

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Internationalization of China's E-Commerce Higher Education: A Review between 2001 and 2019

*Wenying Huo, Mingxuan Wu and Jeffrey Soar*

## Abstract

The purpose of this chapter is to review the development of China's higher education in electronic commerce (e-commerce) and explore the requirements of the internationalization of China's e-commerce higher education. The Benefit-Driving Model (BDM) was adopted to explain the reasons for the internationalization of China's e-commerce higher education. The literature review spans 20 years from 2001 when the first 13 e-commerce programs were offered from China's 597 universities. By 2019, 328 e-commerce programs were offered by 831 universities. There is a sustainable growth from 2001 (2.17%, 13 of 597) to 2019 (39.47%, 328 of 831). Currently, six universities offer two e-commerce programs with different majors. Eight universities established specialized e-commerce schools. There are also six jointly founded or cooperative e-commerce programs run in China with overseas universities. This research may be valuable for any international organization interested in collaboration with China's e-commerce higher education. A limitation is that this research focuses only on bachelors of e-commerce programs. Further research will explore factors for success in jointly founded e-commerce programs with China's e-commerce educators.

**Keywords:** e-commerce, e-commerce education, e-commerce program, higher education, internationalization

## 1. Introduction

In the past 40 years, China's higher education has undergone the transition from elite focused education to popular and mass education. In 2019, the number of students enrolled in China's higher institutions was 8.2 million, and the enrollment rate was 79.53% [1]. Students enrolled in the bachelor program of China's universities were 4.22 million, while the enrollment rate was 43.3% [2]. It is estimated that the enrollment rate of higher education will reach over 60% by 2035 [3]. Comparing the number of students enrolled (0.27 million) and the enrollment rate (5%) in 1977 [4], this is a remarkable growth in the development of China's higher education.

Since the Internet started to become popular with the public in 1994, the electronic commerce (e-commerce) market has evolved from a simple counterpart of brick and mortar retail to a shopping ecosystem; when looking at the e-commerce landscape, a relatively mature market with established players and a clear set of rules can be seen [5]. Among them, China's e-commerce market is expected to grow

by 20% annually over the 5 years since 2018 which is twice as fast as the United States or the United Kingdom [6]. Thus, the e-commerce industry requires quality talent in e-commerce.

However, a number of issues are challenging China's e-commerce higher education. Business managers feel that it is still difficult to find satisfying e-commerce talents. In the meantime, e-commerce graduates found that it was difficult to get appropriate job positions. Research shows that managers' Knowledge has become one of the critical success factors (CSFs) for small and medium enterprises (SMEs) for adopting e-commerce successfully [7]. Staff competency is also vital to successful e-commerce adoption [7, 8]. The probability of the acceptance of e-commerce is linked to higher individuals' awareness and knowledge of e-commerce [9]. Regular training may help staff in better understanding new and updated systems adopted for business processes. Organizations with strong technical expertise and e-commerce knowledge that provide e-commerce training are most likely to realize e-commerce implementation success [10].

The purpose of this chapter is to review the development of China's e-commerce higher education and explore the requirements of the internationalization of China's e-commerce higher education. The Benefit-Driving Model (BDM) was adopted to explain the reasons for the internationalization of China's e-commerce higher education. The literature review spans a 20-year review since 2001 when the first 13 e-commerce programs were offered. The following section will explain the benefits driving the internationalization of China's higher education. The third section will review China's e-commerce higher education. The fourth section will discuss the internationalization of China's e-commerce higher education. The fifth section will provide four suggestions for the further development of China's e-commerce programs. The last section will focus on conclusions, research limitations, and further research.

## **2. The benefit of driving the internationalization of China's higher education**

Wu and Yu [11] developed the Benefit-Driving Model (BDM) for illustrating the factors influencing the internationalization of China's higher education. The BDM will explain the reasons for the internationalization of China's e-commerce higher education. According to BDM, there are three driving factors pushing China to open her educational market linked to three prominent benefits for China (see **Figure 1**).

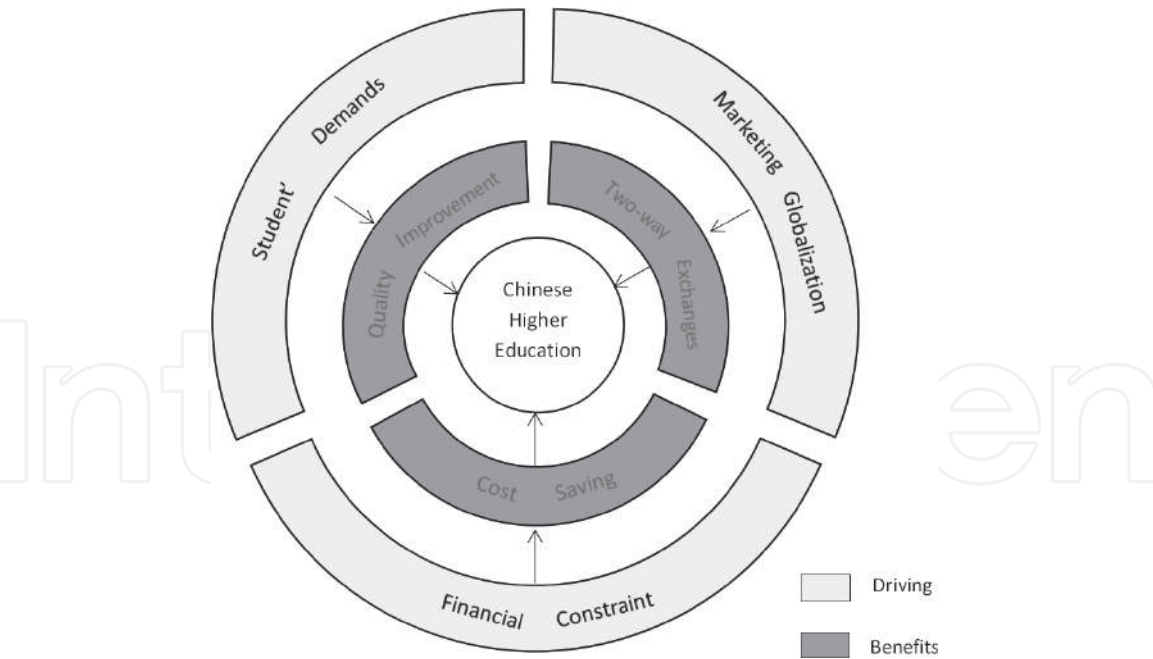
### **2.1 The first driver for the growth is students' demands**

China's students' demands drive quality improvement in China's higher education. Students' desires are to obtain advanced educational training so as to improve competitive capacity and to increase career opportunities. This pressure from students pushes China's universities to improve their educational quality and to catch up with the recent advances in higher education [11].

By 2019, 21 joint-founded universities had been successfully established in China (see Appendix A). There are about 450,000 students in international cooperational programs in China, which are 1.4% of the number of students enrolled in China's universities [12].

### **2.2 The second driver is marketing globalization**

The second driver is the marketing globalization, which benefits the two-way exchanges. International cooperative programs do not only provide opportunities



**Figure 1.**  
*The Benefit-Driving Model ([11], p. 211).*

for students to accept overseas higher education in China, but opening an educational market can also attract overseas students to study in China. Many international cooperative programs and several cooperative universities are operating in China. This is an explicit trend that China has increasingly become one of the international education providers.

