.* **2014**, 277–288. [CrossRef]
29. Chan, C.M.L. *From Open Data to Open Innovation Strategies: Creating E-Services Using Open Government Data*; IEEE: New York, NY, USA, 2013; pp. 1890–1899.
30. Eckartz, S.; van Broek, T.D.; Ooms, M. Open Data Innovation Capabilities: Towards a Framework of How to Innovate with Open Data. In *Electronic Government*; Scholl, H.J., Glassey, O., Janssen, M., Klievink, B., Lindgren, I., Parycek, P., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 47–60.
31. Worthy, B. The Impact of Open Data in the UK: Complex, Unpredictable, and Political. *Public Adm.* **2015**, *93*, 788–805. [CrossRef]
32. Maier-Rabler, U.; Huber, S. “Open”: The changing relation between citizens, public administration, and political authority. *JeDEM eJ. eDemocracy Open Gov.* **2011**, *3*, 182–191. [CrossRef]
33. Bates, J. The strategic importance of information policy for the contemporary neoliberal state: The case of Open Government Data in the United Kingdom. *Gov. Inf. Q.* **2014**, *31*, 388–395. [CrossRef]
34. Ruijter, E.; Détienné, F.; Baker, M.; Groff, J.; Meijer, A.J. The Politics of Open Government Data: Understanding Organizational Responses to Pressure for More Transparency. *Am. Rev. Public Adm.* **2020**, *50*, 260–274. [CrossRef]
35. Ubaldi, B. *Open Government Data Towards Empirical Analysis of Open Government Data Initiatives*; OECD: Paris, France, 2013. [CrossRef]
36. Jetzek, T.; Avital, M.; Bjørn-Andersen, N. Generating Value from Open Government Data. ICIS 2013 Proceedings. 2013. Available online: <https://aisel.aisnet.org/icis2013/proceedings/GeneralISTopics/5/> (accessed on 27 October 2020).
37. Jetzek, T.; Avital, M.; Bjørn-Andersen, N. Generating value from open government data. In *Proceedings of the 34th International Conference on Information Systems ICIS 2013, Milano, Italy, 15–18 December 2013*; pp. 1737–1756.
38. Attard, J.; Orlandi, F.; Auer, S. Value Creation on Open Government Data. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*; IEEE: New York, NY, USA, 2016. [CrossRef]
39. Oakland, W.H. *Chapter 9 Theory of Public Goods*; Elsevier: New York, NY, USA, 1987; pp. 485–535.
40. Cornes, R.; Sandler, T. *The Theory of Externalities, Public Goods, and Club Goods*; Cambridge University Press: Cambridge, UK, 1996. [CrossRef]
41. Hettich, W.; Winer, S.L. *Democratic Choice and Taxation*; Cambridge University Press: Cambridge, UK, 1999. [CrossRef]
42. Mueller, D.C. *Public Choice III*; Cambridge University Press: Cambridge, UK, 2003. [CrossRef]
43. Hankla, C.; Martinez-Vazquez, J.; Rodríguez, R.P. *Local Accountability and National Coordination in Fiscal Federalism*; Edward Elgar Publishing: Cheltenham, UK, 2019. [CrossRef]
44. Roemer, J.E. *Political Competition: Theory and Applications*; Harvard University Press: Cambridge, UK, 2011.

45. Bates, J. The Domestication of Open Government Data Advocacy in the United Kingdom: A Neo-Gramscian Analysis. *Policy Internet* **2013**, 118–137. [CrossRef]
46. Dos Santos Brito, K.; Da Silva Costa, M.A.; Garcia, V.C.; De Lemos Meira, S.R. Assessing the benefits of open government data: The Case of meu congresso nacional in Brazilian Elections 2014. In *ACM International Conference Proceeding Series*; Zhang, J.K.Y., Ed.; Association for Computing Machinery: New York, NY, USA, 2015; pp. 89–96.
47. Purwanto, A.; Zuiderwijk, A.; Janssen, M. Citizen engagement with open government data. In *Transforming Government: People, Process and Policy*; Emerald Publishing Limited: Bingley, UK, 2020; pp. 1–30. [CrossRef]
48. Hong, S. Electoral Competition, Transparency, and Open Government Data. In *Proceedings of the 21st Annual International Conference on Digital Government Research*, Seoul, Korea, 15–19 June 2020.
49. Kochi, I.; Rodríguez, R.A.P. Voting in federal elections for local public goods in a fiscally centralized economy. *Estud. Econ.* **2011**, 26, 123–149.
50. Bergstrom, T.C.; Goodman, R.P. Private Demands for Public Goods. *Am. Econ. Rev.* **1973**, 63, 280–296.
51. Beron, K.J.; Murdoch, J.C.; Vijverberg, W.P.M. Why Cooperate? Public Goods, Economic Power, and the Montreal Protocol. *Rev. Econ. Stat.* **2003**, 286–297. [CrossRef]
52. Jackson, P.M.; Atkinson, A.B.; Stiglitz, J.E. Lectures on Public Economics. *Econ. J.* **1981**, 573. [CrossRef]
53. Tresch, R.W. *Public Finance: A Normative Theory*; Academic Press: Cambridge, MA, USA, 2002.
54. Mas-Colell, A.; Whinston, M.D.; Green, J.R. *Microeconomic Theory*; Oxford University Press: Oxford, UK, 1995.
55. Rubinfeld, D.L. The economics of the local public sector. In *Handbook of Public Economics*; Elsevier: New York, NY, USA, 1987; pp. 571–645.
56. Mason, H.; Wiggins, C. A Taxonomy of Data Science. Available online: <http://www.dataists.com/2010/09/a-taxonomy-of-data-science/> (accessed on 27 October 2020).
57. Kotsiantis, S.B.; Kanellopoulos, D.; Pintelas, P.E. Data preprocessing for supervised learning. *Int. J. Comput. Sci.* **2006**, 1, 111–117.
58. Liu, H.; Motoda, H. *Feature Extraction, Construction and Selection: A Data Mining Perspective*; Kluwer Academic Publishers: Norwell, MA, USA, 1998.
59. Garg, A.; Tai, K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *Int. J. Model. Identif. Control* **2013**, 18, 295–312. [CrossRef]
60. George, G.; Osinga, E.C.; Lavie, D.; Scott, B.A. Big Data and Data Science Methods for Management Research. *Acad. Manag. J.* **2016**, 59, 1493–1507. [CrossRef]
61. John Lu, Z.Q. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2010**, 173, 693–694. [CrossRef]
62. Cao, L. Data Science: Challenges and Directions. *Commun. ACM* **2017**, 60, 59–68. [CrossRef]
63. Salimans, T. Variable selection and functional form uncertainty in cross-country growth regressions. *J. Econom.* **2012**, 171, 267–280. [CrossRef]
64. Greene, W.H. *Econometric Analysis*; Pearson Education India: Bengaluru, India, 2003.
65. Long, J.S.; Ervin, L.H. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *Am. Stat.* **2000**, 54, 217–224.
66. Agunbiade, D.A.; Adeboye, N.O. Estimation of Heteroscedasticity Effects in a Classical Linear Regression Model of a Cross-Sectional Data. *Prog. Appl. Math.* **2012**, 4, 18–28.
67. Zhu, L.; Fujikoshi, Y.; Naito, K. Heteroscedasticity checks for regression models. *Sci. China Ser. A Math.* **2001**, 44, 1236–1252. [CrossRef]
68. Rao, C.R.; Toutenburg, H. *Linear Models*; Springer Series in Statistics: Cham, Switzerland, 1995; pp. 3–18. [CrossRef]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features

Antonio Maria Rinaldi ^{*,†} , Cristiano Russo [†]  and Cristian Tommasino [†]

Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Via Claudio, 21, 80125 Napoli, Italy; cristiano.russo@unina.it (C.R.); cristian.tommasino@unina.it (C.T.)

* Correspondence: antoniomaria.rinaldi@unina.it; Tel.: +39-081-768-3911

† These authors contributed equally to this work.

Received: 18 September 2020; Accepted: 26 October 2020; Published: 28 October 2020

Abstract: In recent years the information user needs have been changed due to the heterogeneity of web contents which increasingly involve in multimedia contents. Although modern search engines provide visual queries, it is not easy to find systems that allow searching from a particular domain of interest and that perform such search by combining text and visual queries. Different approaches have been proposed during years and in the semantic research field many authors proposed techniques based on ontologies. On the other hand, in the context of image retrieval systems techniques based on deep learning have obtained excellent results. In this paper we presented novel approaches for image semantic retrieval and a possible combination for multimedia document analysis. Several results have been presented to show the performance of our approach compared with literature baselines.

Keywords: content base image retrieval; semantic information retrieval; deep features; multimedia document retrieval

1. Introduction

The main aim of a search engine is to satisfy user information needs [1] retrieving relevant information for the user [2]. Relevance may be divided in two main classes called objective (system-based) and subjective (user-based) relevance respectively [3–5]. The objective relevance takes into account the direct match between the topic of the retrieved document and the desired topic, according to a user query. Several studies on human relevance [6–8] show that several criteria are involved in output evaluation of information retrieval process. The subjective relevance is related to the concepts of aboutness and appropriateness of retrieved information, so it depends on the user and his/her judgment. Relevance can also been divided into five typologies [9]: an algorithmic relevance between the query and the set of retrieved information objects; a topicality-like type, associated with aboutness; cognitive relevance, related to the user information need; situational relevance, depending on the task interpretation; and motivational and effective, which is goal-oriented. Moreover, relevance has two main features—multidimensional, because different users can grasp and evaluate differently the acquired information and dynamic because their skills and knowledge about a specific information can change over time. The information must be represented to be analyzed. We can use symbols to represent information and these symbols are called signs [10]. A sign can be defined as, “something that stands for something else, to someone in some capacity” [11]. A sign cannot be limited to words, but also images, sounds, videos and more. Starting from these considerations, the design of a modern information retrieval system takes into account that information can be represented in different forms, in order to improve the efficiency and effectiveness of the whole retrieval process. The existence of different kinds of information representation increases the semantic gap between low-level features and high-level concepts. Moreover, in a general content based approach

the query is unknown to the system if it is not represented with low-level features. The main effect of the semantic gap is that a query expressed in terms of low-level features can return wrong results if the conceptual content is not given. For this reason, the representation of low-level features is a crucial task in information retrieval systems.

In the last years, the extracting of low-level features is an important topic highly investigated in literature. New techniques based on deep-learning approach have been presented from the scientific community. In particular, novel feature called deep feature are extracted as an output of deep neural networks (DNNs).

In our work we propose novel techniques to analyze and combine semantic information and visual features for multimedia web documents retrieval. Our techniques are the base of a whole framework for multimedia query posing and document analyzing. Our approach has been extensively tested and compared with well-known semantic and content-based retrieval techniques.

The paper is organized as follows—in Section 2 we present a literature overview putting in evidence the differences with our approach; Section 3 gives a top-level view of our system, while in Section 4 we discuss the used strategy for our experiments; eventually, discussion and conclusions are reported in Section 5.

2. Related Work

Retrieval of multimedia objects had a contribution by several fields of scientific communities such as artificial intelligence, pattern recognition, computational vision and deep learning. In this section, we present the topics of our work, introducing and analyzing several approaches and techniques for text and image retrieval.

A user tends to handle high-level concept representations, such as keyword and textual descriptors, to interpret images and measure their similarity [12]. The difference in knowledge representation between low-level features and user semantic knowledge is referred as semantic gap [13,14]. The authors of References [15,16] proposed an interesting survey on low-level features and high-level semantics for image analysis and retrieval. Moreover, an interesting approach is performed at the query level [17]. Authors represent a query by three levels: (i) low-level features, such as text, colour and shape; (ii) derived features with the same degree of logical inference; (iii) abstract features. In the authors' opinion there are two ways to mitigate the semantic gap. The first consists in a visual and textual query posing performed by the user; on the other hand, the use of a multimedia Knowledge base to drive the multimedia analysis. In this paper we investigate the second approach and propose a framework based on semantic similarity to analyze the multimedia contents. In Reference [18], authors discuss semantic similarity measures and divide them into four types: (i) path-based; (ii) information content-based; (iii) features-based and (iv) hybrid methods. Path-based measures express the similarity between two concepts as a function of path length between concepts in a knowledge structure (e.g., taxonomies or semantic networks). The main idea of the information content-based measures is that each concept represent a well defined information in a knowledge structure and two or more concepts are similar if they share common information. The features-based measures are independent from knowledge structure and they derive similarity exploiting the properties of the knowledge structure itself. An hybrid measure combines the measure types described above and different relations between concepts such as is-a and part-of. In our work we use a path-based and information content-based semantic similarity measures.

The features used to represent images have a low dimension compared with the original raw data, they are generally composed by a series of numerical values and can be represented through different data formats (e.g., a vector). The authors of Reference [19] propose a comprehensive review on feature extraction for content-based image retrieval systems (CBIRs) and in Reference [20] is presented a review on SIFT and different features extracted from convolutional neural networks (CNN) for image retrieval. In general, we can divide features in global, local and deep features. The global features aim to represent an image as a whole considering, for example, colour, shape and texture.

Local features describe the key-points of an image object, generally used in object detection and recognition. Deep features are a new approach based on DNNs. In some case, deep features are the output of the last level of the DNN or other levels with additional operators [20]. Many authors consider the second to last level of the DNN as a deep feature applying an aggregation layer based on *Global Max Pooling* or *Global Average Pooling*.

There are several descriptors presented in literature and the remainder of this section. We introduce different baselines used in our work, as discussed in the following sections. *PHOG* (*Pyramid of Histograms of Orientation Gradients*) is described in detail in [21], where the authors propose a descriptor based on HOG (*Histograms of Orientation Gradients*). Its goal is the representation of an image through its local shape and their spatial arrangement. A local feature is an image pattern which differs from its immediate neighbourhood. It is usually related to a change of an image property or several properties simultaneously [22]. Example of local features are SIFT [23] and SURF [24]. *ORB* [25] was proposed as a valid alternative to mitigate the high computational cost of SIFT and SURF; it is a fusion of FAST keypoint detection and BRIEF descriptor. In this paper, we use PHOG as a global descriptor due to its good results reported in the discussed literature. In addition, we use ORB as a local feature because it has a comparable accuracy with SIFT and SURF but it presents a fast computation [26]. In the last years, the progress of Computer Vision-based on Deep Learning achieved impressive performances in terms of accuracy and image understanding [27]. In this context, different DNNs have been proposed, and specific architectures called Convolution Neural Network presents the best results. A Convolution Neural Network (CNN/ConvNet) is a type of artificial neural feed-forward network inspired by the organization of the animal visual cortex. A CNN is organized in different layers: (i) input layer; (ii) convolution; (iii) normalisation; (iv) pooling; (v) full connection [27].

Over the years, the scientific community proposed different CNN architectures. In this section we briefly introduce the ones used in our work. *VGGNet* [28], developed by *VGG* (*Visual Geometry Group*) of the University of Oxford and presented at ImageNet LSVRC (Large Scale Visual Recognition Competition) in 2014 in the classification task, it is one of the CNNs more used in research works becoming one of the most cited in the literature. *ResNet* [29] won the first place in the *ILSVRC 2015* classification competition with a top 5 error rate of 3.57%. *ResNet* is a residual network. Residual networks solve some learning problems when many levels are added to the convolutional network in some points because in this condition, the previous CNNs architecture the performance degrades quickly. In 2014, Google researchers introduced the Inception network [30], which ranked first in the *ILSVRC 2014* competition. The issues addressed by the authors to design Inception are mainly related to the difficulty of choosing the kernel size, the creation of too deep networks which generate overfitting and reduce the computational cost. The authors present different versions of Inception, and in each version, they apply a set of optimizations to improve accuracy and decrease the computational complexity. *MobileNet* [31,32] is a neural network proposed by Google researchers. This neural network can be used on mobile devices or in cases where processing power is limited.

In our system, we use the CNNs reported above for feature extraction. In References [33,34] the authors propose a literature review of deep learning in Content-based Information Retrieval Systems (CIBR) and, according to them we use deep descriptors extracted from the second last layer of a CNN and apply max or average pooling.

Over the years, in literature can find many frameworks for CBIR and Information retrieval systems based on ontology and multimedia features. The authors of Reference [35] proposed an approach to support multi-modal image retrieval based on the Bayes point machine to associate words and images. In Reference [36], authors use the latent semantic indexing together with both textual and visual features to extract the underlying semantic structure of a web page. The authors of Reference [37] propose an iterative similarity propagation approach to explore the inter-relationships between web images and their textual annotations for image retrieval. In Reference [38], it is introduced a semantic combination technique to efficiently fuse text and image retrieval systems in multimedia information retrieval. In Reference [39], the authors report a description of an ontology model, the integrates

domain specific features, and processing algorithms focused on the domain specified by the user. YaSemIR [40] is a free and open-source semantic IR semantic system based on Lucene, which uses conceptual labels to annotate documents and questions. In Reference [41], authors discuss and present a study about different multimedia retrieval techniques based on ontologies in the semantic web. They compare these techniques to highlight the advantages of text, image, video and audio-based retrieval systems. The authors of Reference [42] propose a recommendation system for e-business applications. The recommendation strategy aimed is based on a hybrid approach, combining intrinsic characteristics of objects, past behaviours of users in terms of usage patterns and user interest expressed by ontologies to computes customized recommendations. In Reference [37], the authors present two methods to improve the Image recovery system performance. The first proposed method defined the most efficient way to GLCM texture for the recovery process. It also increased the recovery precision, combining the most efficient GLCM structure with DWT decomposition. The second proposal combined colour and texture characteristics to improve the method recovery services. This method combined the HSV colour with the most efficient GLCM texture features and with the GLCM and DWT texture features.

The authors of Reference [43] propose an efficient “bag-of-words” model that uses deep local descriptors of the convolutional neural network. The selection of high-quality descriptors provides a simple and effective way to choose the most discriminating local descriptors, which significantly improves the accuracy of retrieval. They evaluate Different methods of pre-processing the descriptors, and they report that the RootCFM is to be the best. The model uses a large visual codebook combined with the inverted index for efficient storage and fast recovery. The authors of Reference [44] show a Semantic Event Retrieval System, that includes high-level concepts and uses concept selection based on semantic embeddings. In Reference [45], the authors propose a new multimedia embedding for few-example event recognition and translation.

In this article, we propose a novel framework for multimedia web document retrieval system combining semantic similarity measures based on a formal and semantic multimedia knowledge base and different image descriptors. In particular, deep descriptors have been computed using more performance aggregation functions, reducing their dimension and improving the accuracy of the result considerably.

3. The Proposed System

In this section, we present the architecture of the implemented system and describe in detail the modules. The Figure 1 shows the system at a glance.

The system consists of two main subsystems. The first one has in charge the creation and population of the database from a document collection; moreover, it extracts images and their descriptors and normalizing the text. The second one implements the retrieval process, using semantic measure and image retrieval techniques. *Multimedia Web Documents Repository Processor* creates the document collection used in the retrieval tasks. The Figure 2 shows an activity diagram of this subsystem. Web documents are processed to obtain a structured object, then is stored in a NoSQL document-based database. In particular, the subsystem extracts image and text from the document normalizing the text, while the images are processed to extract features. The result of the this tasks is then merged into a single structured object if the document has at least one image and the text is composed of at least 25 terms. We remove documents without images because we want evaluate a combination of textual and visual descriptors.

The *Text normalization* is a fundamental step in an information retrieval systems because it allows to obtain documents with only clean text without stopwords and other kind of “noise” as for example misspelling terms [1]. Moreover, we use stemming and lemmatization tasks to have each word in its basic form. As shown in the Figure 3 the *Tags Removal* module removes the HTML code from the text. The second module transforms the text into lowercase text and removes special chars (i.e., all chars except number, letter, _ and '). *Stopwords Removal* processes the normalized

text removing words without any particular meaning (e.g., articles, adverbs, conjunctions, ...). *Stemming-Lemmatization* module analyzes the output of the previous blocks and performs text transformations. Stemming transforms words into their canonical form to achieve more effective matching. The problem in using stemming is that words with different meanings can be associated with the same root. To solve this issue, we use a more sophisticated algorithm which uses lemmas instead of the root. The use of lemmas is more efficient because it represents the canonical form of the word (e.g., for verbs it uses infinitive form). The last step before document storing is performed by the *Out of vocabulary words removal* module. It removes all words that do not have sense in English language (e.g., misspelling words).

The semantic similarity module is in charge of compute the similarity on the textual document information. In our work we use different kinds of similarity metrics as described in the previous section.

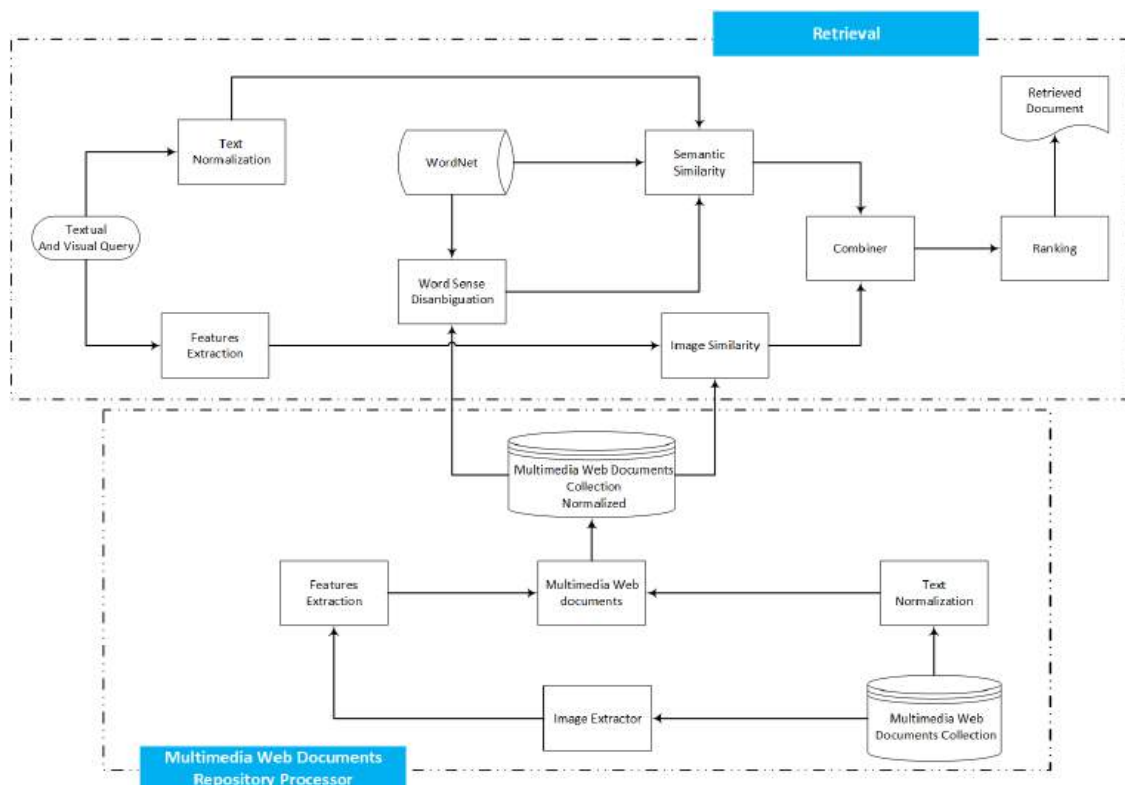


Figure 1. Architecture: Proposed Framework.

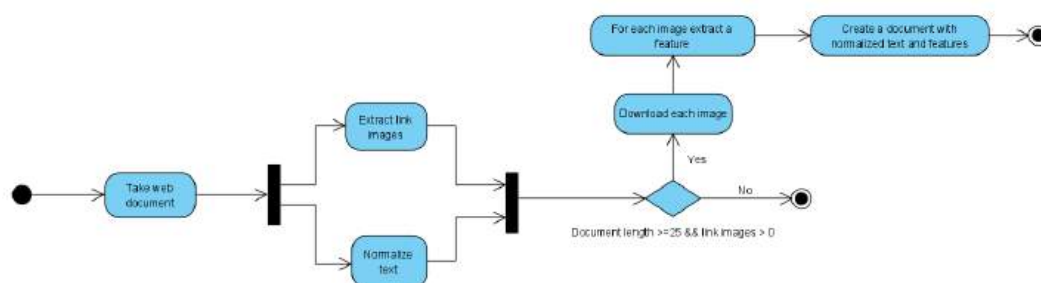


Figure 2. Activity diagram: Multimedia web Documents Repository Processor.

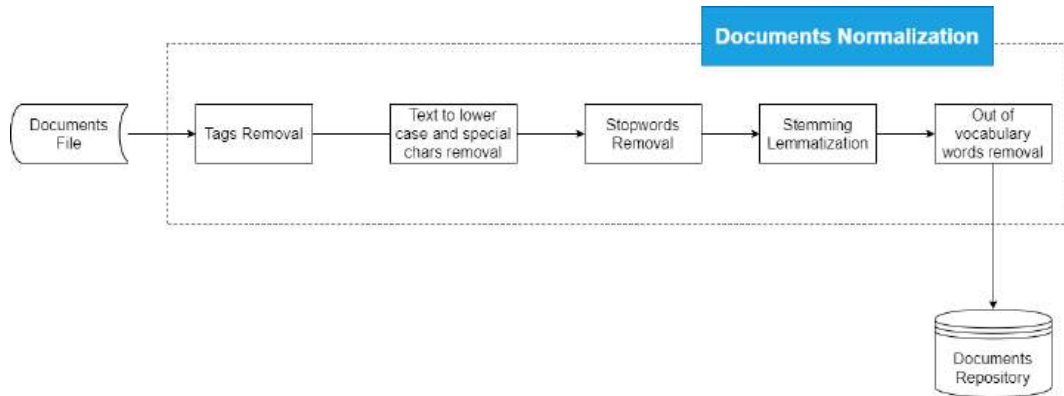


Figure 3. Text Normalization.

The semantic measures based on path are:

- Shortest Path based Measure: this measure only takes into account $len(c_1, c_2)$. It assumes that the $sim(c_1, c_2)$ depend on closeness of two concepts are in the taxonomy, and that a similarity between two terms is proportional to the number of edges between them.

$$sim_{path}(c_1, c_2) = 2 \cdot deep_{max} - len(c_1, c_2). \quad (1)$$

- Wu & Palmer's Measure: it introduced a scaled measure. This similarity measure considers the position of concepts c_1 and c_2 in the taxonomy relatively to the position of the most specific common concept $lso(c_1, c_2)$. It assumes that the similarity between two concepts is the function of path length and depth in path-based measures.

$$sim_{WP}(c_1, c_2) = \frac{2 \cdot depth(lso(c_1, c_2))}{len(c_1, c_2) + 2 \cdot depth(lso(c_1, c_2))}. \quad (2)$$

- Leacock & Chodorow's Measure: it uses the maximum depth of taxonomy from the considered terms.

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2 \cdot deep_{max}}. \quad (3)$$

Semantic measures based on information content (IC) assume that each concept includes much information in a knowledge based (i.e., WordNet [46]). Similarity measures are based on the information content of each concept. The measures are:

- Resnik's Measure: it assumes that for two given concepts, similarity is depended on the information content that subsumes them in the taxonomy

$$sim_{resnik}(c_1, c_2) = -\log p(lso(c_1, c_2)) = IC(lso(c_1, c_2)). \quad (4)$$

- Jiang's Measure: Jiang's Measure uses both the amount of information needed to state the shared information between the two concepts and the information needed to fully describe these terms. The value is a semantic distance between two concepts. Semantic similarity is the opposite of the semantic distance.

$$sim_{jiang}(c_1, c_2) = \frac{2 \cdot IC(lso(c_1, c_2))}{IC(c_1) + IC(c_2)}. \quad (5)$$

Image Extractor module fetches the images in each HTML document. The retrieved images are analyzed and filtered by size and format; in our case, we consider only JPEG images but we can easily consider other format. The same module extracts visual features from these images. In our framework, we consider three types of features—local, global and deep. We use ORB as local feature

due to its performance compared with other similar descriptors [26]. The same consideration is made regarding the global descriptor PHOG [47]. We consider four deep descriptors derived from CNNs—VGG16, ResNet50, InceptionV3, MobileNetV2 due to the novelty in the use of this kind of features. For each CNNs we have implemented a global average pooling and a global max pooling. In this work the CNNs are pre-trained on ImageNet. Multimedia Web Document Composer module builds the documents to store in our NoSQL database (i.e., MongoDB). In particular, we use a json document consists of the following fields:

- Name: web site name;
- Url: web site url;
- Images: array like structure, each element is a json like structure (see Figure 4);
 - Image name;
 - Image path: image storage path;
 - PHOG: PHOG feature;
 - ORB: 2-D array of ORB key point;
 - Deep Descriptor: array of deep descriptors;

```
_id: ObjectId("5da2e85a9a96f1812d0b1847")
name: "113600"
url: "http://www.cottagecraft.co.uk/"
topic: "Top/Sports/Equestrian/Equipment_and_Apparel"
cleaned_text: "javascript seems disabled browser for best experience site sure turn j..."
title: "Cottage Craft"
summary: "Manufacturers of girths, headcollars, numnahs, and pads. Site offers p..."
images: Array
  0: Object
    path: "/113600_1.jpg"
    size: 617880
    orb: Array
    deep_desc: Array
  1: Object
    path: "/113600_0.jpg"
    size: 617880
    orb: Array
    deep_desc: Array
  2: Object
    path: "/113600_7.jpg"
    size: 242248
    orb: Array
    deep_desc: Array
```

Figure 4. Document Example.

The Retrieval subsystem has been configured in three different search cases. The first case (*Case A*) is the text only query. In Figure 5 the activity diagram is shown. The system normalizes the text of the query posed by the user and considers the domain to extract the concept from the knowledge Base. It computes the semantic similarity between the concept and each term of the documents. In a previous step, we apply a word sense disambiguation algorithm, and for the assignment of the score to the single documents the Equation (6) is used.

$$sim_{doc}(tq, d) = \frac{\sum_{i=1}^N sim(td, d_i)}{N}, \quad (6)$$

where tq is a concept, d is a document, N is the number of tokens in the document and d_i is a document term.

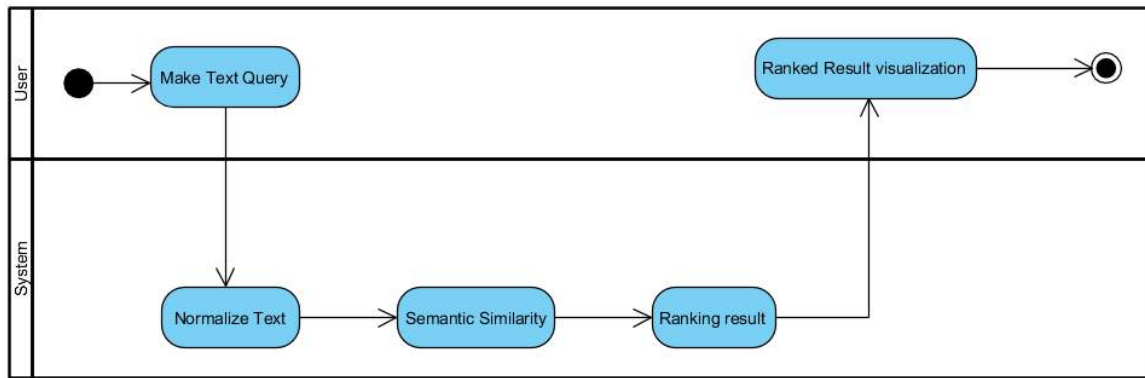


Figure 5. Activity diagram: Textual query (Case A).

The second case (*Case B*) is the visual query. In Figure 6 the activity diagram of this step is shown. The system extracts the feature from the image query and computes its similarity with regards to the document collection images. Semantic similarity score between query image and multimedia document is obtained as average of cosine similarity between query and each image contained in multimedia documents, as expressed in the Equation (7), and ranks the documents based on the results obtained.

$$sim(vq, d) = \frac{\sum_{i=1}^N \cos(vq, d_i)}{N}, \quad (7)$$

where vq is a descriptor of the query image, d represents all visual descriptors extracted from an image in the document, N is the number of images in the document and d_i is a visual descriptor of a single image.

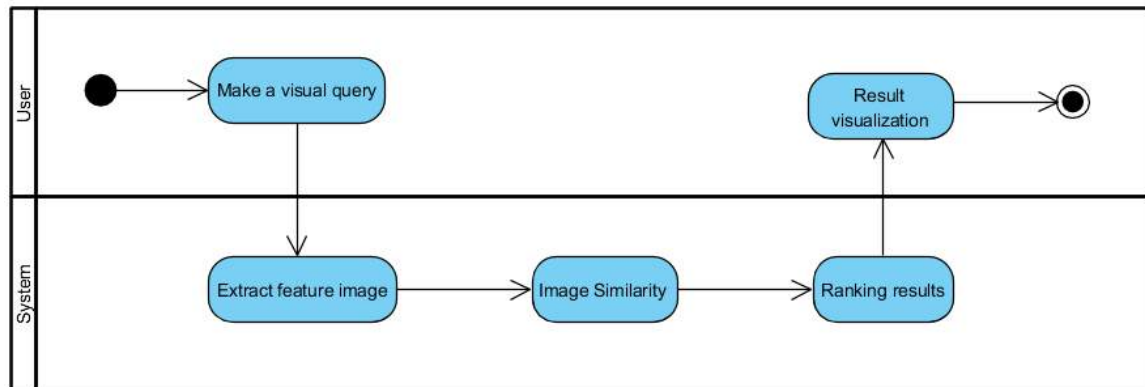


Figure 6. Activity diagram: visual query (Case B).

The third case (*Case C*) is the combined visual and text query and the Figure 7 shows the activity diagram. In this case, the system performs an union of the two cases illustrated above. Then it adds them up to combine the results.

In this paper the *knowledge base* is implemented following a multimedia model proposed in References [48,49]. This formal representation uses signs as defined in Reference [11]. A concept can be represented in various multimedia forms. The structure of the model is composed of a triple $\langle S, P, C \rangle$, defined as:

- S : the set of signs;
- P : the set of properties used to relate signs to concepts;
- C : the set of constraints on the set P .

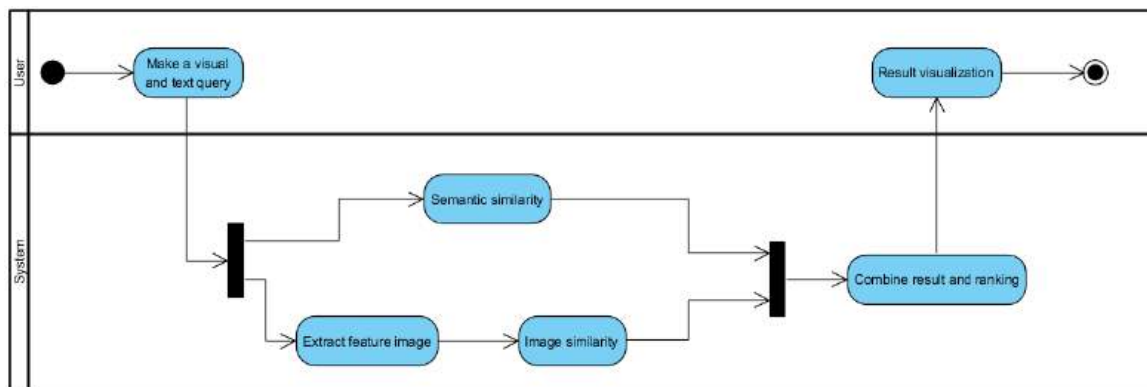


Figure 7. Activity diagram: visual and textual query (Case C).

The knowledge base is an ontology logically represented by a semantic network (SN). We can see a SN as a graph where the nodes are the concepts and the arcs the semantic relations between them. The language chosen to describe this model is the DL version of the Web Ontology Language (OWL) [50], a standard language in the semantic web and we use a NoSql technology to implement a SN, in particular Neo4j graph DB. The hierarchies used to represent the objects of interest in our model are shown in Figure 8.

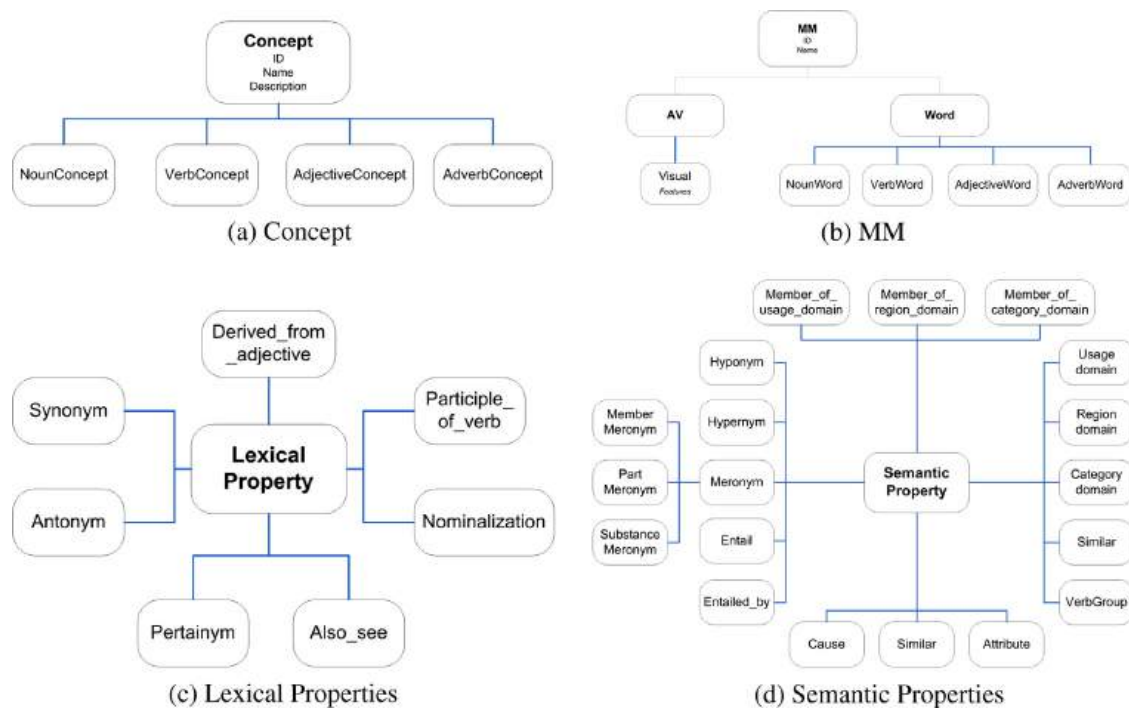


Figure 8. Concept, Multimedia, Semantic Properties.

The knowledge base has been populated using ImageNet [51]. Imagenet is based on WordNet adding multimedia contents to a portion of it. The multimedia nodes added to the graph contain only the meta-data and the raw image data is stored in the document-based data structure previously described. The split of image metadata from its raw content is performed to improve the global performance of our *knowledge base* due to the problems related to graph db to store a manage large documents. Therefore we design and implement an hybrid technology solution based on document based and graph db. We explicit point out that this is a novelty of our work. Figure 9 shows a sketch of our knowledge base.

The *Word Sense Disambiguation (WSD)* is a fundamental process in semantic similarity computation because it associates each lemma of the document with the correct concept (i.e., the right sense). Our implementation uses the Lesk algorithm [52]. The assumption at the base of this algorithm is that words have neighborhoods which tend to share a topic. A simplified version of the algorithm has been proposed in Reference [53].

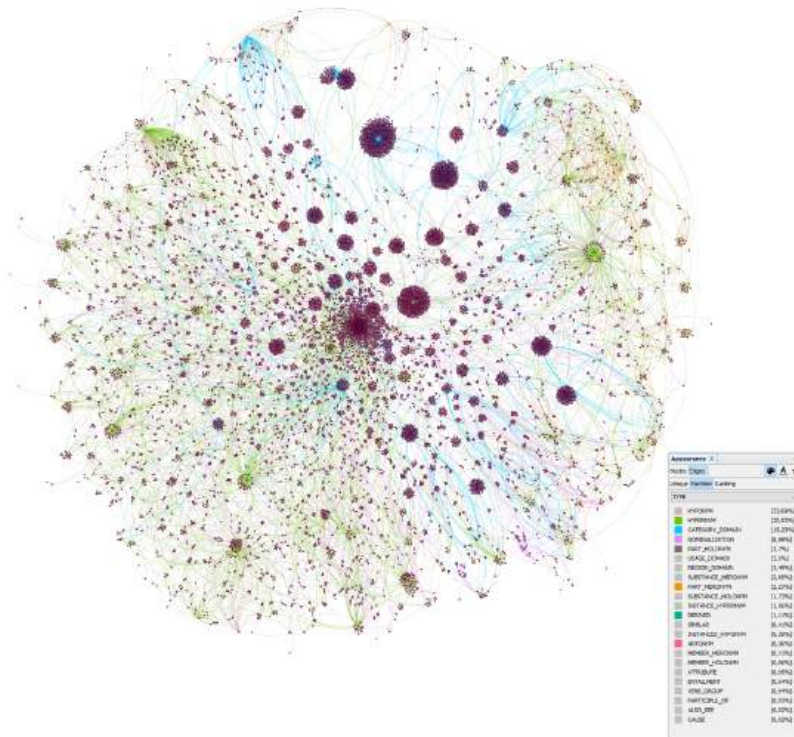


Figure 9. Knowledge Base.

The *Image Similarity Module* works with two models: in case of features extracted with *PHOG* or with a *CNN*, we use the cosine distance, while in case of *ORB*, we use the best matcher. This difference is due to the dimensionality of the descriptors. We apply *global max pooling* or *global average pooling* to obtain a feature expressed as a mono-dimensional array [54] to use the cosine similarity with features extracted from the *CNNs*. In the case of *ORB*, we use the best match as suggested in Reference [55].

The main aim is to improve the performance of the system ensuring the best accuracy of results and the best precision. Several studies have shown that the use of different combined techniques could achieve better results because one technique could compensate the lacks of the others. In Figure 10 is shown a combining process. In this work, we use a the sum as combining function [56].

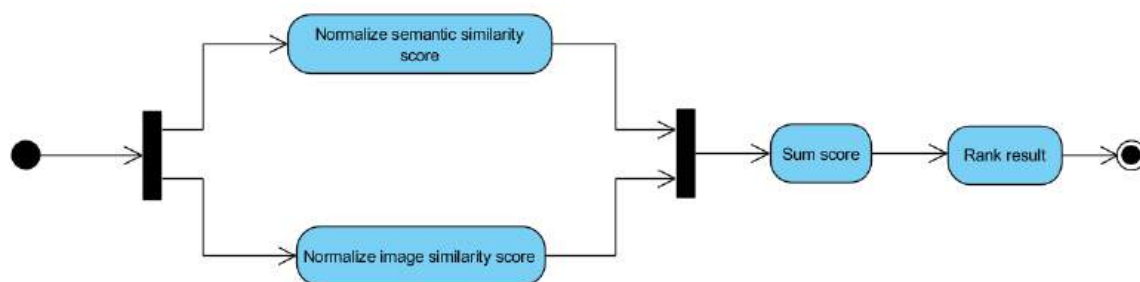


Figure 10. Activity diagram: sum combiner.

4. Experimental Results

In this section, we present the used document collections, the testing strategy and the obtained results.