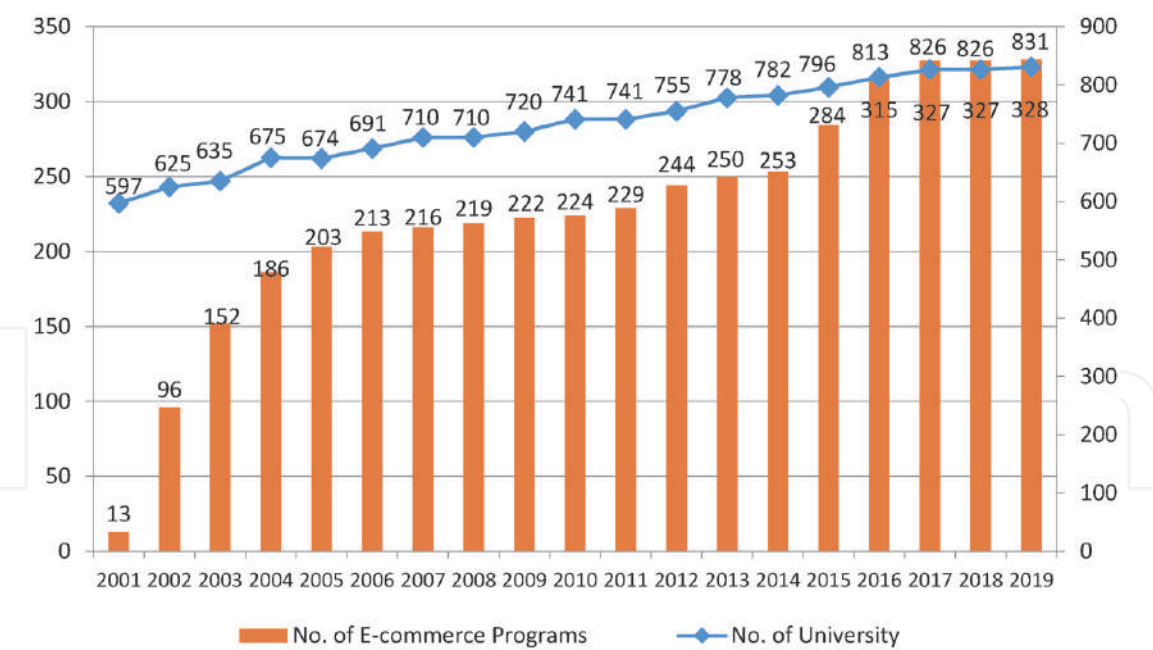
China has undergone a transition from a one-way education outflow to a two-way student exchange market. The number of Chinese students studying abroad in 2018 was over 662,100 [13, 14]. At the same time, the number of overseas students from 196 countries studying in China had increased to more than 492,200 including 258,122 studying at China's universities [14]. In 1950, there were only 33 foreign students from Eastern European countries studying in China [15].

### 2.3 The third driver is financial constraint

The driver of financial constraint provides an opportunity for international higher education providers to joint-found or cooperate international programs in China. Cooperative programs may reduce the costs of moving overseas. Expensive tuition fees prevent many Chinese students from studying overseas. Cooperative programs provide the opportunity for those students who wish to access the advanced educational resources offered by overseas higher educational institutions. The students just pay about \$5000 per year for enrolled in such joint-founded or cooperated programs in China [16]. It saves approximately 70% of tuition fees compared to studying overseas.

## 3. China's e-commerce higher education

The world's first undergraduate e-commerce program was offered by Acadia University, Canada, in September 2000 [17], where the University of California, San Diego offered a master's degree in e-commerce in 1998 [18]. China's universities started to recruit students in bachelor of e-commerce programs in September 2001 [19]. Thirteen of China's universities offered a bachelor of e-commerce program in 2001. As one of the international pioneers, China's education sector



**Figure 2.**  
The development of China’s e-commerce higher education between 2001 and 2019 (Source: Data in 2001–2015 from Wu et al. [20]; Data in 2016–2019 from MOE [3, 21, 22]).

Schools	No.	%
Economics Management	120	36.59
Business	62	18.90
Management	33	10.06
MIS	31	9.45
Computer Science or IT	23	7.01
Business Administration	15	4.57
Management Science	11	3.35
Economics and Trading	8	2.44
E-commerce	8	2.44
Transportation and Logistics	5	1.52
International	4	1.22
Business Planning	1	0.30
Intellectual Property	1	0.30
Tropical Agriculture and Forestry	1	0.30
Science, Technology, and Art	1	0.30
Innovation and Entrepreneurship	1	0.30
Arts, History, and Law	1	0.30
Humanities	1	0.30
Big Data Engineering	1	0.30
Sum	328	100

**Table 1.**  
E-commerce programs by 16 schools.



has been involved in e-commerce programs since the beginning of the twenty-first century [20].

3.1 Run with different majors

By September 2019, 831 universities had been established in China and 328 universities (39.47%, 328 of 831) offered e-commerce programs. This is a sustainable growth since 2001 (2.17%, 13 of 597) (see **Figure 2**). They are currently provided by 19 different schools including Economics Management; Business; Management; Management Science; Management Information System; Computer Science or IT; Business Administration; Economics and Trading; E-commerce; Transportation and Logistics; Business Planning; Intellectual Property; Tropical

University	School	Major	Reference
Zhongnan University of Economics and Law	Business Administration	Marketing Management	ZUEL [23]
	IT and Security	Computer Science and IT	ZUEL [24]
Capital University of Economics and Business	Business Administration	E-commerce	CUEB [25]
	IS	MIS	CUEB [26]
Shandong University	Business	E-commerce	SDU [27]
	Management	E-commerce	SDU [28]
Shenyang University of Technology	Management	Logistics Management and E-commerce	SUT [29]
	Business	E-commerce	SUT [30]
Zhejiang Wanli University	Logistics and E-commerce	E-commerce	ZWU [31]
	Law	E-commerce and Law	ZWU [32]
Southwest University	Computer and Information Science	E-commerce	SWU [33]
	Applied Technology	E-commerce	SWU [34]

**Table 2.**  
*Offering two programs within the university.*

No.	University	School	Established	Reference
1	Henan University of Economics and Law	E-commerce and Logistics Management	2009	HUEL [35]
2	Henan College of Animal Husbandry and Economics	Logistics and E-commerce	2015	HNUAHE [36]
3	Luoyang Normal University	E-commerce	2015	LYNU [37]
4	Jiujiang University	E-commerce	2015	JJU [38]
5	Zhejiang Wanli University	Logistics and E-commerce	2016	ZWU [39]
6	Nanyang Institute of Technology	E-commerce	2016	NYIST [40]
7	Zhejiang Technology and Business University	Management Engineering and E-commerce	2017	ZJSU [41]
8	Zhejiang International Studies University	Cross-border E-commerce	2018	ZISU [42]

**Table 3.**  
*E-commerce schools established in China.*

Agriculture and Forestry; International; Science, Technology and Art; Innovation and Entrepreneurship; Arts, History, and Law; Humanities; and Big Data Engineering (see **Table 1**).

Most of China's e-commerce programs focus on the field of business and management (73.47%, 241 of 328) including Economics Management (36.59%, 120 of 328), Business (18.90%, 62 of 328), Management (10.06%, 33 of 328), Business Administration (4.57%, 15 of 328), Management Science (3.35%, 11 of 328), and only 9.45% (31 of 313) and 7.01% (23 of 313) focus on MIS and Computer Science (IT) in 2019, respectively (see **Table 1**).

### 3.2 Two e-commerce programs offered within the same university

Six universities offer two e-commerce programs with different majors by different schools within the same university (see **Table 2**). These specialization majors in e-commerce include Marketing Management, Computer Science, MIS, Logistics Management, and Law, which are offered by the schools of Business Administration, Business and Management, Management, Logistics and E-commerce, IT and Security, IS, Computer and Information Science, Applied Technology, and Law.

### 3.3 Eight e-commerce schools established

As the first mover, Henan University of Economics and Law established a school of E-commerce and Logistics Management in 2009. Three universities followed and established e-commerce schools in 2015. Henan College of Animal Husbandry and Economics established a school of Logistics and E-commerce; Luoyang Normal University and Jiujiang University established a school of E-commerce; in 2016, Zhejiang Wanli University established a school of Logistics and E-commerce; and Nanyang Institute of Technology established a school of E-commerce. In 2018, Zhejiang University of International Studies established the school of Cross-border E-commerce. Thus, eight China's universities established e-commerce schools as shown in **Table 3**.

## 4. Internationalization of China's e-commerce higher education

The University of Nottingham, UK, in partnership with Zhejiang Wanli University, China, launched the first overseas joint-founded university – the University of Nottingham, Ningbo China in Autumn 2004 in China [43]. Many countries have since then exported their advanced higher education programs to China. It is predicted that the coverage of cooperative educational programs is likely to continue to increase substantially.

The first joint-founded e-commerce programs run in 2004. Clearly, China's e-commerce higher education took steps to keep up with the internationalization of education, while China is embracing the world's economy and markets since entering the twenty-first century. In the internationalization's review of China's e-commerce programs, six joint-founded or cooperative e-commerce programs are run in China with overseas universities (see **Table 4**).