We used three datasets—*20 Newsgroups* [57], *PASCAL VOC2012* [58] and *DMOZ* [59]. The specific characteristic in terms of document features, contents and size allow us to evaluate in deep all the component of our framework. In particular, we use the first dataset to evaluate and select retrieval metrics based on semantic similarity, the second one to evaluate and select the deep features in *CBIR* context and the last one to perform the final tests in which we combine the different strategies.

20 Newsgroups is one of the most used document collections in literature. It is available in *scikit-learn* and managed by Python language. This version allows us to get the dataset by removing headers, footers and quotes. The 20 newsgroups collection contains twenty topics, as shown in Table 1.

Table 1. Number of documents for each category in 20 Newsgroups dataset.

Topic	Documents
alt.atheism	799
comp.graphics	973
comp.os.ms-windows.misc	985
comp.sys.ibm.pc.hardware	982
comp.sys.mac.hardware	961
comp.windows.x	980
misc.forsale	972
rec.autos	990
rec.motorcycles	994
rec.sport.baseball	994
rec.sport.hockey	999
sci.crypt	991
sci.electronics	981
sci.med	990
sci.space	987
soc.religion.christian	997
talk.politics.guns	910
talk.politics.mideast	940
talk.politics.misc	775
talk.religion.misc	628
Total	18,828

The document collection has been pre-processed by the normalization module. The query set used for the evaluation of semantic similarity metrics is composed of ten queries, for each one has been recognized the set of relevant categories of 20Newsgroups. For example for the query “mars” as planet the corresponding category is “sci.space”.

The *PASCAL VOC2017* is an image database composes by 20 objects as shown in Table 2.

The query set used for the evaluation of the visual descriptors is composed of thirteen classes each consisting of five images. The classes have a direct correspondence with the classes pre-assigned in the dataset.

DMOZ has been one of the most popular and rich multilingual open-source web directories. The project initially called ODP—Open Directory Project was born in 1998. The purpose of *DMOZ* was to collect and index URLs to create a directory of hierarchically organized web contents. From the *DMOZ* dump both image and text has been extracted as described in Section 3. In Tables 3 and 4 are reported the statistics downstream of this process.

We use as evaluation metrics Precision-Recall curve and Mean Average Precision (MAP) [1].

The Precision is expressed as:

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (8)$$

Table 2. Number of images for each category VOC2017 dataset.

Object	Images
Aeroplane	1340
Bicycle	1104
Bird	1530
Boat	1016
Bottle	1412
Bus	842
Car	2322
Cat	2160
Chair	2238
Cow	606
Diningtable	1176
Dog	2572
Horse	964
Motorbike	1052
Person	8174
Pottedplant	1054
Sheep	650
Sofa	1014
Train	1088
Tvmonitor	1150
Total	23,080

Table 3. Number of multimedia web document for each category in DMOZ collection.

Top Category	Num.
Sports	1087
Society	824
Computers	758
Shopping	1537
Arts	785
Business	1895
Health	1025
Games	385
News	459
Science	509
Total	9264

Table 4. Number of multimedia web document for each level of DMOZ collection.

Level	Num.
I	2
II	54
III	1023
IV	2353
V	2485
VI	1699
VII	918
VIII	641
XI	81
X	8
Total	9264

It is the percentage of retrieved relevant documents compared with all retrieved documents. Recall is expressed as:

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}. \quad (9)$$

It is the percentage of relevant retrieved documents compared with all relevant documents in the document collection.

The Precision-Recall curve is obtained as an interpolation of the Precision values for 11 standard Recall values ranging from 0 to 1 with step 0.1. Interpolation is estimated with the following criteria:

$$P_{interp}(r) = \max_{r_i \geq r} p(r_i). \quad (10)$$

We use The Precision-Recall curve to compare different retrieval algorithms considering the whole set of retrieved documents. Moreover, one of the most used measures in literature to measure the web information retrieval performances is the Mean Average Precision on the first top k results. We choice MAP, because in a web search engine the user is generally interested in the first k results. Mean Average Precision is expressed as:

$$AveP = \frac{\sum_{k=1}^n (P(k) \cdot rel(k))}{number\ of\ relevant\ documents} \quad (11a)$$

$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}. \quad (11b)$$

In this work, we consider the MAP@10 because it is mostly used on web document retrieval considering the first 10 results.

As previously stated, the measure of semantic similarity has been performed considering: *path similarity*, *Leacock-Chodorow Similarity*, *Wu-Palmer Similarity*, *Resnik Similarity* and *Jiang-Conrath Similarity*. We use the *Lesk* algorithm as disambiguation technique. The query set used consists of twelve text queries, which have polysemic meaning. An example of a polysemic query is “mars” which in WordNet has two meanings, “Mars” as god and “Mars” as a planet.

In this context we are interested in text analysis and the used document collection is 20 newsgroups. In the Figure 11 there is shown the Precision-Recall curve, where we can observe that *Jiang-Conrath Similarity* is the best measure, but in Figure 12 the *path similarity* where the best measure with respect to MAP@10.

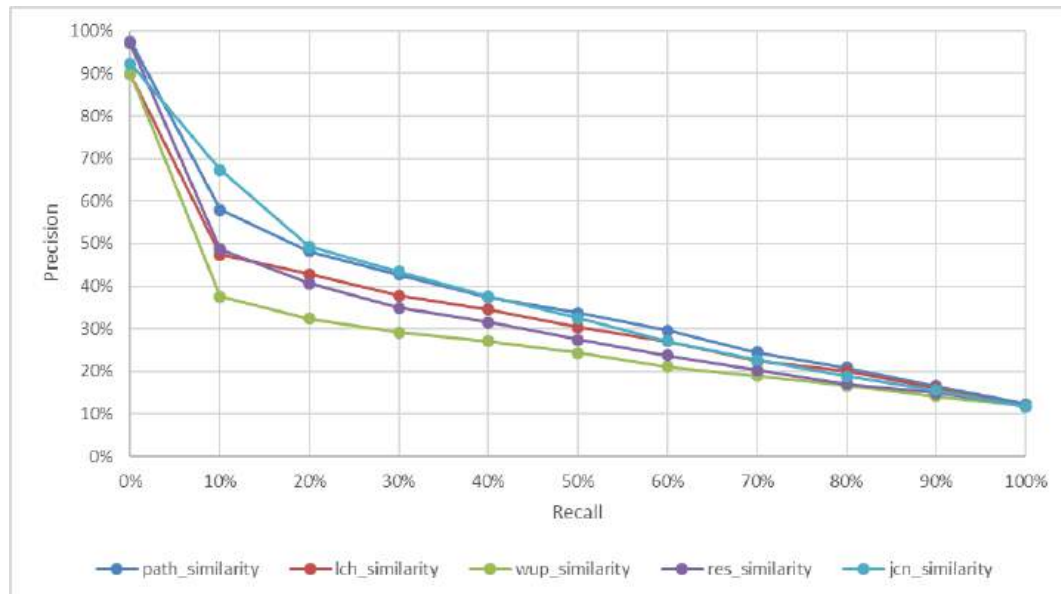


Figure 11. Precision-Recall Curve for Textual Semantic Similarity Measures.

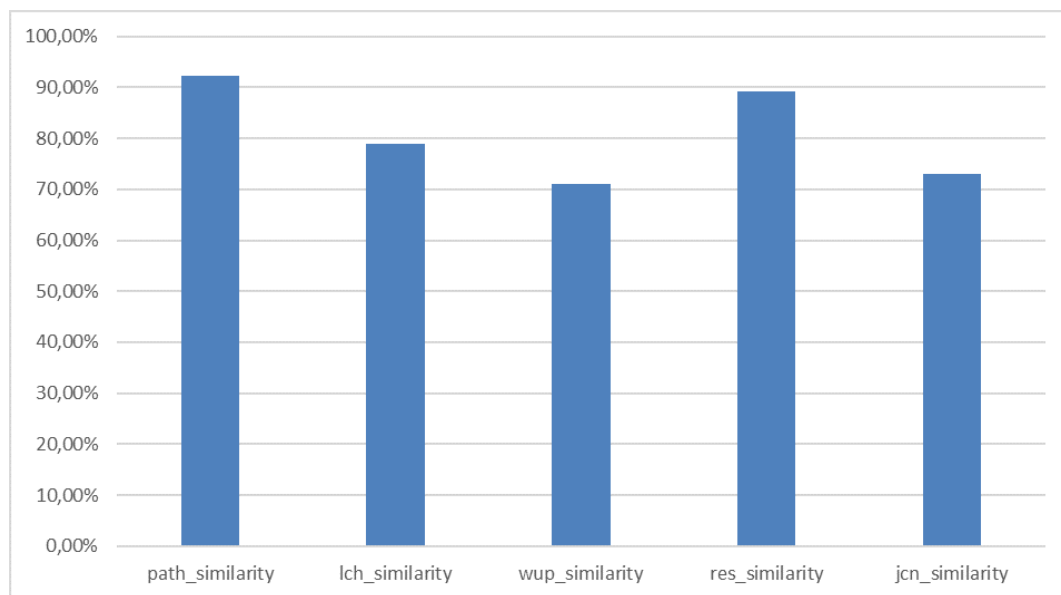


Figure 12. MAP@10 for Textual Semantic Similarity Measures.

According to results shown in Figures 11 and 12 we choose path similarity and Resnik similarity, the first one is path-based type. We consider two measures on semantic textual similarity due to the very small difference obtained by experiments. In this way we can investigate more deeply in the use of a combination of these measures and multimedia descriptors.

The used visual descriptors are: *PHOG*, *ORB* and deep descriptors. In particular, the deep descriptors are:

- VGG-16 with global max pooling and global average pooling;
- ResNet-50 with global max pooling and global average pooling;
- Inception V3 with global max pooling and global average pooling;
- MobileNet V2 with global max pooling and global average pooling.

The used query set consists in sixty-five images, divided into ten classes. The used dataset is PASCAL VOC2012.

Figure 13 shows the Precision-Recall curve, where we can see that the descriptor extracted from *ResNet-50 with global average pooling* has the best result. Moreover, in the Figure 14 the *MobileNet V2 with global average pooling* is the best considering MAP@10.

The chosen descriptor, according to Figures 13 and 14, is the one extracted with MobileNet V2 with global average pooling. In this case we consider only one descriptor due to the high difference in accuracy obtained by the results comparison.

The evaluation strategy has been defined by different test cases performed on DMOZ data set to have a complete analysis of our framework in a real scenario. The test cases are:

- *Case A*: text query:
 - Path similarity;
 - Resnik similarity;
- *Case B*: visual query, using deep descriptor obtained with MobileNet V2 with global average pooling;
- *Case C*: visual and text query:
 - Path similarity and deep descriptor;
 - Resnik similarity and deep descriptor;

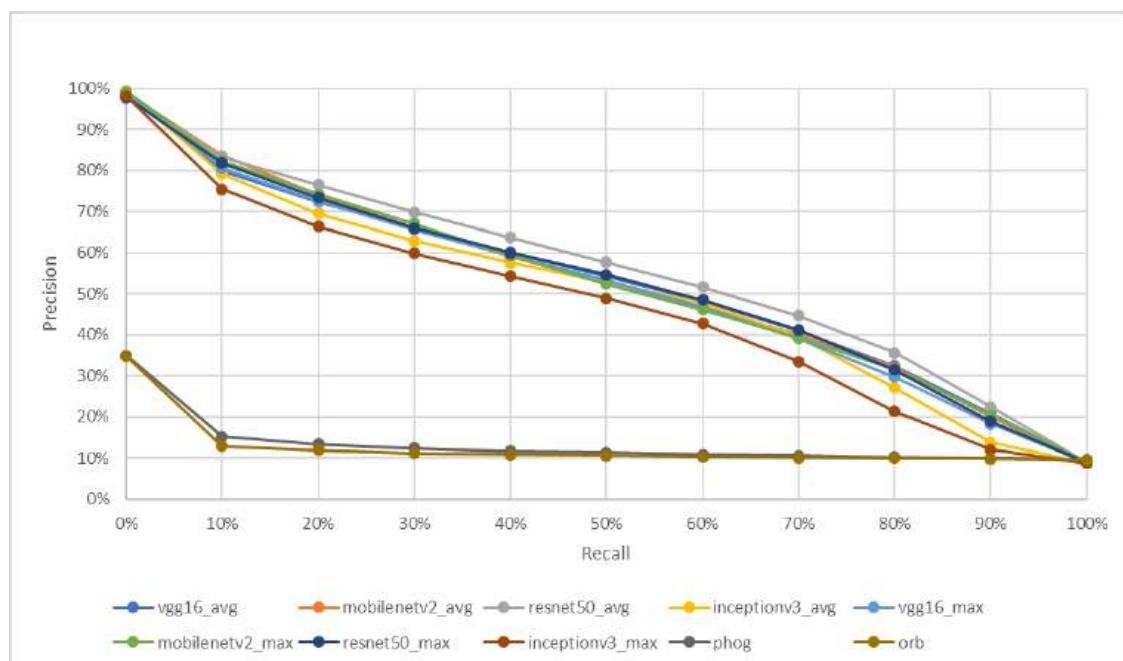


Figure 13. Precision-Recall Curve for Image Descriptors.

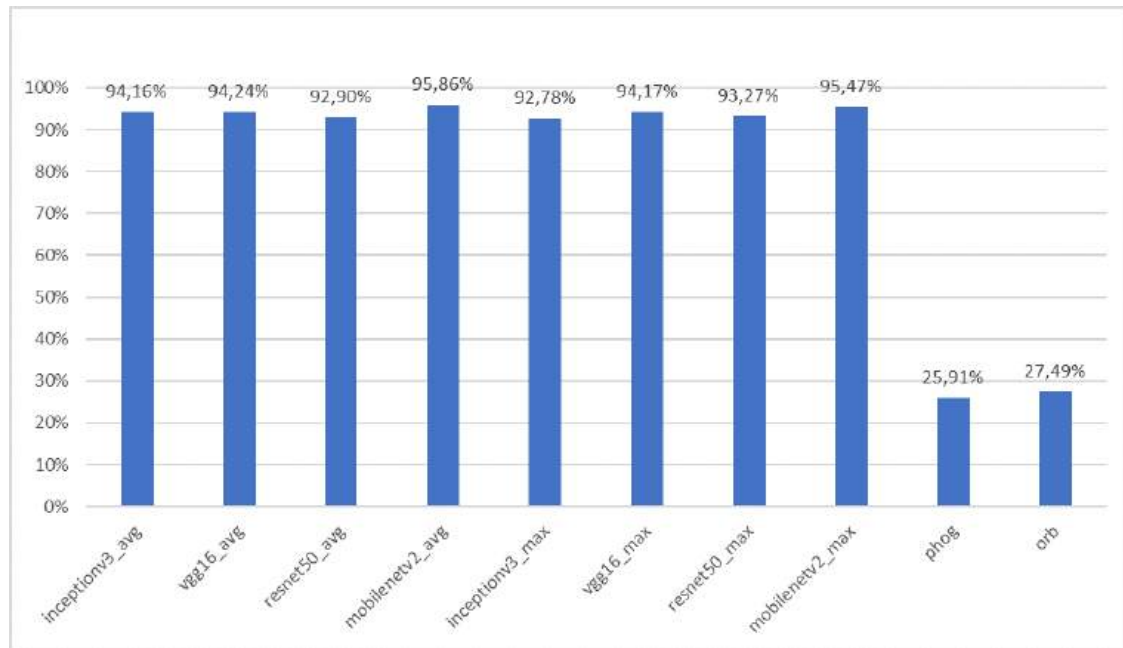


Figure 14. MAP@10 for Image Descriptors.

In the Figures 15 and 16 are reported in all experiments results on the DMOZ collection. Considering the results the best measure is the combination of the path similarity with the deep descriptor. The experiments show that on a real and general web document collection as DMOZ the path similarity is better than Resnick. We argue that the previous results depend on the specific used data sets. Moreover, the combination with the deep descriptor improves the results in both case.

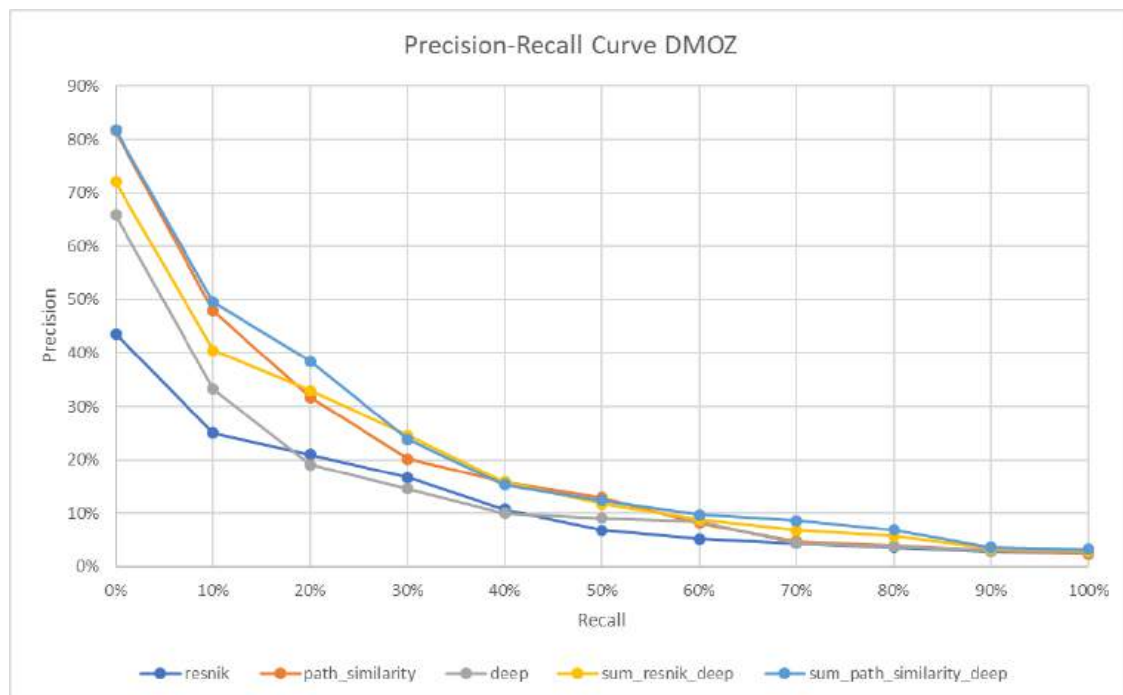


Figure 15. Precision-Recall Curve DMOZ.

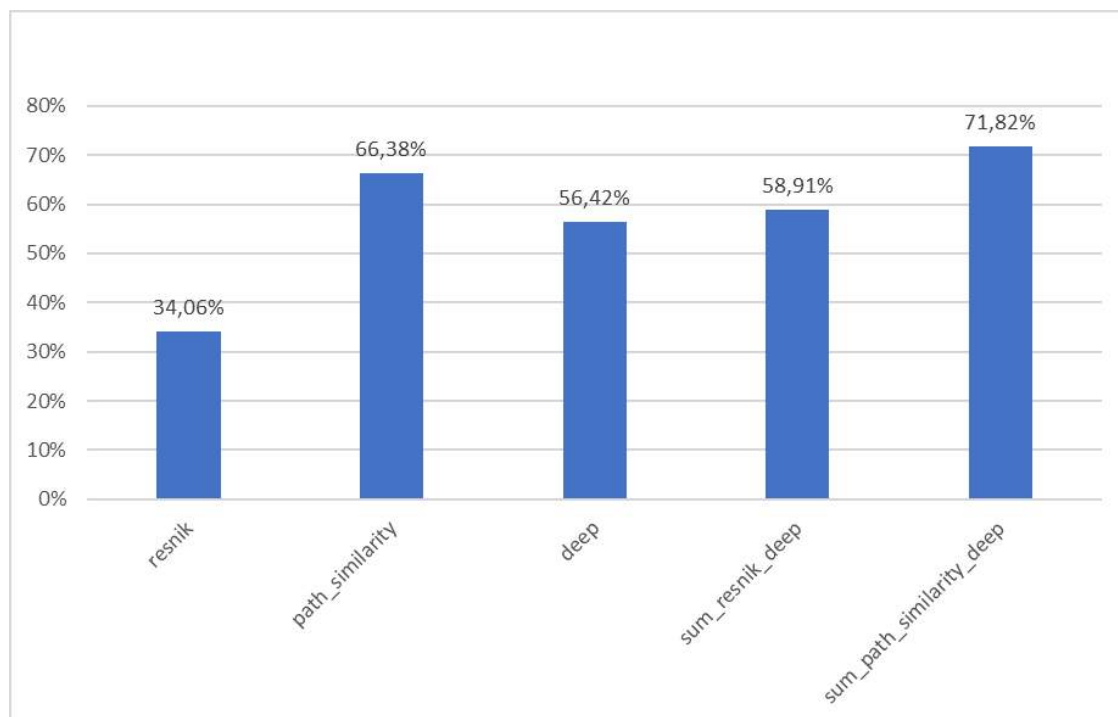


Figure 16. MAP@10 DMOZ.

5. Conclusions and Future Work

In this paper, we proposed a novel and complete framework for multimedia information retrieval using semantic retrieval techniques combined with content-based image descriptors. The paper presents several novelties. The use of a formal multimedia knowledge base allows us to have excellent results improving the precision of used techniques compared to literature. The documents retrieved to the user are more relevant for the query due to the possibility of discriminating the different meanings of the words used in the query. In this context, the implementation of an automatic WSD task in the information retrieval task improve considerably the performance of whole retrieval process. We also presented a system based on an hybrid big data technology that integrates graph-based knowledge representation and document based approach. Moreover, different kinds of image descriptors have been added in our knowledge based to improve the representation of concepts. A deep testing shows very promising results in the combination of textual semantic metrics and deep image descriptors. Finally, we implement and test a visual query with automatic concept extraction which simplifies the query posing process obtaining comparable results with other test cases. We explicitly point out that the modularity of the proposed framework allows an easily extension of our system with other functionalities. Our future works will focus in the definition and implementation of a novel multimedia semantic measure considering both textual and multimedia information stored in our knowledge and different combination strategies. Moreover, we will investigate on the automatic generation of stories starting from the retrieved relevant documents to improve the serendipity for the user and the integration of specific knowledge domains [60,61]. In future works we will investigate on efficient techniques to improve the number of multimedia contents in ImageNet. We think that this task could improve our results, having a more comprehensive knowledge base. Moreover, we will design and implement a novel WSD method based on the analysis of the relationships among concepts in a document using our knowledge base. In addition, the use of structured text as HTML fields can enhance the efficiency and effectiveness of our retrieval methodology. We will improve the evaluation of our novel approach and additional methods using a larger query set and a document collection with high polysemic document categories. Our approach will be compared with other similar baselines proposed in literature in order to prove the effectiveness of our framework.

Author Contributions: Conceptualization: A.M.R.; investigation: C.R. and C.T.; methodology: A.M.R. and C.T.; software: C.R. and C.T.; supervision: A.M.R.; validation: A.M.R.; Writing—original draft, C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd ed.; Addison-Wesley Publishing Company: Boston, MA, USA, 2011.
2. Rinaldi, A.M. An ontology-driven approach for semantic information retrieval on the web. *ACM Trans. Internet Technol. (TOIT)* **2009**, *9*, 10. [CrossRef]
3. Saracevic, T. Relevance: A review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.* **1975**, *26*, 321–343. [CrossRef]
4. Swanson, D.R. Subjective versus objective relevance in bibliographic retrieval systems. *Libr. Q.* **1986**, *56*, 389–398. [CrossRef]
5. Harter, S.P. Psychological relevance and information science. *J. Am. Soc. Inf. Sci.* **1992**, *43*, 602–615. [CrossRef]
6. Barry, C.L. Document representations and clues to document relevance. *J. Am. Soc. Inf. Sci.* **1998**, *49*, 1293–1303. [CrossRef]
7. Park, T.K. The nature of relevance in information retrieval: An empirical study. *Libr. Q.* **1993**, *63*, 318–351.
8. Vakkari, P.; Hakala, N. Changes in relevance criteria and problem stages in task performance. *J. Doc.* **2000**, *56*, 540–562. [CrossRef]
9. Saracevic, T. Relevance reconsidered. In Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2), Seattle, WA, USA, 13–16 October 1996; ACM: New York, NY, USA, 1996; pp. 201–218.
10. Miller, K. *Communication Theories*; Macgraw-Hill: New York, NY, USA, 2005.
11. Danesi, M.; Perron, P. *Analyzing Cultures: An Introduction and Handbook*; Indiana University Press: Bloomington, IN, USA, 1999.
12. Rinaldi, A.M.; Russo, C. User-centered information retrieval using semantic multimedia big data. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2304–2313.
13. Smeulders, A.W.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1349–1380. [CrossRef]
14. Chen, Y.; Wang, J.Z.; Krovetz, R. An unsupervised learning approach to content-based image retrieval. In Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, Paris, France, 4 July 2003; Volume 1, pp. 197–200.
15. Rui, Y.; Huang, T.S.; Chang, S.F. Image retrieval: Current techniques, promising directions, and open issues. *J. Vis. Commun. Image Represent.* **1999**, *10*, 39–62. [CrossRef]
16. Liu, Y.; Zhang, D.; Lu, G.; Ma, W.Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognit.* **2007**, *40*, 262–282. [CrossRef]
17. Eakins, J.; Graham, M. Content-Based Image Retrieval. 1999. Available online: <http://www.leeds.ac.uk/educol/documents/00001240.htm> (accessed on 2 September 2020).
18. Meng, L.; Huang, R.; Gu, J. A review of semantic similarity measures in wordnet. *Int. J. Hybrid Inf. Technol.* **2013**, *6*, 1–12.
19. Wang, S.; Han, K.; Jin, J. Review of image low-level feature extraction methods for content-based image retrieval. *Sens. Rev.* **2019**, *39*, 783–809. [CrossRef]
20. Zheng, L.; Yang, Y.; Tian, Q. SIFT meets CNN: A decade survey of instance retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1224–1244. [CrossRef]
21. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 401–408.

22. Mikolajczyk, K.; Tuytelaars, T. Local Image Features. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A.K., Eds.; Springer: Boston, MA, USA, 2015; pp. 1100–1105.
23. Introduction to SIFT (Scale-Invariant Feature Transform). Available online: https://docs.opencv.org/master/da/df5/tutorial_py_sift_intro.html (accessed on 1 September 2020).
24. Introduction to SURF (Speeded-Up Robust Features). Available online: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_surf_intro/py_surf_intro.html (accessed on 1 September 2020).
25. ORB (Oriented FAST and Rotated BRIEF). Available online: https://docs.opencv.org/3.4/d1/d89/tutorial_py_orb.html (accessed on 1 September 2020).
26. Karami, E.; Prasad, S.; Shehata, M. Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images. *arXiv* **2017**, arXiv:1710.02726.
27. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.; Asari, V.K. A state-of-the-art survey on deep learning theory and architectures. *Electronics* **2019**, *8*, 292. [CrossRef]
28. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
32. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
33. Wan, J.; Wang, D.; Hoi, S.C.H.; Wu, P.; Zhu, J.; Zhang, Y.; Li, J. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM International Conference on Multimedia, Mountain View, CA, USA, 18–19 June 2014; pp. 157–166.
34. Leng, C.; Zhang, H.; Li, B.; Cai, G.; Pei, Z.; He, L. Local Feature Descriptor for Image Matching: A Survey. *IEEE Access* **2019**, *7*, 6424–6434. [CrossRef]
35. Chang, E.; Goh, K.; Sychay, G.; Wu, G. CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. Circuits Syst. Video Technol.* **2003**, *13*, 26–38. [CrossRef]
36. Zhao, R.; Grosky, W.I. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *IEEE Trans. Multimed.* **2002**, *4*, 189–200. [CrossRef]
37. Wang, X.J.; Ma, W.Y.; Xue, G.R.; Li, X. Multi-model similarity propagation and its application for web image retrieval. In Proceedings of the 12th Annual ACM International Conference on Multimedia, New York, NY, USA, 10–16 October 2004; pp. 944–951.
38. Clinchant, S.; Ah-Pine, J.; Csurka, G. Semantic combination of textual and visual information in multimedia retrieval. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Trento, Italy, 18–20 April 2011; pp. 1–8.
39. Giordano, D.; Kavasidis, I.; Pino, C.; Spampinato, C. A semantic-based and adaptive architecture for automatic multimedia retrieval composition. In Proceedings of the 2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Madrid, Spain, 13–15 June 2011; pp. 181–186.
40. Buscaldi, D.; Zargayouna, H. Yasemir: Yet another semantic information retrieval system. In Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, San Francisco, CA, USA, 28 October 2013; pp. 13–16.
41. Kannan, P.; Bala, P.S.; Aghila, G. A comparative study of multimedia retrieval using ontology for semantic web. In Proceedings of the IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM-2012), Nagapattinam, Tamil Nadu, India, 30–31 March 2012; pp. 400–405.
42. Moscato, V.; Picariello, A.; Rinaldi, A.M. Towards a user based recommendation strategy for digital ecosystems. *Knowl.-Based Syst.* **2013**, *37*, 165–175. [CrossRef]

43. Cao, J.; Huang, Z.; Shen, H.T. Local deep descriptors in bag-of-words for image retrieval. In Proceedings of the on Thematic Workshops of ACM Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 52–58.
44. Boer, M.H.D.; Lu, Y.J.; Zhang, H.; Schutte, K.; Ngo, C.W.; Kraaij, W. Semantic reasoning in zero example video event retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *13*, 1–17. [CrossRef]
45. Habibian, A.; Mensink, T.; Snoek, C.G. Videostory: A new multimedia embedding for few-example recognition and translation of events. In Proceedings of the 22nd ACM International Conference on Multimedia, Mountain View, CA, USA, 18–19 June 2014; pp. 17–26.
46. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41.10.1145/219717.219748. [CrossRef]
47. Purificato, E.; Rinaldi, A.M. Multimedia and geographic data integration for cultural heritage information retrieval. *Multimed. Tools Appl.* **2018**, *77*, 27447–27469. [CrossRef]
48. Rinaldi, A. A multimedia ontology model based on linguistic properties and audio-visual features. *Inf. Sci.* **2014**, *277*, 234–246. [CrossRef]
49. Rinaldi, A.M.; Russo, C. A semantic-based model to represent multimedia big data. In Proceedings of the 10th International Conference on Management of Digital EcoSystems, Tokyo, Japan, 25–28 September 2018; pp. 31–38.
50. Web Ontology Language. Available online: <https://www.w3.org/OWL/> (accessed on 1 September 2020).
51. ImageNet. Available online: <http://www.image-net.org/> (accessed on 1 September 2020).
52. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, Toronto, ON, Canada, 8–11 June 1986; pp. 24–26.
53. Vasilescu, F.; Langlais, P.; Lapalme, G. Evaluating Variants of the Lesk Approach for Disambiguating Words. Available online: <http://www.iro.umontreal.ca/~felipe/Papers/paper-lrec-2004.pdf> (accessed on 27 October 2020).
54. Tolias, G.; Sicre, R.; Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. *arXiv* **2015**, arXiv:1511.05879.
55. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
56. Kittler, J. Combining classifiers: A theoretical framework. *Pattern Anal. Appl.* **1998**, *1*, 18–27. [CrossRef]
57. 20 Newsgroups Scikit-Learn. Available online: https://scikit-learn.org/0.15/datasets/twenty_newsgroups.html (accessed on 1 September 2020).
58. Visual Object Classes Challenge 2012 (VOC2012). Available online: <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/> (accessed on 1 September 2020).
59. DMOZ Website. Available online: <https://dmoz-odp.org/> (accessed on 1 September 2020).
60. Caldarola, E.; Rinaldi, A. A multi-strategy approach for ontology reuse through matching and integration techniques. *Adv. Intell. Syst. Comput.* **2018**, *561*, 63–90.
61. Rinaldi, A.M.; Russo, C. A matching framework for multimedia data integration using semantics and ontologies. In Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 363–368.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Visualization, Interaction and Analysis of Heterogeneous Textbook Resources

Christian Scheel ^{1,*†‡}, Francesca Fallucchi ^{1,2,‡}  and Ernesto William De Luca ^{1,2,†‡}

¹ Georg Eckert Institute for International Textbook Research Member of the Leibniz Association, 38114 Braunschweig, Germany; f.fallucchi@unimarconi.it (F.F.); deluca@gei.de (E.W.D.L.)

² Department of Innovation and Information Engineering, Guglielmo Marconi University, 00193 Rome, Italy

* Correspondence: scheel@gei.de

† Current address: Celler Straße 3, D-38114 Braunschweig, Germany

‡ These authors contributed equally to this work.

Received: 8 October 2020; Accepted: 17 October 2020; Published: 21 October 2020

Abstract: Historically grown research projects, run by researchers with limited understanding of data sustainability, data reusability and standards, often lead to data silos. While the data are very valuable it can not be used by any service except the tool it was prepared for. Over the years, the number of such data graveyards will increase because new projects will always be designed from scratch. In this work we propose a Component Metadata Infrastructure (CMDI)-based approach for data rescue and data reuse, where data are retroactively joined into one repository minimizing the implementation effort of future research projects.

Keywords: textbook research; digital humanities; digital infrastructures; data analysis

1. Introduction

The availability of Big Data has boosted the rapidly emerging new research area of Digital Humanities (DH) [1,2] where computational methods have been developed and applied to support solving problems in the humanities and social sciences. In this context, the concept of Big Data has been revised and has another connotation, which regards the data being too big or complex to be analyzed manually by a close reading [3]. Educational media research suffers from this overwhelming availability of information and will likely get a boost by the DH, as it will be possible to analyze the sheer amount of historical textbooks on international level.

For instance, educational media research investigates sanctioned knowledge by comparing textbooks and identifying modified, missing or added information. Such modifications can have a big impact for the formation of the young generation. Hence, textbooks are also gaining importance in the historical research [4]. Because there are millions of digitized textbook pages available, researcher's search for "popular knowledge", as it reflects views of the world, thought flows and desired knowledge has to be supported by digital research tools. To be able to work with processed data, the digital research tools rely on digitization efforts and people who make implicit knowledge explicit by describing said resources. In this direction, we developed Toolbox [5] a complete suite for analysis in Digital Humanities. The system offers functionalities that allow text digitalization, Optical Character Recognition(OCR), language recognition, digital libraries management, topic modeling, word counting, text enrichment and specific reporting elements, all in a flexible and highly scalable architecture.

From a digital research point of view, data derived from textbooks and educational media provide interesting challenges. First, processed data were often not meant to be reused. Hence, the data are hard to retrieve and has to be processed again to be of any value for the research community. Second, if processed data are available, they are syntactically heterogeneous (text, images, videos, structured

data in different formats, such as XML, JSON, CSV and RDF), they are described in different metadata standards and often written (or cataloged) in different languages, without any use of standards or controlled vocabulary. Third, the data are semantically rich, covering different “views of the world” taken from textbooks of different countries and epochs. Lastly, the data are implicitly interlinked across different data sources, but only accessible by individual and often outdated interfaces.

The separate storage of data silos is not helpful if humanistic researchers want to deal with such data and address semantically complex problems or interesting methodological problems. As a non-university research institution, the Georg Eckert Institute for the International Textbook Research (GEI) (<http://www.gei.de/home.html>).

Conducts and facilitates fundamental research into textbooks and educational media primarily driven by history and cultural studies. For this purpose, the GEI provides research infrastructures such as its renowned research library and various dedicated digital information services. Hence, the institute develops and manages both digital and social research infrastructures. As such, the GEI realizes a unique position in the international field of textbook research. In the digital humanities, the investigation of research questions is supported by a range of increasingly sophisticated digital methods such as automatic image and text analysis, linguistic text annotation, or data visualization. Digital tools and services combined with the increasing amount of resources available through digital libraries such as the German Digital Library, the Deutsches Textarchiv, Europeana and research infrastructures such as Common Language Resources and Technology Infrastructure (CLARIN) or Digital Research Infrastructure for the Arts and Humanities (DARIAH) provide digital support for textbook analysis.

Analogous to the work done in [6] we identified three generations of portals that develop historically grown and individually processed research projects. First, the research focus in semantic portal development was on data harmonization, aggregation, search and browsing (“first generation systems”). At the moment, the rise of Digital Humanities research has started to shift the focus to providing the user with integrated tools for solving research problems in interactive ways (“second generation systems”). The future portals not only provide tools to solve problems, but can also be used for finding research problems in the first place, for addressing them and even for solving them automatically by themselves under the constraints set by the humanities; however, to reach or even think about the possibilities of such “third generation system”, some challenges like semantic interoperability and data aggregation have to be approached first.

In order to being able to embed institute’s data into these resources, it has to be separated from existing historically grown research tools, to be joined in a single repository. Research projects, tailored for specific research questions, often result in graphical interfaces only usable for satisfying one given information need. Nevertheless, the underlying data are not limited to such use cases and could often also be used for searching, visualizing and exploring data. In this work, we show how overlaps and missing overlaps of these data silos can be disclosed with the Component Metadata Infrastructure (CMDI) [7] approach in order to retroactively disclose planning deficits in each project. Generalizing these shortcomings helps projects to be more focused on data reuse, user group multilingualism, the provision of standardized interfaces and the use of unified architectures and tools.

After describing the problem in detail and giving an overview about the lessons learned from visualization, data exploration and interactive data approaches, we show how to overcome the dispersion produced by data silos, a multitude of metadata formats and outdated tools using the CMDI, suggested by CLARIN. CMDI can help to not only emphasize the common characteristics in the data, but also keep the differences. Concluding, we show that the visualization, data exploration and interactive data approaches can be applied to the newly created repository, gaining additional research value from the newly known interconnections between the formerly separated data.

2. Problem Description

In the recent past, the Georg Eckert Institute has generated many data silos whose origins lie in historically grown and individually processed research projects. The data available in search indices or databases are fundamentally different, but have many common characteristics (such as title, persons, year and link to resource). Because the institute prescribes the research direction, the data from the research projects are thematically related, which is reflected not only in the common characteristics but also in their characteristic values. The separate storage of data silos is not desirable because, firstly, data is kept twice and, secondly, no project can benefit from the other.

In the following, the data, their similarities and their significance for data harmonization and interconnection are described, followed by data approaches (visualization, exploration and interaction), which can be observed on this data, to raise a common understanding of their potential benefits for a harmonized data repository.

2.1. Recording Characteristics and Characteristic Values

We started to explore each project's underlying data, in order to merge them and to get rid of data silos. Initial investigations had shown that the data structure was always very flat, even when complex objects were described. Whenever certain characteristics were present in most projects, but could not be satisfied by another project's data sources, the question was how and where to extract or substitute it from other projects' underlying data. We organized different workshops and analyzed project documentations together with the experts of the research field and with the users of the corresponding research tools. We learned that the observed differences between the common characteristic expressions resulted from missing knowledge about former and current projects. Hence, having common vocabulary, coming from standards or standard files, has never been an option. For merging projects' data and applying standards or standard files, we analyzed the following twelve project's resources for their characteristics:

- edu.docs (<http://www.gei.de/en/departments/digital-information-and-research-infrastructures/edumeres-the-virtual-network-for-international-textbook-research/edudocs-publications-from-educational-media-research.html>) (202 resources)
- edu.reviews (<http://www.gei.de/en/departments/digital-information-and-research-infrastructures/edumeres-the-virtual-network-for-international-textbook-research/edureviews-the-platform-for-textbook-reviews.html>) (371 resources)
- edu.data (<http://www.gei.de/en/departments/digital-information-and-research-infrastructures/edumeres-the-virtual-network-for-international-textbook-research/edudata-textbook-systems-worldwide.html>) (2796 resources)
- edu.news (<http://www.gei.de/en/departments/digital-information-and-research-infrastructures/edumeres-the-virtual-network-for-international-textbook-research/edunews-the-latest-from-educational-media-research.html>) (4064 resources)
- Curricula\Workstation (<https://curricula-workstation.edumeres.net/en/curricula/>) (7687 resources)
- K10plus (<https://www.bszgbv.de/services/k10plus/>) (search index of the library; 183,295 resources)
- GEI.de (<http://www.gei.de>) (the institute's website; 546 resources)
- GEI|DZS (<https://gei-dzs.edumeres.net/en/search/>) (2641 resources)
- WorldViews (<http://worldviews.gei.de>) (57 resources)
- GEI Digital (<http://gei-digital.gei.de/viewer/>) (5200 resources)
- Pruzzenland (<https://www.pruzenland.eu>) (116 resources)
- Zwischentöne (<https://www.zwischentoene.info/themen.html>) (461 resources)

Below, we report an analysis of the most important bibliographic metadata used to record resource information in the different projects. When preparing the harmonization of data from different data sources, it is important to focus on the similarities of these resources, in order to not getting overwhelmed by individual differences. Additionally, these similarities are most likely the data which can interconnect the resources.

We formalize a bibliographic dataset (D) as follows. D is a set of 8-tuple $d \in D = (id; url; t; p; c; T; s; l)$ where:

id are unique identifiers of the resource or to other resources.

url is a link to the resource.

t is the title of the resource.

p is the published/publisher of the resource.

c is the created/changed information of resource.

T represents the topics of the resource, a set of three resources $T = k, sa, dt$ where

k are the keywords or tags.

sa are the subject areas.

dt are places or descriptive terms.

s is the information related to level of education, school type, country of use or subject of the resource.

l is the language of the resource.

2.1.1. Identifiers

The most straight forward way of data harmonization is looking for the same identifiers within different resources and hence, identifying two descriptions d_1 and d_2 of the same resource or resources which are linked to each other. Although there is often a field “id” in the data, this attribute is not necessarily the data to look for, because it often just separates this resource from other resources of the same source. Talking with experts about exemplary resources will provide knowledge about fields which contain identifiers.

2.1.2. URL

When merging data from different sources, the URL should be used to reference the original data. The URL describes a fixed web address, which can be used to view an entry in the corresponding project. Accordingly, all URLs are different and cannot be limited by a prescribed vocabulary. We observed indirect URLs, where links led to a descriptive overview pages, generated by the containing projects. For data harmonization, these URLs had no value, because the information presented on these pages was already part of d . Within 7 of the 12 projects there was at least one link that led to the original resource. With the remaining five projects, the URL could be assembled with the help of static character strings and available information (e.g., from identifiers). The total coverage of this metadata was 99.83%.

2.1.3. Title

Titles are a very short textual description of an entry and are often combined with the URL to create a human readable link to the original resource. Intuitively one would assume that every entry in every project has a title. However, this is only the case for 99.62% of the resources. In all but four projects, there was a complete title assignment. Further investigations have shown that some of these documents were not missing the title in the original data source, but must have been lost when preparing the data for searching and presentation for the project’s interface. For some resources, such as maps, a title was not always necessary.

2.1.4. Published, Publisher

Knowing when and by whom an entry was published is an important feature. Three of the twelve projects did not have this feature at all. This includes the institute’s website, Pruzzenland and edu.data.

The information on the publisher was also missing for two other projects: edu.news and Zwischentöne. In case of missing publisher information, the publisher often was the institute itself. The total coverage of “published” is 92.31% and that of “publisher” 91.33%.