Beijing University of Posts and Telecommunications and the Queen Mary University of London were firstly joint-founded the bachelor of e-commerce and law in 2004. Three joint-founded programs were then followed and run by Zhengzhou Institute of Light Industry jointed with Edinburgh Napier University, UK; Jilin University of Finance and Economics jointed with Charles Sturt University, Australia; and Beijing Normal University jointed with Hong Kong Baptist University in

China's university	Cooperative university	Country	Program	Year	Reference
Beijing University of Posts and Telecommunications	Queen Mary University of London	United Kingdom	E-commerce and Law	2004	BUPT [44]
Zhengzhou Institute of Light Industry	Edinburgh Napier University	United Kingdom	E-commerce	2005	ZZULI [45]
Jilin University of Finance and Economics	Charles Sturt University	Australia	E-commerce	2005	JLUFE [46], MOE [47]
Beijing Normal University	Hong Kong Baptist University	Hong Kong	E-business Management and Information Systems	2005	UIC [48]
Nankai University	Neoma Business School	France	E-commerce	2017	Nankai [49]
Guizhou University of Finance and Economics	Marshall University	United States	E-commerce	2017	Xuezhangbb [50]

**Table 4.**  
*Joint-founded e-commerce programs.*

2005. Two programs run in 2017 including Nankai University jointed with Neoma Business School in France and Guizhou University of Finance and Economics jointed with Marshall University in the United States.

These programs were jointly founded in 2004 with five different countries and regions including the United Kingdom (2), Australia (1), France (1), the United States (1), and Hong Kong, China (1). China's e-commerce higher education sector has been involved in the internationalization of higher education.

## 5. Suggestions for the development of China's e-commerce programs

For the better development of China's e-commerce programs further, the following four suggestions will be provided and discussed.

### 5.1 Learning curriculum from international experience

Although there are 328 e-commerce programs run in 2019, there are only six joint-founded or cooperative e-commerce programs run with overseas universities. **Table 5** shows the international pioneers in e-commerce education [20]. China's e-commerce educators could learn the experience of curriculum development from these international pioneers.

### 5.2 Integrating e-commerce courses into postgraduate programs

CEO and senior staff IT/e-commerce/e-commerce marketing knowledge play critical roles for SMEs successfully in adopting e-commerce [51]. If the decision-maker is knowledgeable about the issues and reliability problems on the Internet, he/she is likely to make a more informed decision about e-commerce adoption [52]. The higher the managers' knowledge of e-commerce, the higher the probability of the acceptance of e-commerce [52]. Senior business management knowledge is highly relevant to e-commerce success.



Year	Country	E-commerce program provider	Program
July 1998	United States	University of California, San Diego	Master degree program in e-commerce
July 1999	United Kingdom	University of Portsmouth Business School	MA marketing with e-commerce
January 2000	United States	Boston University's Metropolitan College	Master of science in e-commerce
September 2000	Canada	Acadia University	Bachelor of computer science with a specialization in e-commerce
March 2001	Australia	Central Queensland University	Master of e-commerce
October 2001	United Kingdom	Liverpool University	Undergraduate degree in e-business
2001	New Zealand	The University of Waikato	Undergraduate degree in e-commerce

**Table 5.**  
*The international pioneers in e-commerce education.*

5.3 Developing teaching materials based on industry requirements

Although many real business cases have been discussed, teaching materials still lag the business and industry requirements. Although some China's universities have established the number of joint programs with industries, China's e-commerce education needs improvement in business practices, industry requirements, and industry involvement [20]. Innovative technologies have not yet been introduced into e-commerce education, such as virtual reality (VR), augmented reality (AR), and mixed reality (MR). Apple has developed ARKit as its own Augmented Reality platform for iOS, and Google has developed ARCore as its own Augmented Reality platform for Android [53]. The emerging innovative technology of VR, AR, and MR may be widely used for developing immersive e-commerce systems and enhancing customer online experience. It should be encouraged to adopt the real industry project into teaching materials and study assessments.

5.4 Offering specialization major in cross-border e-commerce

Despite a slowing Chinese economy, a shift in purchasing power from the U.S. and Europe to China and Southeast Asia has begun [5]. China's cross-border retail e-commerce sales are projected by eMarketer to reach \$245 billion by 2020 [54]. China has announced another 24 cities as pilot zones for cross-border e-commerce to boost exports in December 2019 [55]. Cross-border e-commerce has thus expected as one of the dominating industry sectors and contributors to impetus the development of China's economy. There is only Zhejiang Foreign Studies University that offers a specialization major in Cross-border e-commerce.

6. Conclusions, limitations, and further research

Since 2004, Chinese higher educational institutions have taken steps to catch up with the internationalization of education in terms of collaboration with overseas

universities. The purpose of this chapter is to review the development of e-commerce higher education in China and address the requirements of the internationalization of China’s e-commerce higher education.

The Benefit-Driving Model (BDM) was adopted to address the reasons for the marketability of internationalization of China’s e-commerce higher education. A 20-year review of China’s e-commerce program found that there was sustainable growth from 2001 (2.17%, 13 of 597) to 2019 (39.47%, 328 of 831). Three hundred and twenty-eight e-commerce programs are run by 19 different schools. Six universities offer two e-commerce programs with different majors. Eight universities established specialized e-commerce schools. There are also six joint-founded or cooperative e-commerce programs run in China with overseas universities. There are opportunities to improve including adopting the learning curriculum from international experience, integrating the e-commerce courses into postgraduate programs, developing the teaching materials based on industry requirements, and offering the specialization major in cross-border e-commerce.

Although this research focused only on China’s e-commerce higher education, the increasing demand will also affect international higher education providers. This research should be also of interest for any international education organizations attracted to China’s e-commerce higher education.

Acknowledgements

This research is sponsored by the Fund for Shanxi “1331 Project” Collaborative Innovation Center, China.

Appendix

21 Sino-foreign cooperative universities

Cooperative university	China's university partner	Foreign university partner	Country/ religion	Established
University of Nottingham Ningbo China	Zhejiang Wanli University	University of Nottingham	British	2004
United International College	Beijing Normal University	Hong Kong Baptist University	Hong Kong, China	2005
Xi'an Jiaotong-Liverpool University	Xi'an Jiaotong University	University of Liverpool	British	2006
NYU Shanghai	East China Normal University	New York University	United States	2012
Duke Kunshan University	Wuhan University	Duke University	United States	2013
Wenzhou Kean University	Wenzhou University	Ken University	United States	2014
The Chinese University of Hong Kong, Shenzhen	Shenzhen University	Chinese University of Hong Kong	Hong Kong, China	2014
Shenzhen MSU-BIT University	Beijing Institute of Technology	Moscow State University	Russia	2016

Cooperative university	China's university partner	Foreign university partner	Country/ religion	Established
Guangdong Technion – Israel Institute of Technology	Shantou University	Israel Institute of Technology	Israel	2016
SWUFE-UD Institute of Data Science at Southwestern University of Finance and Economics	Southwestern University of Finance and Economics	University of Delaware	United States	2019
FESTU Transport Institute of Dalian Jiaotong University	Dalian Jiaotong University	Far Eastern State Transport University,	Russia	2019
Kyiv College at Qilu University of Technology	Qilu University of Technology	Kyiv National University of Technologies and Design	Ukraine	2019
MSU Institute, Nanjing Agricultural University	Nanjing Agricultural University	Michigan State University	United States	2019
FedUni Information Engineering Institute, Hebei University of Science and Technology	Hebei University of Science and Technology	Federation University Australia	Australia	2019
Aulin College, Northeast Forestry University	Northeast Forestry University	University of Auckland	New Zealand	2019
Portland Institute, Nanjing University of Posts and Telecommunications	Nanjing University of Posts and Telecommunications	Portland State University	United States	2019
Ulster College at Shaanxi University of Science & Technology	Shaanxi University of Science and Technology	University of Ulster	United Kingdom	2019
Detroit Green Technology Institute, Hubei University of Technology	Hubei University of Technology	University of Detroit Mercy	United States	2019
SDU-ANU Joint Science College, Shandong University	Shandong University	Australian National University	Australia	2019
Don College of Shandong Jiaotong University	Shandong Jiaotong University	Don State Technical University	Russia	2019
Chengdu University of Technology Oxford Brookes College	Chengdu University of Technology	Oxford Brookes University	United Kingdom	2019