2.1.5. Created, Changed

Since the project’s individual search indices have never been deleted and providing this service ever since, we were able to gain two pieces of information that are of great value for a common representation. The values “created” and “changed” managed by the search index were not found in the underlying databases. However, they describe very well and independently from the publication date when an entry was added to the corresponding project. It can also be considered as a substitute for publication date if this information is missing. Information on “changed” was available in 11 of 12 projects and on “created” in 10 of 12. The total coverage of “changed” was 94.64% and that of “created” 10.89%, because the research library resources are missing this information.

2.1.6. Topic

By topic we mean keywords, subject areas, places or descriptive terms. Even if they were not necessarily a descriptive topic term in the original project, the total coverage is 95.26%. Interestingly, it was news related project (edu.news) where such descriptive information is missing. This shows retroactively an error with the conception of this project, because keywords and geographical information are common information in news articles. Fortunately, the keywords and topics often have been linked with external knowledge bases (e.g., GND).

2.1.7. Level of Education, School Type, Country of Use, School Subject

Because it was important for educational media research, but is not part of traditional cataloging, the institute decided to establish a classification for textbook characteristics. The research, the textbook collection and hence the local classification scheme of the Georg Eckert Institute are primarily focused on educational sciences, history, geography, political science and religious sciences [8–10]. As these characteristics are specific characteristics of textbooks and related media, this information had the greatest overlap between the projects. However, recent projects have shown the need to map “level of education” and “school subject” into the UNESCO International Standard Classification of Education (ISCED) to be able to cover international educational media [4].

2.1.8. Language

Information about the language of the entries were often given implicitly, like when the whole data source was written in one language. The language in which the entries were written is unknown in half of the projects.

2.2. Data Approaches

Research projects in our increasingly data- and knowledge-driven world are dependent on applications that build upon the capability to transparently fetch heterogeneous yet implicitly connected data from multiple, independent sources. Even though, all projects have been driven by the respective research goals, the resulting tools generally show how research could benefit from data-driven visualization, exploration and interaction approaches. The data inspection described in Section 2.1 made it obvious that there would be no short term solution for harmonizing underlying data of the projects, so that the projects could switch to the new data repository. Instead it revealed the long-term need to research and develop new projects that could interact together with the very large amounts of complex, interlinked, multi-dimensional data, throughout its management cycle, from generation to capture, enrichment in use and reuse and sharing beyond its original project context. Furthermore, the possibility of traversing links defined within a dataset or across

independently-curated datasets should be an essential feature of the resulting tools and thus to ensure the benefits for the Linked Data (LD) [11] community.

In the following, we further describe the reuse and reusability of the products of the different projects analyzing the benefit from data-driven visualization, exploration and interaction approaches in more details.

2.2.1. Visualizing Data

The design of user interfaces for LD, and more specifically interfaces that represent the data visually, play a central role in this respect [12–14]. Well-designed visualizations harness the powerful capabilities of the human perceptual system, providing users with rich representations of the data. Dadzie and Pietriga illustrate in [15] the design and construction of intuitive means for end-users to obtain new insight and gather more knowledge. As a cultural institution, the GEI digitizes and interlinks its collections providing new opportunities of navigation and search. However, it is comprehensive that the data are sparse, complex and difficult to interact with, so that a good design and support of the systems is indispensable. Moreover, it is difficult to grasp their distribution and extent across a variety of dimensions.

An important promise in connection with the digitization efforts of many institutions of cultural heritage is increased access to our cultural heritage [16]. Aggregators, such as the Digital Public Library of America and Europeana expand this ambition by integrating contents from many collecting institutions so as to let people search through millions of artifacts of varied origins. Due to the size and diversity of such composite collections, it can be difficult to get a sense of the patterns and relationships hidden in the aggregated data and the overall extent of the collection [17]. Therefore, new ways of interaction and visualization possibilities can help in finding relevant information more easily than before [18,19]. We developed different tools and visualizations for accessing educational media data in various project.

An example is given by the platform GEI-Digital (<http://gei-digital.gei.de/viewer/>), which is a first generation system that provides more than 4300 digitized historical German textbooks in the fields of history, geography and politics, including structural data (e.g., table of contents and table of figures) and OCR processed text from more than one million pages. Both textbooks from the Georg Eckert Institute and textbooks from other partner libraries were digitized and integrated. GEI-Digital aggregates the entire collection of German textbooks until 1918. In the course of digitization, a total of 250,000 metadata were recorded, whereby the indexing follows the specific needs of textbook research. Thus, in addition to information about the publisher and year of publication, subjects and grades were recorded as meta data. However this tool does not provide any visually appealing information, except from the presentation of the scanned textbook pages and figures, which offers researchers the opportunity to print sections and work directly on these copies [20].

To overcome this deficit, the prototypical visualizations of “GEI-Digital visualized” (<http://gei-digital.gei.de/visualized/>) have been developed in cooperation with the Potsdam University of Applied Sciences in the Urban Complexity Lab as part of a research commission from the Georg Eckert Institute for International Textbook Research. Through the visualization of the metadata and interactive combination possibilities, developments on the historical textbook market with its actors and products can be made visible. This tool illustrates the prerequisites and possibilities of data visualization, while being limited to only data coming from GEI-Digital [21]. Letting researchers use this tool and observing their interaction, we analyzed the added value given by data visualizations in combination with library content, on the one hand, and the research purposes on the other (see Figure 1).

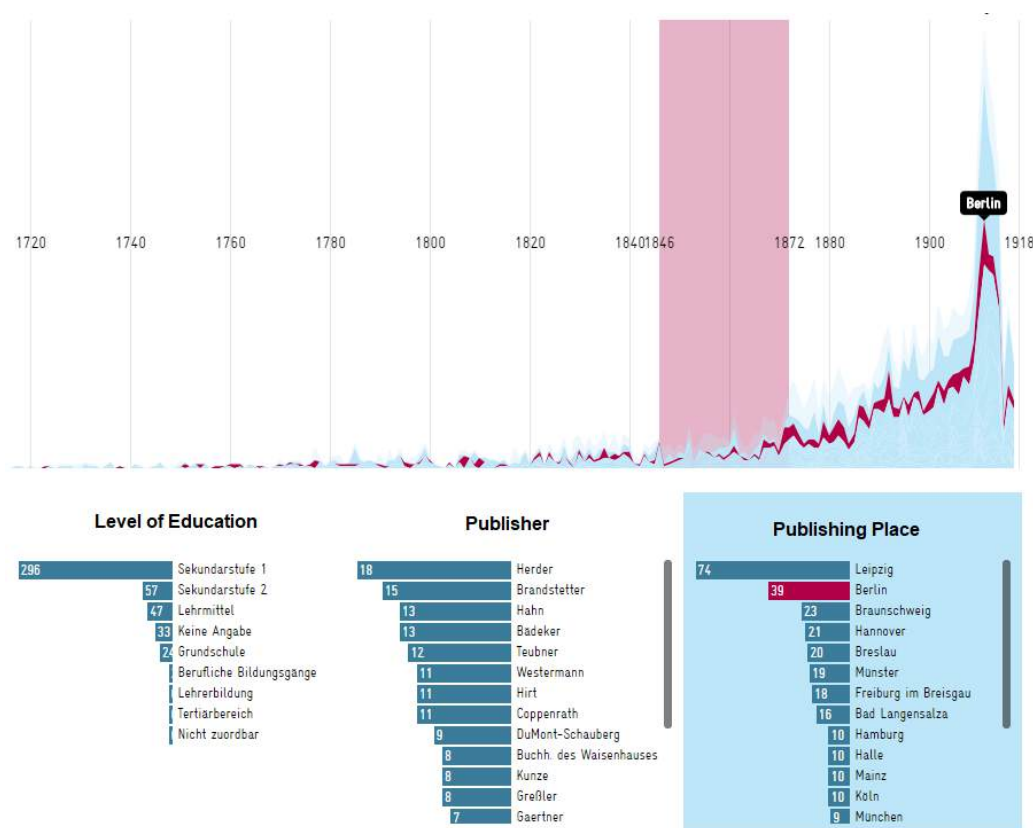


Figure 1. Screenshot of the Georg Eckert Institute for the International Textbook Research (GEI)-Digital-Visualized tool.

2.2.2. Data Exploration

Within the “Children and their world” Explorer, we implemented a second generation tool, which shows how texts, included in the corpus, can be exported and used in other DH tools for further detailed analysis (see Figure 2). Researchers can work with a set of texts and look for ways to reveal structural patterns in their sources, which were, until now, impossible to analyze within a classical hermeneutical way. This interdisciplinary DH project deals with world knowledge of the 19th-century reading books and children’s books. The digital information (a sub corpus of the GEI-Digital textbook collection combined with the Hobrecker collection [22], a children’s book collection) has been combined to implement specific tools for semantic search and statistical text analysis, which can support researchers to better formulate their research questions and to support the serendipity effect, which can be given by the use of digital tools. To this end, approximately 4300 digitized and curated 19th-century historical textbooks have been annotated at the page level using topic modeling and automatic enrichment with additional meta data. These extensions enable a free browsing possibility and a complex content and meta data driven search process on textbooks. For supporting the research goals of this project, a sub set of the books were manually annotated by the supposed target gender (male, female, both or unknown) or the targeted religious confession. The International TextbookCat research instrument (see Figures 3 and 4) does not only provide a welcome extension to the library OPAC system, but also is a discovery tool that dramatically improves the (re)search possibilities within underlying textbook collections. In contrast to the content driven “Children and their world” Explorer, which is dependent on the digitization process, the International TextbookCat is solely based on metadata and hence provides access to much more textbooks. It employs the local classification system (see Section 2.1.7) in order to categorize textbooks according to applicable country, education level and subject. Additional categories of federal state and school type are provided for German textbooks. The project extends the textbook collection with the inventories

of international partners, combining the textbook databases of three institutions: the Georg Eckert Institute (165,231 resources), the University of Turin (25,661 resources) and the National Distance Education University in Spain (66,556 resources), in order to create a joint reference tool [23]. Workflows and system architecture have been developed that in the long-term will enable further institutions to participate with relatively little effort on their part. An additional functionality is given in the statistics view. Diagrams illustrate features and compositions of the collection or currently selected set (see Figure 4), which on its own can be seen as an visualization approach. Researchers can use this feature for the development or verification of their hypothesis and research questions.



Figure 2. Screenshot of the Digital Humanities Tool “Children and their world” Explorer.

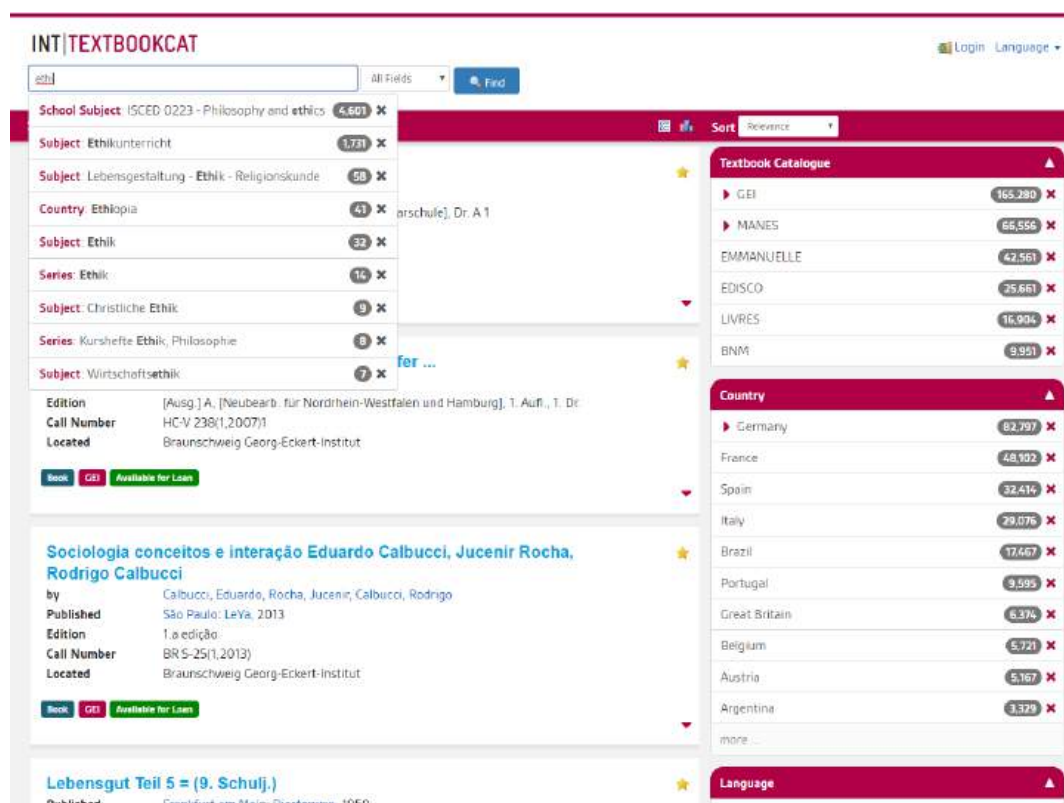


Figure 3. Screenshot of the International TextbookCat Research Tool.

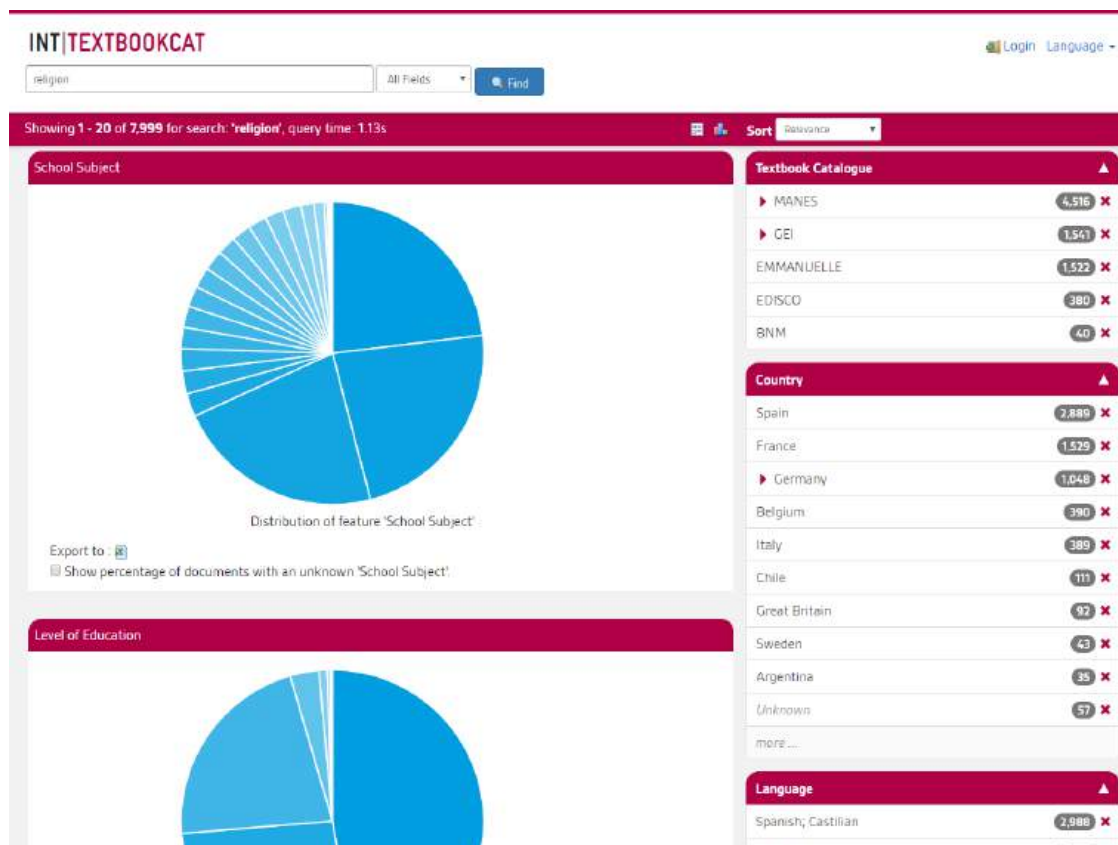


Figure 4. Presentation of the statistics given a query in the International TextbookCat Research Tool.

2.2.3. Data Interaction

The presented approaches encourage the user to interact with their underlying data in order to examine research questions or just in hope of the serendipity effects which could lead to new hypotheses. However, this interaction has no effect on the data itself. None of the examined projects used interaction data to improve itself. However, some projects encouraged the researcher to interact with the institute in order to create reviews, recommend new textbooks to purchase, prioritize books in the digitalization queue, etc.

Recently, we researched about the most desired features for textbook annotation tools and then created SemKoS, an annotation tool prototype, which supports data interaction and creation, based on digitized textbooks. In order to maximize the acceptance of the tool, we did a survey in which we found out what researchers expect from such a tool [20]. As a direct consequence, it was decided that annotations should be made directly on the scanned book pages and not on the corresponding recognized texts, as this supports a working method similar to the traditional work on a book. As it can be seen in Figure 5, text (representing entities) can be linked to a knowledge base, resulting in a better contextual understanding of the textbooks and the development of better data approaches in the future.

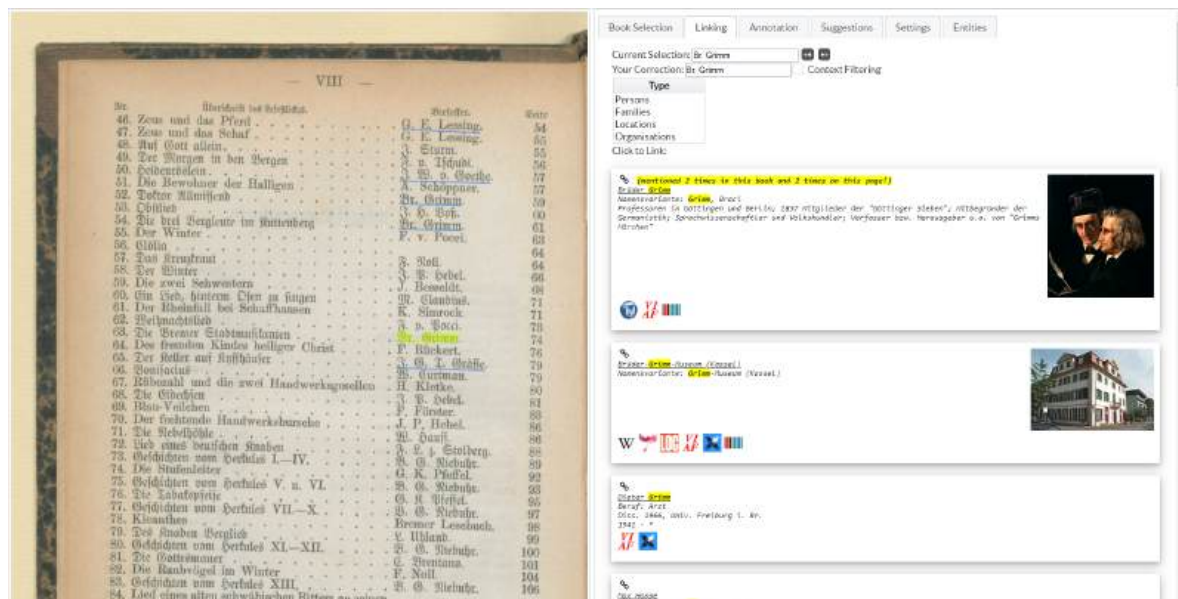


Figure 5. Screenshot of the SemKoS entity linking tool.

3. Creating a Middleware for Continuously Accessing and Joining Data

Developments in the direction of data integration and homogenization has been made within the project WorldViews [24] with its aim of establishing a middleware to facilitate data storing and reuse within the GEI context and beyond that. The project data, being stored in a standard format and accessible through standard interfaces and exchange protocols, will serve as a use case to test the data infrastructure's improvement to facilitate the data's long term sustainability and reuse. To intensify the connection to the Cultural Heritage World (like Deutsche Digitale Bibliothek or Europeana), creating a theoretical knowledge model, covering all types of resources, was essential. The data were found to be in various formats and stem from international and multilingual sources (see Section 2.1). Furthermore, in most project sources, the data were not static. Hence, the middleware had to be able to continuously access and join the data, where joining included cleaning up, mapping to a common representation and representing it in CMDI.

3.1. Accessing the Data

Since the research projects were driven by historical focused research questions, ignorant of the possibilities of later disclosure and reuse of the data, the data structure has been very neglected. Access to the data could be obtained in three ways:

1. Browsing a web based user interface. The projects often offer a search, where the resources data is presented in a "detailed view".
2. Analyzing the internal database of the architectures that make the web presence possible.
3. Analyzing the search indices generated by the provided search functionality.

In an attempt to identify commonalities between the research projects, researchers from the institute took the path (1). In particular, the search masks were examined here, since its drop-down lists often showed the complete assignment of a property (controlled vocabulary). In addition, it was always tracked back where this information originally came from. At the same time, computer scientists tackled (2) and (3), where (2) was too time consuming due to the multitude of different architectures and the associated unmanageable variety of data. The data of the (separately kept) search indices (3) were comparatively easy to access, since they were kept in the same architecture (Solr) (<https://lucene.apache.org/solr/>) and hence, could be automatically accessed via interfaces.

3.2. Mapping the Data to a Common Representation

Defining the mapping of data available in the institute's digital information systems and services was the most time consuming part of data harmonization, because every single feature expression needed a representation in CMDI. In [25–27] we described our previous works on CMDIification process for GEI textbook resources. Often researchers and users were needed to link each expression from the data to the common representation, because these persons knew the real meaning of expressions and would not make assumptions. This process led to a set of mapping rules (like “map language:‘English’ to iso_639_3:‘eng’ ”), which could always be adjusted, extended and applied again, because mapping tools do not manipulate the original data, but the representations in CMDI.