IntechOpen

## Author details

Wenying Huo<sup>1</sup>, Mingxuan Wu<sup>1,2\*</sup> and Jeffrey Soar<sup>3</sup>

<sup>1</sup> Shanxi Normal University, China

<sup>2</sup> School of Engineering and Technology, CQUniversity, Australia

<sup>3</sup> University of Southern Queensland, Australia

\*Address all correspondence to: [robert\\_wumx@hotmail.com](mailto:robert_wumx@hotmail.com)

## IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Xixik. The Number of College Entrance Examination Applicants and Enrollment Over the Years. 2019. Available from: <http://114.xixik.com/gaokao/> [Accessed: 18 November 2014]
- [2] Sina. The Enrolment Rate of Bachelor Programs Students Enrolled in Chinese Universities was 43.3% in 2018. 2019. Available from: [https://k.sina.com.cn/article\\_6902197713\\_19b672dd100100of19.html?from=edu](https://k.sina.com.cn/article_6902197713_19b672dd100100of19.html?from=edu) [Accessed: 18 December 2019]
- [3] EDU. Work Highlights of the Higher Education Department of the Ministry of Education in 2018. 2018. Available from: [https://www.edu.cn/edu/zhe ng\\_ce\\_gs\\_gui/zheng\\_ce\\_wen\\_jian/zonghe/201803/t20180330\\_1592935.shtml](https://www.edu.cn/edu/zhe ng_ce_gs_gui/zheng_ce_wen_jian/zonghe/201803/t20180330_1592935.shtml) [Accessed: 30 March 2018]
- [4] Koolearn. Summary of College Entrance Examination Enrollment and Admission Rate from 1977 to 2018. 2019. Available from: <https://gaokao.koolearn.com/20190228/1208329.html> [Accessed: 15 March 2019]
- [5] Statista. eCommerce Report 2019. 2019. Available from: <https://www.statista.com/study/42335/e-commerce-report/> [Accessed: 05 December 2019]
- [6] Aliza. A What China Reveals about the Future of Shopping. 2017. Available from: <https://www.alizila.com/china-reveals-future-shopping> [Accessed: 17 April 2019]
- [7] Wu MX, Gide E, Jewell R. The EBS management model: An effective measure of E-commerce satisfaction in SMEs in service industry from management perspective. *Electronic Commerce Research*. 2014;14(1):71-86
- [8] Awiagah R, Kang J, Lim JL. Factors affecting E-commerce adoption among SMEs in Ghana. *Information Development*. 2016;32(4):815-836
- [9] Choshin M, Ghaffari A. An investigation of the impact of effective factors on the success of E-commerce in small- and medium-sized companies. *Computers in Human Behavior*. 2017; 66:67-74
- [10] Migdadi MM, Zaid MKSA, Al-Hujran OS, Aloudat AM. An empirical assessment of the antecedents of electronic-business implementation and the resulting organizational performance. *Internet Research*. 2016; 26(3):661-688
- [11] Wu MX, Yu P. Challenges and opportunities facing Australian universities caused by the internationalization of Chinese higher education. *International Education Journal*. 2006;7(3):211-221. Australia. Reprinted by Icfai Journal of Higher Education, India, Feb 2007 (with permission of International Education Journal)
- [12] Tulane-gscass. The Main Characteristics and Development Problems of the New Era of Sino-foreign Cooperative Education. 2019. Available from: <http://www.tulane-gscass.com/news/project/20190923578.html> [Accessed: 17 October 2019]
- [13] MOE. Statistics of Chinese Students Studying Abroad in 2018. 2019. Available from: [http://www.moe.gov.cn/jyb\\_xwfb/gzdt\\_gzdt/s5987/201903/t20190327\\_375704.html](http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/201903/t20190327_375704.html) [Accessed: 19 December 2019]
- [14] Xinhuanet. Ministry of Education: The Total Number of Chinese Students Studying Abroad Reached 662,100 in 2018. 2019. Available from: [http://www.xinhuanet.com/2019-03/27/c\\_1124291948.htm](http://www.xinhuanet.com/2019-03/27/c_1124291948.htm) [Accessed: 19 November 2019]
- [15] MOE. Overview of Overseas Study in China. 2009. Available from: <http://>



- [www.moe.gov.cn/s78/A20/gjs\\_left/moe\\_850/tnull\\_48305.html](http://www.moe.gov.cn/s78/A20/gjs_left/moe_850/tnull_48305.html) [Accessed: 19 November 2019]
- [16] Eduego. How are the Tuition Fees Charged for Chinese-Foreign Cooperation in Running a School? 2018. Available from: <https://www.eduego.com/wenti/35741.html> [Accessed: 19 August 2019]
- [17] Trudel C, Trudel. World's first undergraduate E-commerce specialization offered by a computer science department. *Journal of Computing Sciences in Colleges*. 2004; **20**(2):254-259
- [18] Weinstein B. Calif. University is First to Offer Master's Degree in E-commerce, May 31, *Boston Globe*. 1998
- [19] Zhu Q, Zhao W. 13 University Have Been Approved To Run Bachelor of E-Commerce Programs. 2001. Available from: <http://www.dzsw.org/news/study/meeting/200106/169.html> [Accessed: 01 August 2004]
- [20] Wu MX, Wang MM, Soar J, Gide E. China's E-commerce higher education: A 15 years review from international viewpoint. *Open Journal of Social Sciences*. 2016;**4**:155-164
- [21] MOE. The Ministry of Education of the People's Republic of China 2017, The Number of Students in Higher Education Schools (Institutions). 2018. Available from: [http://www.moe.gov.cn/s78/A03/moe\\_560/jytjsj\\_2017/qg/201808/t20180808\\_344685.html](http://www.moe.gov.cn/s78/A03/moe_560/jytjsj_2017/qg/201808/t20180808_344685.html) [Accessed: 06 August 2019]
- [22] MOE. Notice of the Ministry of Education on the Announcement of the Archival Filing and Examination Results of Undergraduate Majors of Ordinary Institutions of Higher Learning in 2018. 2019. Available from: [http://www.moe.gov.cn/srcsite/A08/moe\\_1034/s4930/201903/t20190329\\_376012.html](http://www.moe.gov.cn/srcsite/A08/moe_1034/s4930/201903/t20190329_376012.html) [Accessed: 15 July 2019]
- [23] Zhongnan University of Economics and Law (ZUEL). Introduction to Business Administration School. 2018. Available from: <http://gsxy.zuel.edu.cn/4111/list.htm> [Accessed: 20 July 2019]
- [24] Zhongnan University of Economics and Law (ZUEL). Introduction to Information and Security Engineering School. 2018. Available from: <http://xagx.zuel.edu.cn/2096/list.htm> [Accessed: 20 July 2019]
- [25] Capital University of Economics and Business (CUEB). Introduction to Business Administration School. 2015. Available from: <https://cba.cueb.edu.cn/xygk/xyjj/index.htm> [Accessed: 20 July 2019]
- [26] Capital University of Economics and Business (CUEB). Introduction to Information Management and Information Systems. 2015. Available from: <https://ggxy.cueb.edu.cn/bkjjyx/bzyjs/106112.htm> [Accessed: 20 July 2019]
- [27] Shandong University (SDU). Electronic Commerce Professional Settings. 2017. Available from: <https://shxy.wh.sdu.edu.cn/info/1036/1570.htm> [Accessed: 20 July 2019]
- [28] Shandong University (SDU). School of Management Organizational Structure. 2018. Available from: <http://www.glxy.sdu.edu.cn/xygk/zzjg.htm> [Accessed: 20 July 2019]
- [29] Shenyang University of Technology (SUT). School of Management Enrollment Information. 2019. Available from: <https://zsxxw.sut.edu.cn/info/1086/1378.htm> [Accessed: 20 July 2019]
- [30] Shenyang University of Technology (SUT). Business College Enrollment Information. 2019. Available from: <https://zsxxw.sut.edu.cn/info/1086/1385.htm> [Accessed: 20 July 2019]

- [31] Zhejiang Wanli University (ZWU). College of Logistics and Electronic Commerce Major Setting. 2018. Available from: <http://xdwl.zwu.edu.cn/820/list.htm> [Accessed: 20 July 2019]
- [32] Zhejiang Wanli University (ZWU). Specialty Setting of Law School. 2018. Available from: <http://xdwl.zwu.edu.cn/820/list.htm> [Accessed: 20 July 2019]
- [33] Southwest University (SWU). College of Computer and Information Science Undergraduate Major Construction. 2015. Available from: <http://computer.swu.edu.cn/s/computer/jyxx2zyjs/> [Accessed: 20 July 2019]
- [34] Southwest University (SWU). Introduction to the School of Applied Technology. 2018. Available from: <http://gaozhi.swu.edu.cn/> [Accessed: 20 July 2019]
- [35] Henan University of Economics and Law (HUEL). Introduction to the School of E-commerce and Logistics Management. 2017. Available from: <http://dswl.huel.edu.cn/list.jsp?urltype=tree.TreeTempUrl&wbtreeid=1003> [Accessed: 15 July 2019]
- [36] Henan University of Animal Husbandry and Economics (HNUAHE). Introduction to Logistics and Electronics Business School. 2015. Available from: <http://wlyds.hnuah.edu.cn/xygk/xyjj.htm> [Accessed: 15 July 2019]
- [37] Luoyang Normal University (LYNU). Introduction to Electronics Commerce School. 2015. Available from: <http://sites.lynu.edu.cn/dzswxy/xygk/xyjj.htm> [Accessed: 15 July 2019]
- [38] Jiujiang University (JJU). Introduction to Electronics Commerce School. 2018. Available from: <http://dsxy.jju.edu.cn/info/1002/2503.htm> [Accessed: 15 July 2019]
- [39] Zhejiang Wanli University (ZWU). College of Logistics and e-Commerce Professional Settings. 2016. Available from: <http://xdwl.zwu.edu.cn/> [Accessed: 15 July 2019]
- [40] Nanyang Institute of Technology (NYIST). Introduction to the School of Electronic Commerce. 2019. Available from: <http://dsxy.nyist.edu.cn/> [Accessed: 15 July 2019]
- [41] Zhejiang Technology and Business University (ZJSU). Overview of School of Management Engineering and E-commerce. 2017. Available from: <http://sme.zjsu.edu.cn/index.htm> [Accessed: 15 July 2019]
- [42] Zhejiang International Studies University (ZISU). Overview of Cross-border E-commerce College and School of Science and Technology. 2019. Available from: <http://kjxy.zisu.edu.cn/> [Accessed: 15 July 2019]
- [43] UNNC. The University of Nottingham, Ningbo China. 2004. Available from: <http://www.unnc.edu.cn/english/about/> [Accessed: 15 November 2004]
- [44] Beijing University of Posts and Telecommunications (BUPT). Introduction to International College. 2013. Available from: <https://is.bupt.edu.cn/xygk/xyjj.htm> [Accessed: 20 August 2019]
- [45] Zhengzhou University of Light Industry (ZZULI). Zhengzhou Institute of Light Industry and Napier University of Edinburgh Jointly Hold the Introduction to the Undergraduate Education Program of Sino-foreign Cooperation in E-commerce. 2018. Available from: <http://iec.zzuli.edu.cn/2018/1120/c14010a188329/page.htm> [Accessed: 20 August 2019]
- [46] Jilin University of Finance and Economics (JLUFE). Introduction to the School of International Exchange. 2018. Available from: <http://newgjil.jlufe.edu>

cn/About/Introduction/ [Accessed: 20 August 2019]

[47] MOE. Approval of the Cooperation between Changchun Institute of Taxation and Charles Sturt University of Australia for the Bachelor of Commerce in E-commerce. 2000. Available from: [http://www.moe.gov.cn/srcsite/A22/s7065/200006/t20000613\\_165260.html](http://www.moe.gov.cn/srcsite/A22/s7065/200006/t20000613_165260.html) [Accessed: 20 August 2019]

[48] Hong Kong Baptist University, Beijing Normal University (UIC). About UIC. 2019. Available from: <https://uic.edu.hk/en/about-us/introduction/about-uic> [Accessed: 20 August 2019]

[49] Nankai University (Nankai). Nankai University and Nuo Business School, France, Bachelor's Degree in E-commerce/International Business. 2019. Available from: <https://bs.nankai.edu.cn/2019/0810/c10085a190077/page.htm> [Accessed: 20 August 2019]

[50] Xuezhangbb. Guizhou University of Finance and Economics Cooperates with Marshall University in the United States to Offer Undergraduate Education Programs in E-commerce. 2018. Available from: <http://www.xuezhangbb.com/cooperate/987> [Accessed: 20 August 2019]

[51] Wu MX, Jewell R, Gide E. An eyeball diagram: Illustrating the common CSFs in E-commerce business satisfaction for successful adoption of E-commerce systems by SMEs. *International Journal of Electronic Customer Relationship Management*. 2012;**6**(2):169-192

[52] Gorla N, Chiravuri A, Chinta R. Business-to-business E-commerce adoption: An empirical investigation of business factors. *Information Systems Frontiers*. 2017;**19**:645-667

[53] Smith A. What's the Influence of Augmented Reality on the Ecommerce

Industry? 2018. Available from: <https://www.itproportal.com/features/whats-the-influence-of-augmented-reality-on-the-ecommerce-industry/> [Accessed: 08 November 2019]

[54] Aliexpo. Access to China: Learn How to Sell to the China Consumer. 2020. Available from: <https://aliexpo.com.au/why-attend/access-to-china> [Accessed: 08 January 2020]

[55] CIW. China Approved 24 More Cross-border E-commerce Pilot Cities and Relaxed Forex in FTZs. 2020. Available from: <https://www.chinainternetwatch.com/tag/cross-border-e-commerce/> [Accessed: 25 February 2019]

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)





# An Overview of Digital Entrepreneurship in Central and Eastern European Countries

*Mladen Turuk*

## Abstract

The aim of the study is to explore and present an overview of digital entrepreneurship in Central and Eastern European countries and to examine how certain components of the DESI index affect GDP per capita in CEE countries and in what way modern information technologies affect their economies. The paper uses secondary data sources, mostly scientific and professional journals from the studied area, DESI reports, Eurostat data, and other Internet sources. The first part of the paper presents a short introduction on digitization digital entrepreneurship and digital technologies. The second part provides a descriptive analysis of digital entrepreneurship indicators and explores business demography in the ICT sector while the third part refers to the analysis of the DESI index. The panel method on data from 2015 to 2019 was used to show the influence of the different DESI index components on the observed countries' GDP per capita. The hypothesis that the components of the DESI index have a positive impact on GDP per capita has been partially confirmed. DESI rank, Connectivity and Human capital did not prove to be significant, while Use of internet services, Integration of digital technology, and Digital public services proved their significant positive effect.

**Keywords:** digital entrepreneurship, digital economy, digital society, ICT, CEE countries

## 1. Introduction

Digitization does not only change certain segments of business and individual industries – it fully affects all spheres of society and the economy, both technologically and organizationally. Digital technologies based on new platforms can transform the way economies function and impact all sectors of the economy, including traditional ones. Digital technologies “have the potential to create new or expand existing goods and services with digital features – yet possibilities in this regard depend on the characteristics of specific sectors' end products” [1, 2]. Full efficiency and profitability are impossible without digital transformation in which the private sector can and must be a leader. Although the crisis caused by COVID 19 severely affected a number of industries, the economic impact on the technology, media, and telecommunications sectors was largely neutral or even positive for some industry segments. For Europeans to take advantage of the opportunities offered by digital technologies, the European Commission adopted its digital strategy on 19



February 2020. During the coronavirus crisis, this strategy is even more important in creating favorable environment for digital entrepreneurship.

The aim of the study is to explore and present an overview of digital entrepreneurship in Central and Eastern European countries and to examine how certain components of the DESI index affect GDP per capita in CEE countries and in what way modern information technologies affect their economies. Central and Eastern European countries are European Union member states which were once part of the former Eastern bloc. The following countries are Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, and Slovenia.

The study is further structured as follows. The second part provides an overview of digital entrepreneurship. The third part explains research methodology followed by the descriptive analysis of information and communication technology sector of CEE countries and provides an insight into ICT share in GDP, the share of ICT employment in total employment and the share of R&D in the ICT industry in total R&D. Furthermore, it analyses business demography of the ICT sector and provides an overview of enterprises' birth, death, and churn rates. The fourth part of the study provides Digital Economy and Society Index analysis and relates observed DESI components to countries' GDP per capita while final remarks are presented in conclusion.