3.3. Application of CMDI Profiles

When developing CMDI, CLARIN assumed that metadata for language resources and tools existed in a variety of formats, the descriptions of which contained specific information for a particular research community. The data within the research projects of the GEI (see Section 2.1), which have been tailored for the closed educational media research community, supports this assumption. Thus, as in CMDI, components can be grouped into prefabricated profiles. The component registry then serves to share components and profiles across the research projects and eventually to make them available to the research community.

3.4. Proof of Concepts

A variety of independent digital services and projects have been implemented in the past, so that researchers which were interested in cross-search possibilities had to find a way themselves to get the most relevant information related to their research questions, if they wanted to use different services. Moreover, researchers who were not familiar with the institute's services and data had no chance of knowing whether it would be worthwhile to learn how to use the tools.

To tackle this issue, the first application using the newly created joint repository was an institute wide search engine. Having the data in one place and knowing its origin, the task of creating this search engine was just configuring a search indexer and designing the user interface. Researchers are now supported in performing cross-research questions, analyzing the data from different perspectives (as shown in Figure 6 (<http://search.gei.de>)) and analyzing found results in their original services in detail. The flow of data from the search engine point of view can be summarized as follows.

1. Project's data is accessed and retrieved continuously by the middleware.
2. Where applicable, the data are then mapped into standards, controlled vocabularies or codes.
3. The resources are then represented in CMDI and stored into a repository.
4. Repository's data are accessed and retrieved continuously by the indexer.
5. The indexer transforms the CMDI representation into a index document representation and stores it in the search index.
6. The index is accessed by the discovery tool (VuFind), which presents results to the user.
7. The codes are translated into corresponding terms of the selected/detected language.

While these steps look overly complicated, only steps (4) to (6) had to be implemented to create the search engine. In fact, in the future, step (4) should be provided by the repository's Application Programming Interface (API), where XSLT(eXtensible Stylesheet Language Transformations) could transform CMDI into other representations.

A second proof of concept did not result in a tool yet, but in supporting a research question. The newly created repository contained and linked data from GEI | DZS, an interface for filtering for approved textbooks, from the Curricula Workstation, a tool for searching for curricula, and textbooks from the library (via Findex). There was the need to filter textbooks for specific criteria, like school

subject, year of admission and level of education, to link the corresponding curricula to the textbooks. Each curriculum describes the wanted knowledge and if a textbook has been approved, we know for sure that this knowledge is in the book. Showing, that linking actual curricula to textbooks enables a variety of new text based approaches. The connection between these three data sources could not be made before, because the data was never meant to be used in other services than they were made for.

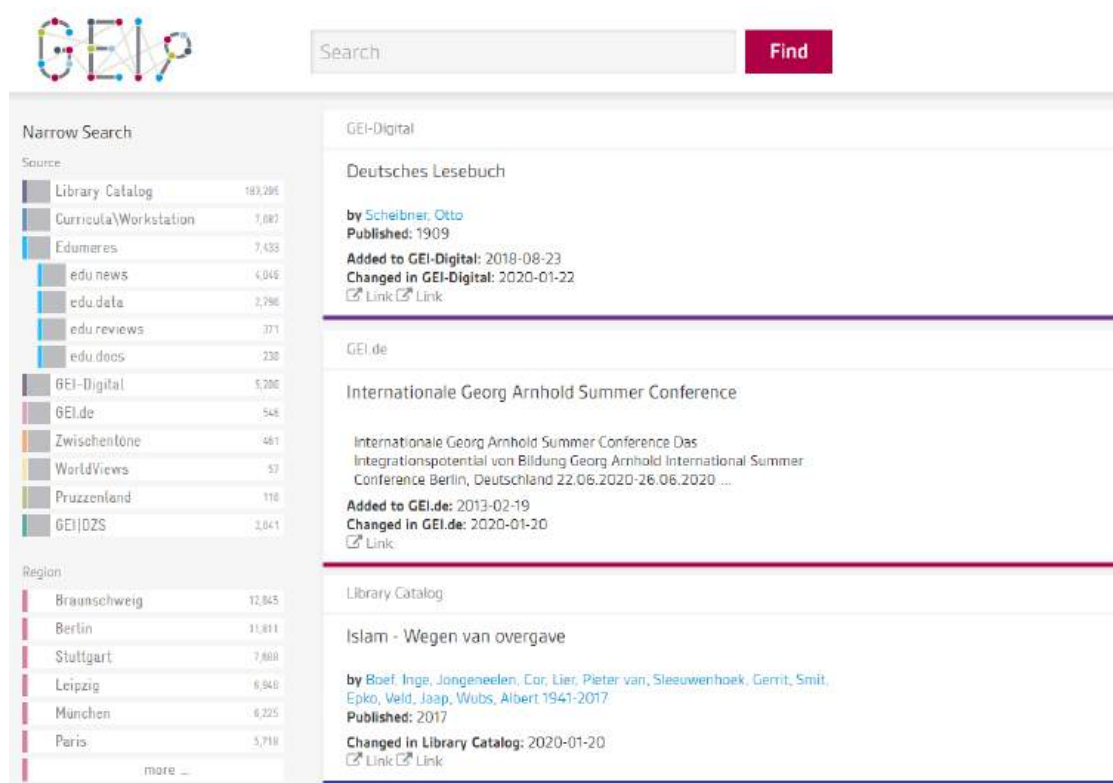


Figure 6. Meta search of data collections.

4. Discussion

In the recent past, the GEI has generated many data silos whose origins lie in historically grown and individually processed research projects. To get rid of these data silos, there was the need to harmonize all project's underlying data. Hence, we started collecting projects' documentations and investigated all options.

After analyzing the data, projects, tools and services of the GEI, we became aware of the great potentials. Not only did we find valuable data, but also tools and services for visualization, exploration and interaction. Even though the tools were designed for different purposes, the general ideas of these applications could be expanded to all the data. For instance, the technology for visually browsing through textbooks (like in GEI-Digital visualized) could be reapplied to visually browsing through textbook admissions or curricula.

The institute approached this undertaking from two perspectives. The researchers (projects' users) reviewed the user interfaces to conclude the data structure and tried to track the data back to its original source. The computer scientists performed a technical approach by analyzing the data in the back-end. While gathering more and more information and understanding where, why and how underlying data were stored, we had to solve new issues on the way. It has been shown that the technical approach can also reveal missing features within a source, while the manual investigation of the source only concluded that this feature exists. Conversely, the manual approach could not detect the existence of properties if they occurred infrequently. From the title of a given document, we could see that the transfer of data from the databases to the search indices did not always have to be complete. Software or planning errors in the corresponding base architectures can lead to more information on

the entries being available than can be found in the search index. Missing fields for publisher and corresponding publication date showed that one should also include implicitly given characteristics in the metadata when planning services, because these can be relevant with a subsequent use. The search indices contained values that could not be found in the databases and the user interface of the projects. Information such as when an entry came into a project or when it was last edited is required if someone wants to know what has changed in the project or if a project is still being managed. When planning new projects, such values should be considered as database entries. A deleted index can be rebuilt at any time, but this information could no longer be reproduced. The lack of information about educational level, school type, country of assignment, subject in edu.reviews' offer is a clear call for the reuse and linking of data, because exactly this information about the reviewed textbooks is contained in the library catalog. Even if the language of the entries is unknown in half of the projects, the technology has now reached the point where the language can be reliably determined and added. This example is representative of many metadata that can be derived from other sources or supplemented in order to make the existing data as complete as possible. Metadata fields such as keywords, subject areas and locations can often also be re-used as general topics. Such a field would be comparable with the GND keywords, which are equally diverse. This means that a field is created here which does not necessarily have to occur in any project. Here the data-driven approach was advantageous, because all topics could be assigned an ID, which links the keywords with the GND. The use of IDs instead of natural language entries also promotes multilingualism, since linked data is often translated into different languages. A decisive advantage when investigating the interface was that the experts always asked themselves: "Where does this data come from?" Even though the indices provide a good overview, they also showed that data were manipulated or lost on the way to the index. Knowing which source they were fed from is indispensable for setting up component registry. The evaluation showed that it was useful to analyze the data in parallel by experts who were familiar with the database and the projects and computer scientists who combined similarly filled index fields pragmatically to create the basis for a common database. The approach of the subject scientists led to detailed investigations of the characteristic values of meaningful characteristics, while the approach of the computer scientists revealed common characteristics. Both approaches complemented each other to enable the generation of CMDI profiles and the transfer of data to component registry.

This work and the succeeding work of representing and linking the analyzed data sources into CMDI and storing it in a repository, enables the implementation of new tools in the future. These tools may be similar to the existing tools, but then with the underlying data always being up to date. The tools may then cover all the data instead of the part they were designed for (compare to the GEI-Digital-Visualized tool). The interconnections may lead to new tools, or, to begin with, link one service with another. Linking also enables the derivation of additional textbook attributes. For instance, if we know the knowledge described in curricula and know the approved textbooks for these curricula, then we know the containing knowledge of the textbooks without any digitization effort.

5. Conclusions

In this work, we identified various reasons to join the data behind digital services. We illustrated the challenges, but also the opportunities of harmonizing such data retroactively, having a single point of access, using the Component Metadata Infrastructure (CMDI), which was especially designed for such an undertaking. To show the many advantages of such data repository, we implemented a prototype service, where the joined data could be accessed via search interface. The actual effort to implement this service was minimal, which was very promising for the implementation of future tools and services.

The overall objective of the presented approach is to limit the effort for implementing new research tools, while maximizing re-use of data and code. The desired effects of data unification are conservation, interlinkage and unification of the data access. Additionally, there is the long term effort to unify the underlying code base, so that implementing a new tool could be just a matter of selecting

existing software components. The complete application to join various data silos, as described in this work, goes through the following phases:

1. Recording the characteristics and characteristic values of the research projects.
2. Creating the CMDI profiles.
3. Transfer the data into the component registry.
4. Prototypical implementation of a tool to show that the profiles are complete and correct.
5. Conversion or re-implementation of research projects via component registry.

Realizing described middleware is a work for years. Our institute's middleware is still being developed and CMDI representations only cover the most commonly used features. Individual features, better duplication detection, noting the source for bits of information, etc. have to be added in the future. In the short term, it was not feasible to recreate old projects using the newly created data repository. First, users would not benefit from this change and second, the newly created interconnection between the resources offer much more possibilities than the projects interfaces needed a complete overhaul.

Tools of the Digital Humanities have shown to be successful in supporting research on books. For instance, our institute provides several tools for doing research on textbooks and curricula. Unfortunately, not all institutes have the possibilities and the qualified staff to set up their own Digital Humanities architecture. Hence, in the future, we will enhance our repository to be ready to cover additional data coming from other educational media research projects, from all around the world.

Author Contributions: Software, C.S.; Writing—original draft, F.F. and C.S.; Writing—review and editing, E.W.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McCarty, W. *Humanities Computing*, 2005th ed.; Palgrave Macmillan: London, UK, 2005.
2. Gardiner, E.; Musto, R.G. *The Digital Humanities: A Primer for Students and Scholars*; Cambridge University Press: New York, NY, USA, 2015. [CrossRef]
3. Schultz, K. The Mechanic Muse—What Is Distant Reading? *The New York Times*, 26 June 2020. Available online: <https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html> (access on 19 October 2020).
4. Scheel, C.; De Luca, E.W. Fusing International Textbook Collections for Textbook Research. In *Digital Cultural Heritage*; Springer: Cham, Switzerland, 2020; pp. 99–107. [CrossRef]
5. De Luca, E.W.; Fallucchi, F.; Ligi, A.; Tarquini, M. A Research Toolbox: A Complete Suite for Analysis in Digital Humanities. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2019; pp. 385–397. [CrossRef]
6. Hyvönen, E. Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semant. Web* **2020**, *11*, 1–7. [CrossRef]
7. Goosen, T.; Windhouwer, M.; Ohren, O.; Herold, A.; Eckart, T.; Ďurčo, M.; Schonefeld, O. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In *Proceedings of the CLARIN 2014 Conference*, Soesterberg, The Netherlands, 24–25 October 2014; pp. 36–53.
8. Maik, F.; Christian, S.; Andreas, W.; De Luca, E.W. Welt der Kinder. Semantisches Information Retrieval als Zugang zu Wissensbeständen des 19. Jahrhunderts. In *Proceedings of the Wissenschaftsgeschichte und Digital Humanities in Forschung und Lehre*, Göttingen, Germany, 7–9 April 2016.
9. Fuchs, E.; Kahlert, J.; Sandfuchs, U. *Schulbuch konkret: Kontexte-Produktion-Unterricht*; Klinkhardt: Bad Heilbrunn, Germany, 2010. (In German)

10. Fuchs, E.; Niehaus, I.; Stoletzki, A. *Das Schulbuch in der Forschung: Analysen und Empfehlungen für die Bildungspraxis*; V&R Unipress GmbH: Göttingen, Germany, 2014.
11. Berners-Lee, T. Linked Data. Available online: <https://www.w3.org/DesignIssues/LinkedData.html> (accessed on 19 October 2020).
12. Valsecchi, F.; Abrate, M.; Bacciu, C.; Tesconi, M.; Marchetti, A. Linked Data Maps: Providing a Visual Entry Point for the Exploration of Datasets. IESD@ISWC. 2015. Available online: http://ceur-ws.org/Vol-1472/IESD_2015_paper_2.pdf (accessed on 19 October 2020).
13. Hu, Y.; Janowicz, K.; McKenzie, G.; Sengupta, K.; Hitzler, P. A Linked-Data-Driven and Semantically-Enabled Journal Portal for Scientometrics. In *The Semantic Web—ISWC 2013*; Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 114–129.
14. Ivanova, V.; Lambrix, P.; Lohmann, S.; Pesquita, C. (Eds.) *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data Co-Located with the 15th International Semantic Web Conference, VOILA@ISWC 2016, Kobe, Japan, 17 October 2016; CEUR Workshop Proceedings*. CEUR-WS.org. 2016; Volume 1704. Available online: <http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-131839> (accessed on 20 October 2020).
15. Dadzie, A.S.; Pietriga, E. Visualisation of Linked Data—Reprise. *Semant. Web* **2016**, *8*, 1–21. [CrossRef]
16. Smith, A. Strategies for Building Digitized Collections. *Microform Digit. Rev.* **2008**, *31*. [CrossRef]
17. Dörk, M.; Pietsch, C.; Credico, G. One view is not enough: High-level visualizations of a large cultural collection. *Inf. Des. J.* **2017**, *23*, 39–47. [CrossRef]
18. Fu, B.; Noy, N.; Storey, M.A. Eye tracking the user experience—An evaluation of ontology visualization techniques. *Semant. Web* **2016**, *8*, 23–41. [CrossRef]
19. Lebo, T.; Rio, N.; Fisher, P.; Salisbury, C. A Five-Star Rating Scheme to Assess Application Seamlessness. *Semant. Web J.* **2015**, *8*, 43–63. [CrossRef]
20. Neitmann, S.; Scheel, C. Digitalisierung von (geistes)wissenschaftlichen Arbeitspraktiken im Alltag: Entwicklung und Einführung eines Werkzeugs zur digitalen Annotation. In *Berliner Blätter—Ethnographische und Ethnologische Beiträge: “Digitale Arbeitskulturen: Rahmungen, Effekte, Herausforderungen”*; 2020; accepted.
21. De Luca, E.W.; Scheel, C. Digital Infrastructures for Digital Humanities in International Textbook Research. In *Digital Cultural Heritage*; Springer: Cham, Switzerland, 2020; pp. 85–97. [CrossRef]
22. Braunschweig, U.; Düsterdieck, P.; Hobrecker, U.B.S.; Bernin-Israel, I. Die Sammlung Hobrecker der Universitätsbibliothek Braunschweig. In *Katalog der Kinder- und Jugendliteratur 1565-1945*; Saur: Munich, Germany, 1985. Available online: <https://books.google.it/books?id=RJJ8vQEACAAJ> (accessed on 20 October 2020).
23. Christian, S.; Claudia, S.; De Luca, E.W. Vereinheitlichung internationaler Bibliothekskataloge. In *Proceedings of the Conference on Learning, Knowledge, Data and Analysis—Lernen. Wissen. Daten. Analysen. (LWDA 2016)*, Potsdam, Germany, 12–14 September 2016; Krestel, R., Mottin, D., Müller, E., Eds., Workshop “Information Retrieval 2016” held by the Special Interest Group on Information Retrieval of the Gesellschaft für Informatik (German Computing Society); 2016, pp. 271–282. Available online: <http://ceur-ws.org/Vol-1670/paper-61.pdf> (accessed on 19 October 2020).
24. Hennicke, S.; Stahn, L.L.; De Luca, E.W.; Schwedes, K.; Witt, A. WorldViews: Access to international textbooks for digital humanities researchers. In *Proceedings of the Digital Humanities 2017*, Montréal, QC, Canada, 8–11 August 2017; Conference Abstracts; McGill University & Université de Montréal: Montréal, QC, Canada, 2017; pp. 254–256. Available online: https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6332/file/Hennicke_Stahn_DeLuca_Schwedes_Witt_WorldViews_2017.pdf (accessed on 19 October 2020).
25. Fallucchi, F.; Steffen, H.; De Luca, E.W. Creating CMDI-Profiles for Textbook Resources. In *Metadata and Semantic Research*; Garoufallou, E., Sartori, F., Siatiri, R., Zervas, M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 302–314.
26. Fallucchi, F.; De Luca, E.W. Connecting and Mapping LOD and CMDI Through Knowledge Organization. In *Metadata and Semantic Research*; Garoufallou, E., Sartori, F., Siatiri, R., Zervas, M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 291–301.

27. Fallucchi, F.; Luca, E.W.D. CMDification process for textbook resources. *Int. J. Metadata Semant. Ontol.* **2020**, *14*, 135–148. [CrossRef]


Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Employing a Chatbot for News Dissemination during Crisis: Design, Implementation and Evaluation

Theodora A. Maniou ^{1,*} and Andreas Veglis ² ¹ Department of Social and Political Sciences, University of Cyprus, Nicosia 1678, Cyprus² Media Informatics Lab, School of Journalism and Mass Communications, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece; veglis@jour.auth.gr

* Correspondence: maniou.theodora@ucy.ac.cy

Received: 5 June 2020; Accepted: 26 June 2020; Published: 30 June 2020

Abstract: The use of chatbots in news media platforms, although relatively recent, offers many advantages to journalists and media professionals and, at the same time, facilitates users' interaction with useful and timely information. This study shows the usability of a news chatbot during a crisis situation, employing the 2020 COVID-19 pandemic as a case study. The basic targets of the research are to design and implement a chatbot in a news media platform with a two-fold aim in regard to evaluation: first, the technical effort of creating a functional and robust news chatbot in a crisis situation both from the AI perspective and interoperability with other platforms, which constitutes the novelty of the approach; and second, users' perception regarding the appropriation of this news chatbot as an alternative means of accessing existing information during a crisis situation. The chatbot designed was evaluated in terms of effectively fulfilling the social responsibility function of crisis reporting, to deliver timely and accurate information on the COVID-19 pandemic to a wide audience. In this light, this study shows the advantages of implementing chatbots in news platforms during a crisis situation, when the audience's needs for timely and accurate information rapidly increase.

Keywords: crisis reporting; chatbots; journalists; news media; COVID-19

1. Introduction

A "bot" (a shortening of "robot") is any automated software that responds to incoming information or data, using predetermined rules and/or artificial intelligence to select a response [1]. Chatbots refer to any software application that engages in a dialog with a human, using natural language. The term is most often used in connection with applications that converse via written language. However, with advances in speech recognition [2], they can be programmed to respond differently to certain keywords or employ machine-learning techniques to adapt their responses based on words included in queries [3–5].

For several years, social, conversational bots have been used to provide benefits to companies, who use them to reduce time-to-response and provide enhanced customer service, increasing satisfaction and engagement [6]. Chatbots consist of a specific AI-based software category developed by companies, to automate communications and management of transactions with their customers [7,8]. Use of this technology in news media platforms, although relatively recent, offers many advantages to journalists and media professionals and, at the same time, facilitates users' interaction with useful and timely information. Initially, when newsbots were interacting with users on social media platforms, such as Twitter and Facebook, they tended to mainly function as re-broadcasters of traditional news content to social media [9], namely to alert, aggregate and monitor content for their users. However, news chatbots have moved a long way forward and are now able to offer a variety of informational services to their users, from the dissemination of information to data categorization and taxonomy.

This study aims to show the usability of a news chatbot during a crisis situation and uses the 2020 pandemic crisis of COVID-19 as a case study. Drawing from Piccolo et al. [10], Ford and Hutchinson [11], and Radziwill and Benton's (2017) [6] studies regarding the use of chatbots in crisis situations, the basic targets of this research are to design and implement a chatbot in a news media platform with a two-fold aim in regard to evaluation: first, the technical effort of creating a functional and robust news chatbot in a crisis situation, both from the AI perspective and interoperability with other platforms, which constitutes the novelty of this approach; and, second, users' perception regarding the appropriation of this news chatbot as an alternative means of accessing existing information during a crisis situation. The news chatbot designed was evaluated in terms of effectively fulfilling the social-responsibility function of crisis reporting to deliver timely and accurate information regarding the COVID-19 pandemic to a wide audience. In this light, this study shows the advantages of implementing chatbots in news platforms during a crisis situation, when the audience's needs for timely and accurate information rapidly increase.

2. Literature Review

2.1. Chatbots in the News Media: Uses and Affordances

In recent years, robotic technologies have established themselves in the news media by offering practical alternatives for journalistic daily routines. Although the transition from bots to chatbots has not been easy, automation techniques have found their way to the news media platforms, eventually resulting in "automated news" and "automated journalism" through the use and exploitation of algorithms.