## **2. Overview of digital entrepreneurship**

Entrepreneurship, in its simplest form, can be described as self-employment [3]. Digital entrepreneurship, on the other hand, diverges from this definition seeing as it involves entrepreneurial pursuits which occur on a digital platform ([4], as in [5], p. 1). Digital entrepreneurship is "an essential driver within the innovation system. It changes the structure, aims, and networking mechanisms of the overall business system and, ultimately, affects the various levels and dimensions of the innovation system" [2], p. 1. Bringing inevitable changes to the innovation system, digital technologies may not only provide new business opportunities but also be disruptive and cause new vulnerabilities" [2], p. 1. „The term 'Digital Entrepreneurship' most commonly refers to the process of creating a new - or novel - Internet enabled/delivered business, product or service. This definition includes both start-ups - bringing a new digital product or service to market - but also the digital transformation of an existing business activity inside a firm or the public sector" [6], p. 1,

Digital entrepreneurship is the practice of pursuing "new venture opportunities presented by new media and internet technologies" [7], p. 8. Digital entrepreneurship is "a subcategory of entrepreneurship in which some or all of what would be physical in a traditional organization has been digitized" [8], p. 4. "Digital entrepreneurship embraces all new ventures and the transformation of existing businesses that drive economic and/or social value by creating and using novel digital technologies. Digital enterprises are characterized by a high intensity of utilization of novel digital technologies (particularly social, big data, mobile, and cloud solutions) to improve business operations, invent new business models, sharpen business intelligence, and engage with customers and stakeholders. They create the jobs and growth opportunities of the future" [9], p. 1. Digital enterprises are different from traditional entrepreneurial ventures because they have different business models and can pursue their products, marketing and distribution activities using digital platforms [10].

Global diffusion of "digital technologies as general use tools has also spurred arguments that it may increase knowledge diffusion through improved communication efficiency, improve consumer engagement, and allow countries to leapfrog

traditional methods of increasing productivity” [11], p. 1. Online business does not just mean having a good communication strategy which is then marketed through various digital channels. Online business also means connecting business entities in an efficient way, enabling digital transformation that facilitates business in a simpler, more accessible, and often cheaper way. Jobs related to digital technologies or digital economy are the most sought after and most stable occupations. Without the new technologies, digital entrepreneurs “would be unable to deliver their products or services, and in some cases, the business model itself could not exist without information technology. The sector of information and communication technology remains a key driver of innovation and a sector with huge growth potential” [12], p. 180.

### 3. Research methodology

The first part of the analysis refers to the descriptive analysis of ICT sector and its business demography. Percentage of the ICT sector in GDP, total employment, and R&D in CEE countries are analyzed. Moreover, business demography in ICT sector in CEE countries is explained through enterprises’ birth rate, death rate and churn rate. The next chapter analyses DESI index in CEE countries. DESI is a composite index that summarizes relevant indicators on Europe’s digital performance. The main components of the DESI index are a) Connectivity (CON), b) Human capital (HC), c) Use of internet services (IS), d) Integration of digital technology (IDT) and e) Digital public services (DPS). In accordance with the stated aim, the following research hypothesis was formulated: *H1. The components of the DESI index have a positive impact on GDP per capita.* In hypothesis testing, GDP per capita is used as a dependent variable in the model, while the components of the DESI index: Connectivity (CON), Human capital (HC), Use of internet services (IS), Integration of digital technology (IDT) and Digital public services (DPS) – represent independent variables in the model. The analysis uses panel method and begins with estimating the equation using OLS, Random, Between, First difference and Fixed models. The next step is to determine which of the above models best specified the equation. For this purpose, the Breusch-Pagan Lagrange multiplier test and the Hausman test are performed. The Breusch-Pagan Lagrange multiplier test tests whether the “OLS” or “Random” model is suitable. The null hypothesis assumes that the variance between entities or industrial activities is zero, i.e. that there are no panel effects, which indicates the use of the least squares method or the OLS method. Furthermore, the Hausman test helps in choosing a “Fixed” or “Random” model. It tests whether the errors are correlated with the regressors. The null hypothesis assumes that the errors are not correlated with the regressors, which would indicate the use of the “Random” model. The null hypothesis of the Breusch-Pagan / Cook-Weisberg test indicates homoskedasticity in the model. Given that in this case the null hypothesis cannot be rejected at all standard levels of significance, it is concluded that there is no heteroskedasticity in the model.

### 4. Descriptive analysis of ICT sector and its business demography

Prior to the introduction of the DESI index (Digital Economy and Society Index), some of the indicators of the intensity of digital entrepreneurship in the total economy were the share of the ICT industry in GDP, the share of the ICT industry in total employment and the share of R&D in the ICT industry in total R&D. One of the prior indicators of digital development was citizens’ Internet

penetration. Specific studies presented in **Table 1** showed positive correlation of internet penetration on GDP growth.

Koutroumpis [15] and Czernich et al. [16] conducted studies of the impact of the internet on economic growth focused mainly on the EU and US OECD countries. These studies found that a 10 per cent increase in internet penetration correlates with a 0.9–1.5 and a 0.3–0.9 percentage point (pp) in gross domestic product (GDP) growth respectively (Hernandez et al., 2016). Other indicators are analyzed below.

The average share of the ICT industry in CEE countries in GDP is 4.39%, with Hungary having the largest share (6.04%) and Lithuania having the smallest share (3.02%). The Republic of Croatia is in the middle with the share of the ICT industry in GDP of 4.40%. The average share of employment in the ICT industry in CEE countries in total employment is 3.00%, with Estonia having the largest share (5.14%) and Romania having the smallest share (2.36%). The Republic of Croatia is just ahead of Romania with the share of employment in the ICT industry in total employment of 2.45%. The average share of R&D in the ICT industry in CEE countries in total R&D is 0.83%, with Lithuania having the largest share (2.49%)

Study	Country/Region	Years	Correlation with GDP growth
Koutroumpis (2009) [15]	22 OECD countries	2002–2007	0.9–1.5 pp
Czernich et al. (2009) [16]	25 OECD countries	1996–2007	0.3–0.9 pp
Garcia Zaballos and Lopez-Rivas (2012) [17]	26 Latin American and Caribbean countries	2003–2009	3.2 pp
Qiang et al. (2009) [18]	120 countries	1980–2006	1.21 pp.
	Developed countries	1980–2006	1.38 pp
	Developing countries		
Scott (2012) [19]	120 countries	1980–2011	1.19 pp.
	Developed countries	1980–2011	1.35 pp
	Developing countries		

**Table 1.**  
*Correlations with GDP growth for every 10-percentage point (pp) increase in internet penetration [13], p. 7, based on [14].*

Country	ICTGDP (%)	ICTEMP (%)	%ICTR&D
Bulgaria	5,72	2,71	0,79
Croatia	4,40	2,45	0,00*
Czech Republic	4,42	3,07	0,33
Estonia	5,14	4,09	2,30
Hungary	6,04	3,56	0,43
Latvia	4,73	3,84	1,15
Lithuania	3,02	2,57	2,49
Poland	3,33	2,47	0,41
Romania	3,53	2,36	0,36
Slovakia	4,30	3,18	0,32
Slovenia	3,68	2,66	0,53

**Table 2.**  
*Percentage of the ICT sector in GDP, total employment, and R&D in CEE countries, 2017 [20].*

Country	BR (%)	DR (%)	CR (%)
Bulgaria	14,20	11,70	25,90
Croatia	10,54	6,27	16,81
Czech Republic	11,40	6,56	17,96
Estonia	18,39	10,90	29,29
Hungary	13,74	0,21	13,95
Latvia	13,66	5,72	19,38
Lithuania	21,95	26,60	48,55
Poland	16,57	9,37	25,94
Romania	15,39	10,27	25,66
Slovakia	17,41	9,59	27,00
Slovenia	12,63	5,01	17,64

**Table 3.**  
*Business demography in ICT sector in CEE countries in 2017 [20].*

and Slovakia having the smallest share (0.32%). Eurostat data is presented in **Table 2**.

The birth rate of a given reference period (usually one calendar year) is the number of births as a percentage of the population of active enterprises. The death rate of a given reference period (usually one calendar year) is the number of deaths as a percentage of the population of active enterprises. The churn rate is equal to the sum of the birth and the death rate. Eurostat data is presented in **Table 3**.

The average company birth rate in the ICT industry in CEE countries is 15.08%. Lithuania (21.95%) has the highest company birth rate, followed by Estonia (18.39%) and Slovakia (17.41%), while Croatia (10.54%), Czech Republic (11.40%) and Slovenia (12.63%) have the lowest company birth rate (11.40%). The average company death rate in the ICT industry in CEE countries is 9.29%. Lithuania (26.60%) has the highest company death rate, followed by Bulgaria (11.70%) and Estonia (10.90%), while Hungary (0.21%), Slovenia (5.01%) and Latvia (5.72%) have the lowest company closure rate. The average churn rate of companies in the ICT industry in CEE countries is 24.37%. Lithuania has the highest churn rate (48.55%), followed by Estonia (29.29%) and Slovakia (27.00%), while Hungary (13.95%), Croatia (16.81%), and Slovenia (17.64%) have the lowest turnover rate.

## 5. DESI index analysis

The European Commission has been monitoring the intensity of the digital economy since 2014 by publishing DESI reports for individual member states. DESI is a composite index that summarizes relevant indicators on Europe’s digital performance and tracks the evolution of EU Member States in digital competitiveness [21]. The main components of the DESI index are a) Connectivity (CON), b) Human capital (HC), c) Use of internet services (IS), d) Integration of digital technology (IDT) and e) Digital public services (DPS).

Connectivity indicators in the DESI index look at both the demand and the supply side of fixed and mobile broadband and consist of: a) Overall fixed broadband take-up (% households), b) At least 100 Mbps fixed broadband take-up (% households), c) Fast broadband (NGA) coverage (% households), d) Fixed Very



High Capacity Network (VHCN) coverage (% households), e) 4G coverage (% households – average of operators), f) Mobile broadband take-up (Subscriptions per 100 people), g) 5G readiness (Assigned spectrum as a % of total harmonized 5G spectrum), and h) Broadband price index (Score 0 to 100). In connectivity, Latvia had the highest score, followed by Hungary and Romania. Bulgaria, Croatia, and Czech Republic had the weakest performance for this dimension of the DESI.

Human capital in DESI index consists of: a) At least basic digital skills (% individuals), b) Above basic digital skills (% individuals), c) At least basic software skills (% individuals), d) ICT specialists (% total employment), e) Female ICT specialists (% female employment), and f) ICT graduates (% graduates). According to the latest data, Estonia is leading in human capital, followed by Croatia and Czech Republic. Romania, Bulgaria, and Latvia rank the lowest.

Use of internet services in DESI index consist of: a) People who have never used the internet (% individuals), b) Internet users (% individuals), c) News (% internet users), d) Music, videos and games (% internet users), e) Video on demand (% internet users), f) Video calls (% internet users), g) Social networks (% internet users), h) Doing an online course (% internet users), i) Banking (% internet users), j) Shopping (% internet users), and k) Selling online (% internet users). Estonia Lithuania and Hungary have the most active internet users. Conversely, Romania, Bulgaria and Poland are the least active.

Integration of digital technology in DESI index consist of: a) Electronic information sharing (% enterprises), b) Social media (% enterprises), c) Big data (% enterprises), d) Cloud (% enterprises), e) SMEs selling online (% SMEs), f) e-Commerce turnover (% SME turnover) and g) Selling online cross-border (% SMEs). The top performers are Czech Republic, Lithuania, and Croatia. At the other end of the scale are Bulgaria, Romania, and Hungary.

Digital public services in DESI index consist of: a) e-Government users (% internet users needing to submit forms), b) Pre-filled forms (Score 0 to 100), c) Online service completion (Score 0 to 100), d) Digital public services for businesses (Score 0 to 100 – including domestic and cross-border), and e) Open data (% of maximum score). The top performers are Estonia, Latvia, and Lithuania. On the other hand, Romania, Slovakia, and Croatia score the lowest.

## 5.1 Hypothesis

The aim of the study is to explore and present an overview of digital entrepreneurship in Central and Eastern European countries and to examine how certain components of the DESI index affect GDP per capita in CEE countries and in what way modern information technologies affect their economies. In accordance with the stated aim, the following research hypothesis was formulated.

*H1. The components of the DESI index have a positive impact on GDP per capita.*

In hypothesis testing, GDP per capita is used as a dependent variable in the model, while the components of the DESI index: Connectivity (CON), Human capital (HC), Use of internet services (IS), Integration of digital technology (IDT) and Digital public services (DPS) – represent independent variables in the model.

The results of the conducted econometric analysis of the hypothesis test are presented below. The analysis begins with estimating the equation using OLS, Random, Between, First difference and Fixed models.

$$GDP_{pc_{i,t}} = \beta_0 + \beta_1 DESI_{rank_{i,t}} + \beta_2 CON_{i,t} + \beta_3 HC_{i,t} + \beta_4 IS_{i,t} + \beta_5 IDT_{i,t} + \beta_6 DPS_{i,t} + u_{i,t}$$



Fixed-effects (within) regression		Number of obs	=	55
Group variable: id		Number of groups	=	11
R-sq: within = 0.8976		Obs per group: min	=	5
between = 0.3018		avg	=	5.0
overall = 0.3162		max	=	5
corr(u_i, Xb) = 0.2042		F(6,38)	=	55.53
		Prob > F	=	0.0003

GD?pc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
DESIrank	14.07801	43.50148	0.32	0.748	-73.98614	102.1422
CUN	-10.14379	7.22938	-1.40	0.169	-24.7789	4.491324
HC	20.33353	15.99207	1.27	0.211	-12.02047	52.69753
IS	64.74837	15.35768	4.22	0.000	33.65837	95.83837
IDT	34.42375	16.42282	2.10	0.043	1.177485	67.67001
DPS	45.60849	5.531885	8.24	0.000	34.40977	56.80721
_cons	5394.138	1574.009	3.43	0.001	2212.724	8585.553
sigma_u	3166.4107					
sigma_e	271.80314					
rho	.59268547	(fraction of variance due to u_i)				

F test that all u_i=0:	F(10, 38) =	358.53	Prob > F =	0.0003
------------------------	-------------	--------	------------	--------

**Table 4.**  
*Stata panel model output.*

The results of the panel model (fixed effects) are presented in **Table 4** above. The hypothesis has been partially confirmed. Three components of the DESI index did not prove to be significant – DESI rank (positive sign), Connectivity (negative sign) and Human capital (positive sign), while three proved to be significant – Use of internet services (IS), Integration of digital technology (IDT) and Digital public services (DPS) (all three with the positive sign).

5.2 Data analysis

In order to prove the model’s reliability and validity five different models were analyzed. The analysis begins with estimating the equation using OLS, Random, Between, First difference and Fixed models. Integration of digital technology (IDT) proved to be significant in three out of five models (OLS, Random and Fixed), while Internet services (IS) and Digital public services (DPS) proved to be significant in two out of five models (Random and Fixed).

The next step is to determine which of the above models best specified the equation. For this purpose, the Breusch-Pagan Lagrange multiplier test and the Hausman test are performed. The Breusch-Pagan Lagrange multiplier test tests whether the “OLS” or “Random” model is suitable. The null hypothesis assumes that the variance between entities or industrial activities is zero, i.e. that there are no panel effects, which indicates the use of the least squares method or the OLS method. In this case, the test result indicates that the null hypothesis can be rejected at all standard levels of significance, which means that in this case it is more appropriate to use the “Random” model.

Breusch and Pagan Lagrangian multiplier test for random effects

$$GDPpc[id,t] = Xb + u[id] + e[id,t]$$

Estimated results:

	Var	sd = sqrt(Var)
GDPpc	1.31e+07	3617.471
e	73876.95	271.8031
u	3551193	1884.461

Test: Var(u) = 0

chibar2(01) = 62.26  
Prob > chibar2 = 0.0000

Furthermore, the Hausman test helps in choosing a “Fixed” or “Random” model. It tests whether the errors are correlated with the regressors. The null hypothesis assumes that the errors are not correlated with the regressors, which would indicate the use of the “Random” model. As in the specific case the result of the Hausman test could not obtain positive test values, the test proved to be inappropriate, but the Fixed model was chosen, given the greater significance of the same.