Lokot and Diakopoulos (2015) [9] argue for the "potentially positive and beneficial utility of automated news and information sharing", including how bots may "contribute to positive effects in the public media sphere if employed ethically and conscientiously" (p.3) [12]. This new wave of automation incorporates many forms of what is commonly referred to as "artificial intelligence" or "cognitive technologies", which aid both the inputs and outputs of journalism [13].

Automation initially aimed to reduce human effort and facilitate the work journalists undertake to get news to the public in the era of big data [8]. The application of chatbots in journalism has shown that they can unburden journalists from daily routine work, reducing pressure to produce quantity and allowing them to concentrate instead on quality, freeing up capacity for in-depth analysis and reporting [5]. This also enables them to consider best practice in journalistic work, like checking multiple sources, reflection and diligence [14]. In addition, chatbots embedded in news platforms assist in the categorization and taxonomy of news, especially in the era of big data, facilitating users' effort to locate specific news topics they are interested in [15].

Moreover, the application of chatbots in the news media reshapes the narrative of journalism: Not only do they allow the personalization of information delivery and immediate interaction among sources and recipients [5], but they also do so through trusting speech, by which they seek to generate emotion and foster loyalty [16] and, consequently, trust.

In the news media environment, trust remains one of the most important features that journalists around the world strive to achieve. Several studies have pointed toward trustworthiness that users ascribe to third-party computerized actors and algorithmic processes used for selecting and curating the news. Thurman et al. (2018) [17], for example, demonstrate a link between respondents' trust in news organizations and their assessment of the utility of algorithmic news selection. They argue that, as trust in news organizations falls, people are less likely to recognize editorial selection by journalists as a good way to receive news. By contrast, agreement that automated personalization is a good way to receive news is less affected by distrust in the media. They conclude that users do not recognize the link between automated news personalization and the operation of news organizations, "believing the technology has a degree of immunity from contamination by a politically compromised or untrustworthy news media" (p.17). Ford and Hutchinson (2019) [11] frame their study in terms of

the newsbot as a medium that mediates existing social relationships (in this case, between audiences and media organizations). Following Hoflich (2013) [18], who suggest that robots can be either a connective or divisive element in the inter-group relationships that they mediate, Ford and Hutchinson find that the friendly newsbot contrasts favorably with users' previous experience with traditional dissemination of news and the journalists who produce it. Shneiderman (2020) [19] introduces a two-dimensional framework of Human-Centered Artificial Intelligence (HCAI) that separates levels of automation/autonomy from levels of human control. This model seeks to produce computer applications that are reliable, safe and trustworthy (the three RST goals). Achieving these goals, especially for complex poorly understood problems, can support, among others, mastery, creativity and responsibility; as Shneiderman (2020) [19] argues, well-designed automation preserves human control where appropriate, thereby increasing performance and enabling creative improvements (p.495). In addition, the design of "shared-control systems", where a virtual human model assists the human in a control task—by simultaneously co-controlling, a system could add value for chatbots' users [20]. It is known that if the assisting virtual human model does not accurately characterize human control, this may lead to increased workload for the human operator, instead of improved performance (Griffiths and Gillespie, 2005).

In recent years, chatbots have been connected to "sensor journalism", the use of sensors embedded in the real world as a source of data driving the composition and distribution of automated news items. Examples of this "sensor-journalism" or "sensor-telling" have covered topics such as pollution and animal welfare [17,21]. Several studies in this area are mainly evidential in (a) exploring whether and why news consumers think automated personalization is a better way to receive news than selection by journalists and editors [17]; (b) questioning received wisdom on the existence of filter bubbles [22]; and (c) even asking whether recommendation engines might promote, rather than limit, diverse news exposure [23].

However, recent developments regarding the use of news chatbots have brought significant implications for individuals, society and communication, resulting in the formation of Human–Machine Communication (HMC) as an area of research [24,25], as did previous changes in technology that were followed by corresponding shifts in communication research. While the underlying theoretical assumption in the majority of communication research is that it is humans who are communicators and machines that function as mediators, within HMC, this assumption is challenged by asking what happens when a machine steps into this formerly human role. Lewis et al. (2019) [26] argue that HMC has to be re-envisaged as a way of approaching the study of technology based on its function as communicator (message source), rather than merely as mediator (message channel).

Moreover, the way(s) news chatbots are used can raise issues regarding data ownership and privacy. In their study regarding the news chatbot created by the Australian Broadcasting Corporation (ABC), Ford and Hutchinson (2019) [11] argue that the way the ABC operated its news chatbot could have consequences for Public Service Media (PSM) accountability principles, since the ABC was dependent on the private infrastructure of Facebook and ChatFuel. They state that such matters will only become more relevant as news is delivered increasingly via smart speakers and controlled by the few global corporations with the expertise, data and resources necessary for leading such developments.

2.2. *Employing News Chatbots for Crisis Reporting: The Case of the COVID-19 Pandemic*

Until the early days of 2020, the global societies of the Western world could conceive of a global pandemic scenario only within the framework of a Hollywood movie. The spread of COVID-19 found mankind unprepared to deal with emergency health crisis situations, although, in the previous decade, several countries around the world had confronted lesser health crises, such as the SARS (Severe Acute Respiratory Syndrome) pandemic in the early 2000s. The COVID-19 pandemic made clear to the whole world that diseases do not recognize national borders, whereas national health systems—even in the most developed countries—are unable to support an effective response to such crises.

In this framework, crisis reporting has become more relevant than ever and in need of more robust and effective practices to fulfil its social-responsibility function: accurate, timely and precise information available to all citizens on practical issues in regard to the global pandemic. As Kruger (2005) argues [27], reporting information “as fully as its standing as a major public health crisis demands, and reporting it in all its various aspects” (p.127). However, crisis reporting regarding health issues in the past had to meet these demands in a way that was in accordance with the technological tools available. The technological explosion of Web 3.0 in recent years seems to have weaponized media organizations with the ability to fulfil their social role with a variety of tools. Among these, chatbots offer a practical solution for both audiences and media professionals to deal with a severe global health crisis, provided they are used to fulfil an audience’s specific needs and demands.

In their study regarding the use of chatbots in crisis situations, Piccolo et al. (2018) [10] find that users expect an ideal chatbot to be high performing (fast, efficient and reliable), smart (knowledgeable and accurate in predictions), seamless (easy and smooth) and personable (understands me and is likeable). Their model was based on Radziwill and Benton’s study (2017) [6], which compiles a set of quality attributes expected for chatbots classified into six categories: (1) performance, reflecting the ability to deal with unexpected input and the appropriate escalation to humans; (2) functionality, which is related to linguistic accuracy; (3) humanity, referring to humanized interactions; (4) affect, which encompasses enjoyability, politeness and personality traits; (5) ethics and behavior, referring to respecting users’ privacy, sensitivity to social concerns and trustworthiness; and (6) accessibility, by detecting users’ meaning or intent and responding to social cues.

Dale (2016) [2] states that, as the technology evolves, in the near future, it could become hard to distinguish a chatbot from a human in a conversation. Louwerse et al. (2005) [28] argue that having “humanlike” behavior can be one of the success criteria in the process of evaluating chatbots. However, in the case of news chatbots, this clear distinction is important, so as to avoid frustration. Design plays a significant role, and, in most cases, it is up to the designer to decide the “form” of the chatbot. For example, creating a chatbot by using a “humanlike” form can make the application more “user-friendly”; however, it can also change user perception of it as “human or machine”. Overall, embedding human elements in a news chatbot can be identified as an important step toward building trust in the chatbot application and in the sociotechnical initiative as a whole [10].

Looking for further success criteria to evaluate chatbots, Zamora (2017) [29] goes back to the virtual agents’ literature from the 1990s, which includes being efficient in responsiveness. To this end, an effective chatbot relies not only on the user interface design but also on the development of a robust AI-based process to support the conversation. The design, though, has to properly address eventual conversation breakdowns [10,30].

Achieving effective interaction between users and news chatbots is an intriguing challenge, depending not only on design but also on users’ preferences and interests. More than that, as pointed out by Zamora (2017) [29], the best interaction mechanism should be chosen according to the context; for instance, replacing text-based input, which is more susceptible to errors, with making selections whenever possible (see also Reference [10]). To create a truly inclusive experience, the design should guide the users throughout the navigation, step by step, facilitating their search and news selection process.

Added to all of this, content plays a significant role in the effectiveness of news chatbots. As Roberts and Doyle (2017) [31] point out, crisis-responding organizations are encouraged to build relationships with the local population around questions of data interoperability, data sharing and understanding how their policies might inhibit or affect data sharing and collection. In regard to the use of chatbots for crisis reporting, users need to be able to have quick access to specific information, as well as access to distinct and clear categories of news. Depending on the type of crisis, this information may vary from instructions for access to hospital and medical care facilities, to road safety instructions and/or emergency contact details of relevant authorities.

3. Research Questions (RQs) and Method

Based on the preceding theoretical analysis, the study sought to address the following research questions:

RQ1. Which functions should news chatbots perform when used in crisis reporting?

RQ2. How can news chatbots be designed and implemented to be trustworthy, reliable and acting in the users' interests when used in crisis reporting?

Based on the ideal features of a chatbot, as earlier analyzed, a news chatbot was created based on the COVID-19 information offered on the web platform of the BBC. The study opted for the implementation of a retrieval based chatbot with predetermined responses based on specific requirements included in this study. Specifically, the scope of the effort was to develop and evaluate a news chatbot that would offer an alternative method of accessing existing information. Moreover, since such a solution is being proposed for crisis situations, a specific one was selected that could be deployed rapidly and does not require sophisticated programming. It is worth noting that chatbot implementations based on natural language interaction can be considered to be overrated for disseminating existing structured (in some sense) information (e.g., symptoms, existing cures, available medication, restriction in movement, etc.).

The specific news organization was selected first because it attracts a global audience through using the English language, and second because it presents not only national information regarding the UK but also global information on the rest of the world, in contrast to other national or local news organizations in Europe that tend to focus more on national/local aspects of the COVID-19 pandemic. This makes it easier for study participants to assess and evaluate the effectiveness of the news chatbot presented, as they are more familiar with the available information.

Once the news chatbot was created, two groups of journalism students (consisting of 45 participants each) were asked to evaluate its performance, both via mobile and computer screens. The questionnaire used for the focus groups' evaluation is presented in Appendix A. The two focus groups were independent and did not interact with each other. The first group was selected from 2nd-year students attending the BA Program of Journalism in the University of Cyprus; the second group was selected from 2nd-year students attending the BA Program of Journalism in the Aristotle University of Thessaloniki, Greece. All participants in both groups were familiar with news chatbot applications, having earlier attended relevant teaching modules. The ratio of men to women was around 1:1, and their ages ranged from 20 to 24 years old. They were deemed appropriate to evaluate the news chatbot mainly due to their familiarization with relevant applications. After all, this study is not concerned with the overall evaluation of using news chatbots; instead, the main target, as already stated, is to design an effective news chatbot to be used during a crisis situation.

Both focus groups were conducted by using the Microsoft Teams Platform for distance learning courses by the same independent researcher/moderator who guided the interactive conversation. The moderator was accompanied by each group's instructor, who remained only as an observer throughout the online conversation. The initial questions were based on the specific features of effective news chatbots, as earlier analyzed in the Theoretical Framework. In the next stage of the study, all evaluation comments were categorized regarding functionality, reliability, design and specific features of the news chatbot (see analysis below) and embedded in the final application.

4. The COVINFO Reporter Chatbot: Design and Implementation

The COVINFO Reporter chatbot was developed over a period of two months (March and April 2020) and was carried out in three steps, namely *design*, *implementation* and *evaluation* of the platform. In the first step, which involved platform design, the authors employed existing experience of chatbot development [8], as well as previous related implementations [10,12]. Figure 1 depicts the actions taken during the development of the chatbot. In step 1, the chatbot's requirements and features were articulated. The second step included the deployment of the COVINFO Reporter, with all the features specified in the first step. The supported conversation was optimized for simplicity of use, aiming at providing all the information requested by users, with minimal conversation steps. During

the implementation step, the COVINFO Reporter was evaluated by a small number of experts who completed a series of actions on the chatbot, acting as potential users obtaining information related to COVID-19. Finally, it was extensively used and evaluated by journalism students from Greece and Cyprus. Based on their comments, improvements were introduced, and possible research directions were identified. The evaluation of the chatbot is presented and discussed in the next section.

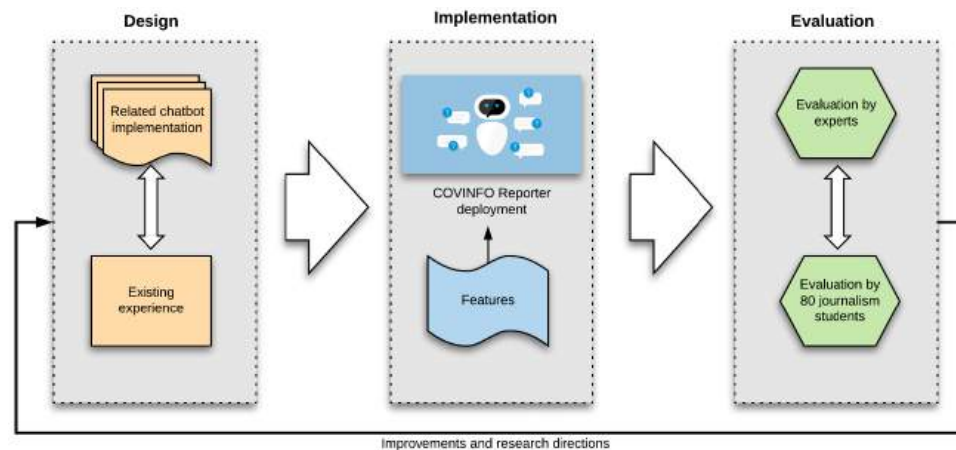


Figure 1. The three steps in the development of the COVINFO Reporter chatbot.

4.1. Design

As mentioned earlier, the design of the COVINFO Reporter chatbot was based on previous experience [8] in which a chatbot based on the Quriobot plugin was embedded in a news article published on a WordPress platform, as a means of alternative narrative. Although the focus of the current project was on facilitating access to multiple news articles, the main idea was the same in both cases: the deployment of an alternative means of accessing existing information. Of course, one must also consider the urgency of the situation of the COVID-19 pandemic, which forces users to seek important information on how to deal with the crisis, which affects all aspects of their daily lives.

For the process of selecting the appropriate chatbot platform, certain parameters were taken into account. Specifically, the study prompted for a simple chatbot platform that can be embedded in a WordPress Content Management System (CMS), since the majority of the news websites is based on this CMS. Moreover, for future extensions of this effort, the possibility of integrating the chatbot into other social media platforms needed to be supported by the selected chatbot platform. Based on the above, the ManyChat chatbot platform (built primarily of Facebook messenger) was excluded. Furthermore, other established chatbot platforms, like Amazon Lex and Google Dialogflow, do not offer their own plugins for WordPress (some related plugging is offered by third parties). In addition, those platforms are built for natural language interaction. Thus, the Quriobot platform was selected based on the facts that it supports WordPress integration and, also, other social media platforms (Facebook Messenger, Viber, Slack, WhatsApp, Snapchat, etc.) for future extensions and its primary focus is on supporting predetermined chat flows.

The definition of the requirements during the design step was supported by previous chatbot implementations. Piccolo et al. (2018) [10] consider the deployment of a chatbot on Facebook Messenger to help people submit reports of violence and misconduct during the 2017 Kenyan elections. During the design phase of the COVINFO Reporter chatbot, this option was examined but rejected as being limited only to users of one messaging platform. However, this functionality can be deployed at a later stage of the chatbot development, since the chosen implementation platform supports a range of messaging platforms. The BBC has also piloted several conversational chatbots based on pre-scripted material written in a multiple-choice format, mainly focused on special events, e.g., the UK's general elections [12,32].

Since the scope of this study was to develop a chatbot that would offer an alternative method of accessing existing information, a retrieval-based chatbot with predetermined responses was selected. The BBC website was chosen as the source of the chatbot information, since, among other characteristics given in the RQs and Method section, it has developed a significant number of web articles (text and video) that examine various aspects of the COVID-19 pandemic [33].

4.2. Implementation

The COVINFO Reporter chatbot aims to enhance journalists' day-to-day workflows. The platform supports the selection by journalists of existing web content from the BBC website, which is made available through the COVINFO Reporter interface and is accessed by users seeking specific information through an alternative narrative. Figure 2 depicts the interactions that take place with involved stakeholders (journalists and users). Journalists create content on the BBC website and also select the articles that can be made available through the COVINFO Reporter chatbot. On the other hand, users employ the COVINFO Reporter interface to access available information.

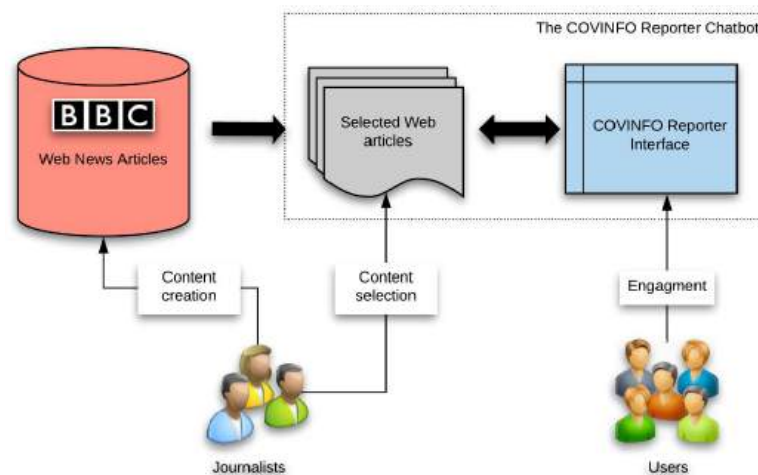


Figure 2. User and journalist interactions with the COVINFO Reporter chatbot.

The Quriobot platform (<https://quriobot.com>) was selected to deploy the COVINFO Reporter chatbot. The Quriobot is relatively simple to install and use, with development managed via its Control Room (<https://control.quriobot.com/>). The chatbot is also able to gather information through interaction with users and thus improve interactivity. Although the COVINFO Reporter chatbot was employed through the URL offered by the Quriobot platform, it can also be embedded in a WordPress website, as well as other channels (e.g., Facebook Messenger). The COVINFO Reporter chatbot can be accessed at <https://botsrv.com/qb/AUTH/COVINFO-CHATBOT>.

The creation of a chatbot is a relatively straightforward process, since the Quriobot platform offers a number of ready-to-use templates that can be adapted to each developer's needs. It also offers the ability to construct a chatbot from scratch, which consists of a number of steps that define the Quriobot's behavior. The Quriobot Control Room conversational designer supports the build of proactive conversations, which are supported by conditional rules and different step types. It can also employ smart jumps between questions, based on answers provided by users.

In the case of the COVINFO Reporter chatbot, 11 steps were programmed, as depicted in the flowchart in Figure 3. Initially, there is a welcome message, and the chatbot introduces itself. Then it asks the user if he/she wants to continue. If the answer is "No", the chatbot terminates. If the reply is "Yes", the chatbot presents a list of available topic categories, and then, upon selection, the selected category is displayed. Next, the user can select and display a selected news article. Then it asks if the user requires access to other information. If the answer is "Yes", the chatbot again displays the categories. If the answer is "No", the conversation is terminated. It is worth noting that all the

questions the chatbot is called to answer have predefined answers. Various steps of the conversation with the COVINFO Reporter chatbot are included in Figure 4.

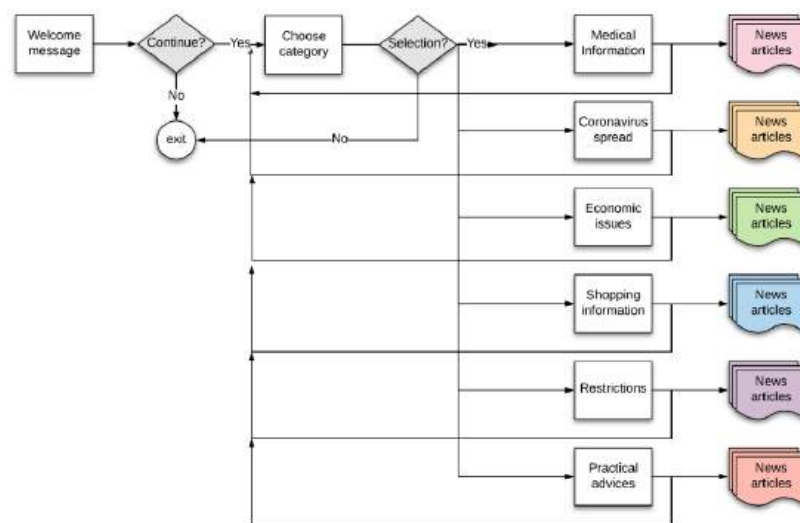


Figure 3. The flow diagram of the COVINFO Reporter chatbot.

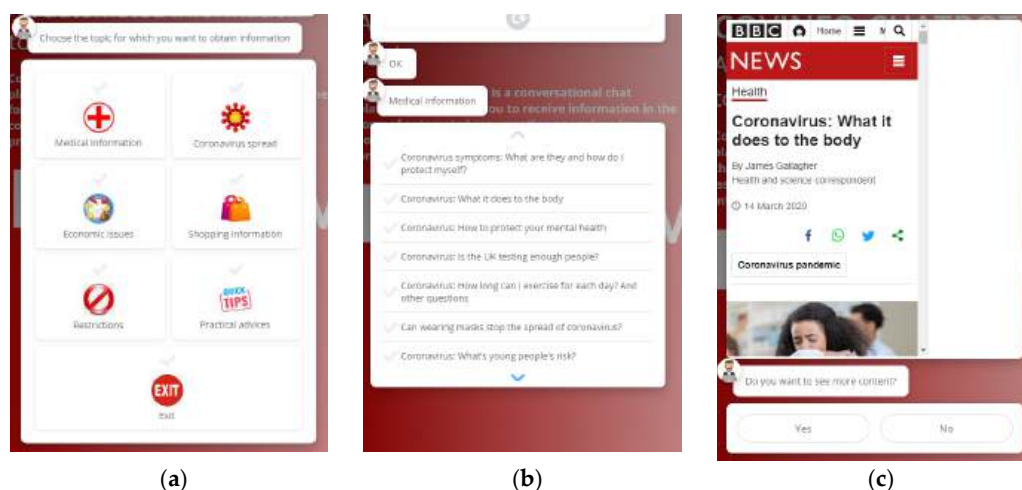


Figure 4. Screenshots of contents (a), available discussion topics (b) and articles (c) in the COVINFO Reporter chatbot.

The Quriobot platform supports ongoing modification and update of chatbots by adding more categories or available articles. It can be accessed from both PCs and smartphones. However, one limitation of this feature is that it cannot be programmed to display differently on different device types. Thus, news articles are displayed in a small-size window appropriate to smartphones, which is, as expected, not very convenient for PC users.

5. Evaluation of COVINFO Reporter Chatbot: Findings and Analysis

Drawing from earlier chatbot evaluation studies [6,10,11], the COVINFO Reporter chatbot was assessed in terms of the following characteristics: (a) performance, (b) reliability, (c) functionality, (d) personalization, (e) interactivity, (f) ethics and behavior, and (g) accessibility. Although the two groups of participants did not interact with each other, both came to the same conclusions and offered similar assessments; therefore, their comments are jointly presented based on the earlier mentioned categorization of the chatbot's characteristics and not on the basis of two separate focus groups.

The *performance* of the COVINFO Reporter chatbot was examined in regard to its ability to respond in a timely and efficient manner, both via a larger screen (tablet or laptop) and a smaller one (mobile phone). All participants deemed it efficient, and most indicated that, in times of crisis, the chatbot can save time when looking for crucial information: “It helped me save time while looking for much needed information, for example guidelines on medical facilities”, stated M.A. (female, 20 years old, Cyprus). The vast majority of participants stated that there were no differences between the two screen categories used. However, they found the mobile phone screen to be more efficient, perhaps because they were more familiar with it: “All the information is there; either you use it via a computer or via a smartphone, but my opinion is that COVINFO Reporter was constructed for a smartphone” (C.G., female, 22 years old, Cyprus). Several problems were initially detected by participants of both focus groups, focusing mainly on the chatbot’s technical performance. “I detected some ‘technical problems’ in regard to the chatbot’s ability to return to the main screen that need to be fixed; otherwise it would be tiring to navigate through it” (A.L., male, 22 years old, Cyprus); “When using it through a computer, it does not seem to be able to use the full-screen mode, only part of the screen” (I.A., male, 23 years old, Cyprus); “I agree, this could entail further problems with people with visual disabilities and older users” (N.C., female, 21 years old, Cyprus).

Functionality was tested in terms of linguistic accuracy and knowledgeable information offered. Most of the participants agreed that the language used is simple and accurate, and the information offered is useful for everybody during the pandemic crisis: “I found everything I was looking for—the information was filtered in a useful way and the news categories were expressed with simple words” (A.L., male, 23 years old, Greece). Although some participants indicated that they enjoyed the fact that the COVINFO Reporter offers the basic information needed, others argued that more information could be added: “At least another information category has to be added regarding information for the pandemic outside the UK, for people who want to know what is happening in other countries” (G.A., male, 20 years old, Greece; A.E., 21 years, female, Cyprus). An important point was made by several participants regarding the colors used: “the use of red color could eventually make us tired while using the chatbot for [a] longer duration of time” (A.B., female, 22 years old, Greece); “the partial use of white color should be avoided in a news chatbot because it can make users feel bored” (G.A., male, 20 years old, Greece); “definitely the use of white color for the fonts should be avoided in the chatbot responses; it makes it difficult to read the information offered” (G.K., female, 21 years old, Greece).

Reliability was measured in regard both to the content offered and to the chatbot’s proper function. The majority of the participants indicated that the chatbot functioned properly, and the timeframe for providing answers to users was adequately calculated. “It functions properly in all categories tested, and this kept me going for longer than I thought I would have stayed inside the application, definitely better than reading a conventional news website” (S.Ma., male, 21 years old, Greece). Regarding the news content, all participants deemed it reliable and trustworthy, as “it is based on the information offered by the BBC web page; therefore, I consider it reliable enough” (P.P., male, 23 years old, Cyprus); “I trust that the information is reliable, because it comes from a reliable source, since it is the source that guarantees reliability, not the robot application” (C.Ch., female, 22 years old, Cyprus).

Personalization refers to the form and depiction of the chatbot. All participants found the selected form as being “representative”, “reliable” and “affective”. To this end, several students emphasized the fact that the picture selected for the COVINFO Reporter chatbot “encompasses politeness and personality traits that can make it seem more human” (E.A., female, 21 years old, Cyprus), “it looks like a real reporter and uses vivid color, which is very pleasant” (C.C., female, 20 years old, Cyprus), whereas it offers “crucial information in a customized form, and this is important in a crisis situation, for all users” (N.K.Th., female, 22 years old, Greece). All participants argued that the depiction of the chatbot in a humanlike form was not confusing: “it is clear that this is a robot application, although it is depicted in a humanlike form” (N.P., male, 21 years old, Cyprus); “the humanlike picture selected cannot confuse the users; this is clearly a robot we are interacting with, although he looks really friendly and polite, exactly as a reporter should look” (G.I., male, 22 years old, Cyprus).

Interactivity was tested in regard to the chatbot's ability to easily interact with users. "It was fun and enjoyable to read the news in this format; it really helped me to move on with my next questions", stated D.Th. (male, 22 years old, Greece). All participants agreed that this was a more enjoyable way to acquire the information they were looking for regarding the pandemic crisis than the way information is offered in a conventional news site: "Even if I am in a hurry and looking for specific information quickly, this way is far more effective, because it is like the chatbot is trying to answer all my questions" (M.A., male, 20 years old, Cyprus).

Ethics and behavior refer to issues regarding users' privacy and sensitivity toward social concerns. All participants found it positive that no personal data were needed, whereas the information offered was in line with social concerns around an issue as serious as a health pandemic. "I liked the fact that there was no need to state personal data, i.e., my e-mail; it made it easier for me to search the specific information I wanted; for example, if someone I infected looks for medical information, he/she do not need to identify themselves", stated I.A. (male, 23 years old, Cyprus). "The chatbot provides information for every user of any age, and for me, this indicates social responsibility for all citizens", argued Ch.K. (female, 21 years old, Greece).

Accessibility was tested in terms of users' ability to easily access the chatbot and navigate through it. All participants in the study agreed that it was easily accessible and "fun to navigate through it, much more than any conventional news website (M.Z., female, 21 years old, Greece). "It was really easy to access and navigate inside the chatbot; everyone can do it, even older users with limited knowledge regarding chatbots; in fact, the chatbot itself guides you to the information you are looking for" (A.S., female, 21 years old, Greece). "The Q and A process escalates smoothly, and in this way, it is easy to access the specific information you are looking for; it is an easier narrative for telling the news" (P.T., male, 23 years old, Greece). "It is a better narrative when you are looking for emergency information, easier to access and navigate through it" (M.P., female, 22 years, Cyprus).

All expected characteristics, as well as achieved results, are depicted in Table 1.

Table 1. Expected characteristics and achieved results.

Characteristics	Achieved Results
Performance	<ul style="list-style-type: none"> • Ability to respond in a timely and efficient manner • Ability to adjust both to a larger screen (tablet or laptop) and a smaller one (mobile phone)
Functionality	<ul style="list-style-type: none"> • Linguistic accuracy • Knowledgeable information • Simple language • Use of bright colors for the fonts
Reliability	<ul style="list-style-type: none"> • Proper technical function • Credible information • Identification of information source(s) • Proper timeframe for providing answers
Personalization	<ul style="list-style-type: none"> • Humanlike picture • Selection of personal characteristics that adhere to a typical reporter
Interactivity	<ul style="list-style-type: none"> • Ability to easily interact with users
Ethics and Behavior	<ul style="list-style-type: none"> • Respect of users' privacy • No personal data shared • Information in line with social concerns regarding the pandemic
Accessibility	<ul style="list-style-type: none"> • Easy navigation • Easy access • Smooth escalation of Q and A process

As this analysis has shown, chatbots used for news dissemination in a crisis situation seem to present certain differences in comparison to commercial chatbots. First, according to users' assessment, they need to be as simple as possible so as any user can access and navigate through the information offered. This is extremely important during a crisis situation, since users need to acquire information fast and easily. In this light, the chatbot's technical ability to respond in a timely manner is of equal importance during a crisis situation. The second basic difference is related to ethics and behavior. In a crisis situation, social concerns and "sensitive" issues may be related to patients' identification and personal data publicized. As such, news chatbots used for access to emergency information (e.g., nearby medical facilities, guidance according to medical protocols, etc.) need to be in line with social concerns and ethical boundaries.

This analysis has also shown that the design and development of chatbots used for news dissemination in a crisis situation is rapidly evolving, following two basic factors: first, the latest technological trends, as well as the available technology, both to the developer and to the target audience. For example, while an international news organization can have access to the means and personnel needed to develop a more perplexed chatbot application, a local news entity does not have the means, nor does it employ the specialized personnel, to develop perplexed applications. In this light, news chatbots that are developed to meet urgent needs and audience demands need to be easily designed and managed, following existing development tools. Accordingly, every developer needs to keep in mind the technology available to the target audience. For example, users in countries of the Western world tend to enjoy more advanced technological tools than users in underdeveloped countries, following the existing digital divide.

Second, news chatbots need to be in line with the specific peculiarities of the crisis situation for which they provide information. Not every crisis situation presents similar characteristics to previous crises, even if they are related to the same social sectors. For example, the Great Recession of 2007 was radically different from previous economic crises mankind had to face. Accordingly, the pandemic crisis of 2020 due to COVID-19 was different compared to the SARS (Severe Acute Respiratory Syndrome) pandemic in the early 2000s. As such, users' needs regarding information and news dissemination may differ, and this has to be taken into consideration during the development of the application.

6. Conclusions

This paper has focused on the design, implementation and evaluation, in terms of effectively fulfilling the social responsibility function of crisis reporting, of a news chatbot used in a crisis situation. In this light, the pandemic crisis of 2020 due to COVID-19 was used as a case study, and the COVINFO Reporter chatbot was developed, which aims to deliver timely and accurate information regarding the crisis. The novelty of the approach is based on the news chatbot's easy implementation for news organizations, as well as on its ability to effectively deliver crucial information to a wide audience (users) in times of crisis.

Interesting conclusions can be drawn from the findings of the study. There is no doubt that automation is already having a significant impact on journalism and the dissemination of news in general. The introduction of chatbots in the media sector has shown that they can significantly reduce journalists' workload, allowing them to concentrate on quality, in-depth analysis and reporting [14]. Chatbots can facilitate an alternative narrative that can be customized based on users' preferences. This is significant in the cases of crisis reporting, where the dissemination of accurate, timely and customized information is very important for the public.

The theoretical study of previous media chatbot projects informed the implementation of the COVINFO Reporter, a working chatbot that disseminates information published by an international media organization. The chatbot was developed on a commercially available chatbot platform (i.e., Quriobot) and can be easily customized and updated. It offers easy and predictive navigation, enabling users to access the information that interests them, without having to navigate through the significant

number of webpages that a media organization site usually includes. Its programming is relatively straightforward and can be easily integrated into the workflow of a typical media organization.

A thorough evaluation of various characteristics (performance, reliability, functionality, personalization, interactivity, ethics and behavior and accessibility) of the COVINFO Reporter chatbot was conducted by two separate groups of participants. The chatbot was positively evaluated in terms of its efficiency, although some participants reported minor technical problems. The preferred platform for accessing it was mobile phones. The majority of the participants was satisfied with the functionality of the chatbot, reporting that the language used was simple and accurate, and the information it provided was useful. The participants agreed that the chatbot was reliable, was functioning properly and provided answers in an acceptable time frame. As far as personalization is concerned, the COVINFO Reporter was reported to be representative, reliable and affective. All participants appreciated the chatbot's interactivity. No problems were reported in terms of users' privacy and sensitivity toward social concerns. Finally, all participants agreed that the COVINFO Reporter's accessibility was very good, and they experienced no problems in navigating the chatbot. Overall, the evaluation of the chatbot was very positive, and the minor problems that were detected were noted and corrected, thus improving its performance.

Future extensions of this work could include additional research into the ways in which chatbots can be employed in crisis reporting, with a focus on their smooth incorporation in the journalistic workflow, with added features and inputs (textual and voice), so as to further assist users looking for specific information. Special attention should be given to chatbots' ability to collect data from users, thus enabling them to be utilized in crowdsourcing schemes, which can be extremely valuable during crisis situations. However, in this case, the use of a method to prevent the spread of "fake news" (misinformation) would be necessary so as to ensure the reliability of the chatbot and the information disseminated. There exist a variety of techniques to monitor complex systems, in which the average behavior of the whole network is compared to particular nodes. One example is the use of deep learning to detect faults in systems through entropy measurement, as proposed by Martinez-Garcia et al. (2019) [34].

Finally, since this implementation is proposed as an interaction with the general public, interaction in natural language was not the first choice. Nevertheless, the incorporation of a news chatbot that would support natural language interaction is considered to be one of the future extensions of this study.

Author Contributions: Conceptualization, T.A.M. and A.V.; methodology, T.A.M. and A.V.; validation, T.A.M. and A.V.; formal analysis, T.A.M. and A.V.; investigation, T.A.M. and A.V.; data curation, T.A.M. and A.V.; writing—original draft preparation, T.A.M. and A.V.; writing—review and editing, T.A.M. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Initial Questionnaire used for the Focus Groups

1. Was the COVINFO Reporter helpful for finding the information you were looking for? In which specific ways?
2. In which points/steps during your search did you find difficulties?
3. Do you feel that there will be groups of users that could find difficulties in using COVINFO Reporter? If yes, please describe these users.
4. How difficult was for you to use it?
5. In your personal opinion, through which device COVINFO Reporter was more appropriate and easy to use? Which device did you use to access it?
6. In your personal opinion, do you believe that COVINFO Reporter should offer more or less information to users?

7. In your personal opinion, is the information provided through COVINFO Reporter reliable? Is it timely or is it outdated? Did you trace information that may seem as ‘fake news’? If yes, please provide specific examples.
8. Based on the information provided by COVINFO Reporter, can you describe the person that manages this chatbot?
9. Would you like to see similar chatbots in the news sites you usually use for your information regarding COVID-19 news and related stories/events?
10. In your opinion, is the information provided by chatbots reliable?

References

1. Mann, C.B. Can conversing with a computer increase Turnout? Mobilization using chatbot communication. *J. Exp. Political Sci.* **2020**, *1*–12. [CrossRef]
2. Dale, R. The return of the chatbots. *Nat. Lang. Eng.* **2016**, *22*, 811–817. [CrossRef]
3. Janarthanam, S. *Hands-On Chatbots and Conversational UI Development: Build Chatbots and Voice User Interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*; Packt Publishing Ltd.: Birmingham, UK, 2017.
4. Shevat, A. *Designing Bots: Creating Conversational Experiences*; O'Reilly Media Inc.: Sebastopol, CA, USA, 2017.
5. Veglis, A.; Maniou, T.A. Chatbots on the rise: A new narrative in Journalism. *Stud. Media Commun.* **2019**, *7*, 1–6. [CrossRef]
6. Radziwill, N.M.; Benton, M.C. Evaluating quality of chatbots and intelligent conversational agents. *arXiv* **2017**, arXiv:1704.04579.
7. Androutsopoulou, A.; Karacapilidis, N.; Loukis, E.; Charalabidis, Y. Transforming the communication between citizens and government through AI-guided chatbots. *Gov. Inf. Q.* **2019**, *36*, 358–367. [CrossRef]
8. Veglis, A.; Maniou, T.A. Embedding a chatbot in a news article: Design and implementation. In Proceedings of the ACM 23rd Pan-Hellenic Conference on Informatics, Nicosia, Cyprus, 28–30 November 2019; pp. 169–172.
9. Lokot, T.; Diakopoulos, N. News bots: Automating news and information dissemination on Twitter. *Digit. J.* **2015**, *4*, 682–699. [CrossRef]
10. Piccolo, L.S.G.; Roberts, S.; Iosif, A.; Harith, A. Designing chatbots for crises: A case study contrasting potential and reality. In Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI), Belfast, UK, 4–6 July 2018.
11. Ford, H.; Hutchinson, J. Newsbots that mediate journalist and audience relationships. *Digit. J.* **2019**, *7*, 1013–1031. [CrossRef]
12. Jones, B.; Jones, R. Public service chatbots: Automating conversation with BBC News. *Digit. J.* **2019**, *7*, 1032–1053. [CrossRef]
13. Marconi, F.; Siegman, A. A Day in the Life of a Journalist in 2027: Reporting Meets AI. *Columbia Journalism Review*. 2017. Available online: <https://www.cjr.org/innovations/artificial-intelligencejournalism.php> (accessed on 9 April 2020).
14. Dörr, K.N.; Hollnbuchner, K. Ethical challenges of algorithmic journalism. *Digit. J.* **2017**, *5*, 404–419. [CrossRef]
15. Veglis, A.; Maniou, T.A. The mediated data model of communication flow: From data journalism to big data. *Kome Int. J. Pure Commun. Inq.* **2018**, *6*, 32–43. [CrossRef]
16. Sánchez-Gonzales, H.; Sánchez-González, M. Bots as a news service and its emotional connections to audiences: The case of Politibot. *Doxa Comun. Rev. Interdiscip. Estud. Comun. Y Cienc. Soc.* **2017**, *25*, 63–84.
17. Thurman, N.; Moeller, J.; Helberger, N.; Trilling, D. My friends, editors, algorithms, and I. *Digit. J.* **2018**, *7*, 447–469. [CrossRef]
18. Hoflich, J.R. Relationships to social robots: Towards a triadic analysis of media-oriented behavior. *Intervalla Platf. Intellect. Exch.* **2013**, *1*, 35.
19. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 495–504. [CrossRef]
20. Martínez-García, M.; Zhang, Y.; Gordon, T. Memory pattern identification for feedback tracking control in human–machine systems. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2019**. [CrossRef]

21. Vicari, J.; Weiss, B. Sensor driven journalism: Combining reporting with the internet of things. Presented at the Algorithms, Automation, and News Conference, Munich, Germany, 22–23 May 2018.
22. Zuiderveen Borgesius, F.J.; Trilling, D.; Möller, J.; Bodó, B.; de Vreese, C.H.; Helberger, N. Should we worry about filter bubbles? *Internet Policy Rev.* **2016**, *5*. [CrossRef]
23. Helberger, N.; Karppinen, K.; D’Acunto, L. Exposure diversity as a design principle for recommender systems. *Inf. Commun. Soc.* **2018**, *21*, 191–207. [CrossRef]
24. Guzman, A.L. (Ed.) *Human–Machine Communication: Rethinking Communication, Technology, and Ourselves*; Peter Lang: New York, NY, USA, 2018.
25. Spence, P.R. Searching for questions, original thoughts, or advancing theory: Human-machine communication. *Comput. Hum. Behav.* **2019**, *90*, 285–287. [CrossRef]
26. Lewis, S.C.; Guzman, A.L.; Schmidt, T.R. Automation, journalism, and human–machine communication: Rethinking roles and relationships of humans and machines in news. *Digit. J.* **2019**, *7*, 409–427. [CrossRef]
27. Krüger, F. Ethical journalism in a time of AIDS. *Afr. J. AIDS Res.* **2005**, *4*, 125–133. [CrossRef]
28. Louwerse, M.; Graesser, A.; Lu, S.; Mitchell, H. Social cues in animated conversational agents. *Appl. Cogn. Psychol.* **2005**, *19*, 693–704. [CrossRef]
29. Zamora, J. I’m sorry, Dave, I’m afraid I can’t do that: Chatbot perception and expectations. In Proceedings of the ACM 5th International Conference on Human Agent Interaction, 2017, HAI ’17, New York, NY, USA, 17–20 October 2017; pp. 253–260.
30. Folstad, A.; Brandtzaeg, P.B. Chatbots and the new world of HCI. *Interactions* **2017**, *24*, 38–42. [CrossRef]
31. Roberts, S.; Doyle, T. Understanding crowdsourcing and volunteer engagement. In *Flood Damage Survey and Assessment: New Insights from Research and Practice*; Mollinari, D., Ed.; John Wiley and Sons Inc.: London, UK, 2017; pp. 121–134.
32. BBC News Labs. Scripting Chatbots Is Hard. Here’s How We Made It Easier for BBC Journalists: In Our Toolkit: BBC News BotBuilder. *Medium*. 2018. Available online: <https://medium.com/bbc-news-labs/bbc-botbuilder-ba8e09b6a2e9> (accessed on 20 April 2020).
33. Cushion, S.; Sambrook, R. Coronavirus: BBC News Is Uniquely Placed to Serve the Nation—How It Does so Will Define Its Future. *The Conversation*. 2020. Available online: <https://theconversation.com/coronavirus-bbc-news-is-uniquely-placed-to-serve-the-nation-how-it-does-so-will-define-its-future-135265> (accessed on 16 April 2020).
34. Martínez-García, M.; Zhang, Y.; Suzuki, K.; Zhang, Y. Measuring system entropy with a deep recurrent neural network model. In Proceedings of the IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 22–25 July 2019; pp. 1253–1256.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Future Internet Editorial Office
E-mail: futureinternet@mdpi.com
www.mdpi.com/journal/futureinternet



MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com



ISBN 978-3-0365-4315-4