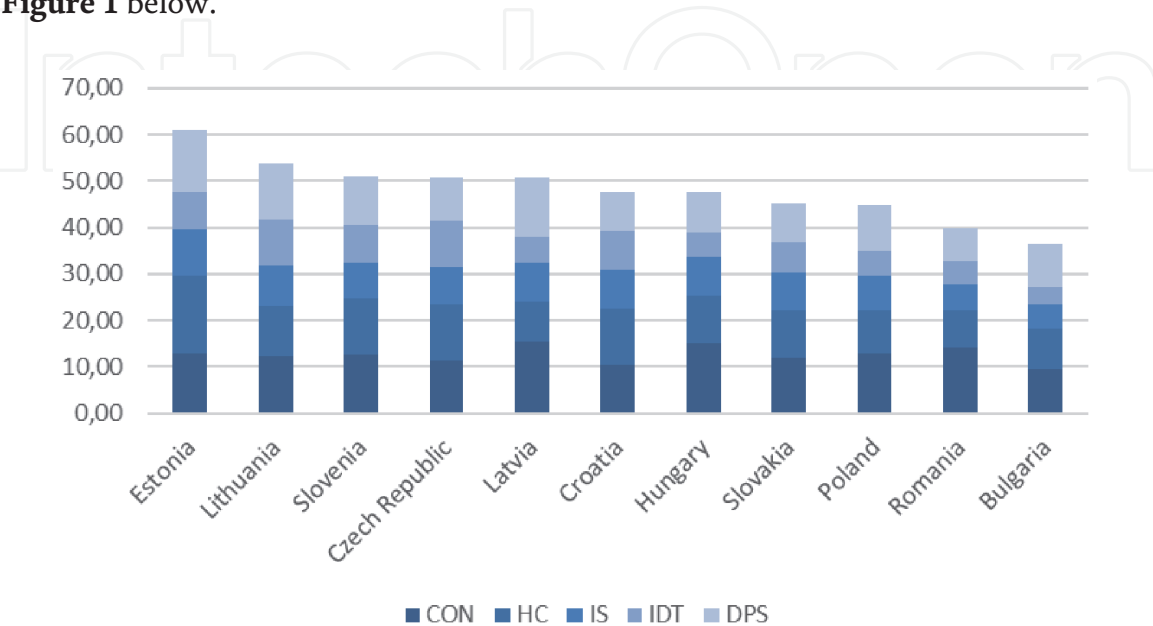
The null hypothesis of the Breusch-Pagan / Cook-Weisberg test indicates homoskedasticity in the model. Given that in this case the null hypothesis cannot be rejected at all standard levels of significance, it is concluded that there is no heteroskedasticity in the model.

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of GDPpc

chi2(1) = 2.62  
Prob > chi2 = 0.1059

Descriptive data analysis is presented below.

The intensity of the individual components of the DESI index is shown in **Figure 1** below.



**Figure 1.**  
DESI index components in CEE countries in 2019 [21–25].

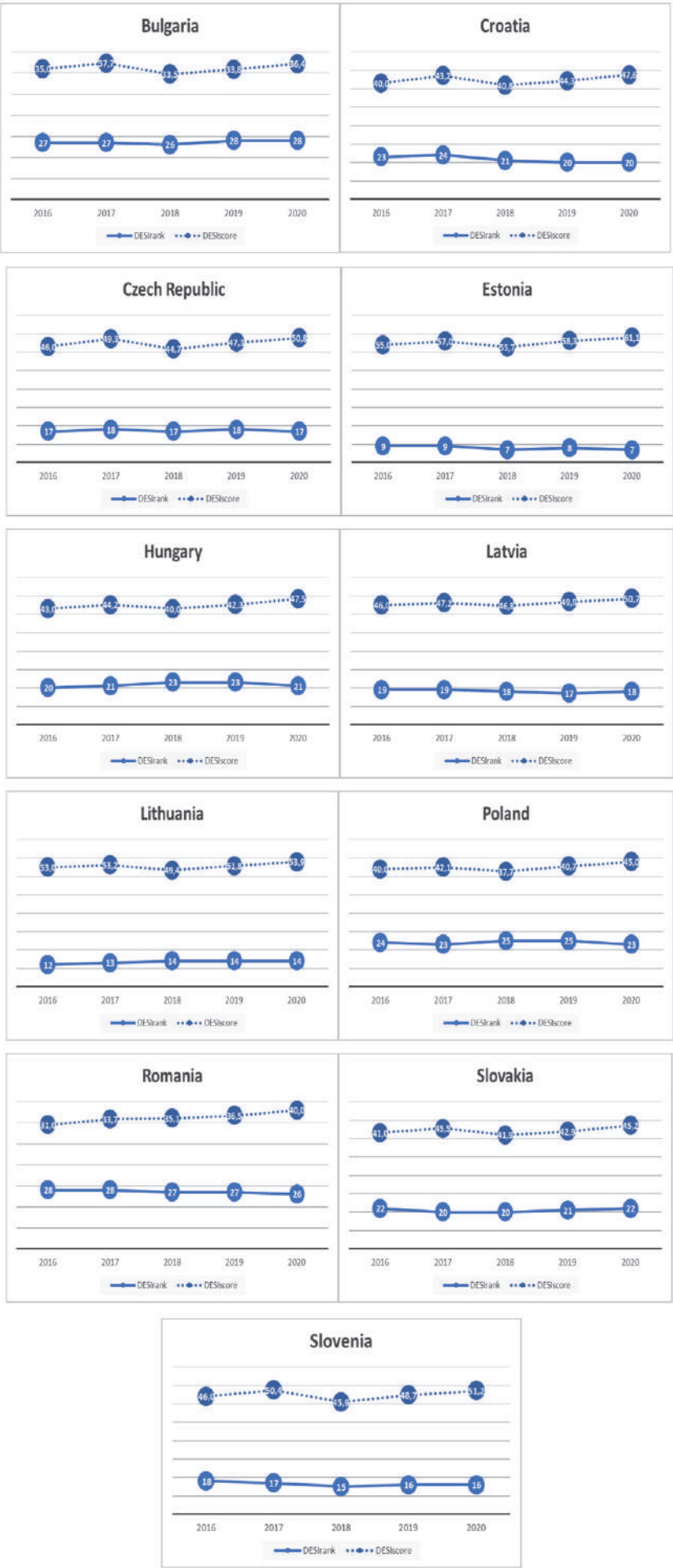


Figure 2.  
DESI rank and DESI score in CEE countries, 2016–2020 [21–25].

Country	GDPpc (EUR)	Rank	CON	HC	IS	IDT	DPS
Bulgaria	6.800	28	38,5	33,9	36,6	17,9	61,8
Croatia	12.480	20	41,2	49,2	55,5	41,5	55,8
Czech Republic	18.000	17	44,9	48,6	54,1	49,6	62,4
Estonia	15.670	7	51,9	66,7	65,4	41,1	89,3
Hungary	13.180	21	59,8	41,8	55,9	25,3	57,8
Latvia	12.490	18	61,8	35,0	54,0	28,3	85,1
Lithuania	13.880	14	48,9	43,8	57,3	49,5	81,4
Poland	12.980	23	51,3	37,3	49,6	26,2	67,4
Romania	9.130	26	56,2	33,2	35,9	24,9	48,4
Slovakia	15.890	22	47,5	41,8	53,4	32,6	55,6
Slovenia	20.490	16	50,2	48,3	51,7	40,9	70,8

**Table 5.**  
*DESI index components for 2020 [21–25].*

DESI ranks and DESI scores for CEE countries from 2016 to 2020 individual countries’ reports are shown on **Figure 2** above. Data from 2020 reports refer to the year 2019. Each country has an increase in DESI score in 2020 (2019) compared to 2016 (2015). Croatia has made the most progress, jumping from 23rd to 20th place within the European Union, while Bulgaria, Hungary and Lithuania have fallen behind in the rankings. The biggest negative shift was made by Lithuania, moving from 12th to 14th place within the European Union. The Czech Republic and Slovakia are countries that have not had a shift in the DESI scale.

DESI index components for 2020, DESI 2020 rank and countries’ GDP per capita are presented in **Table 5** above. Data from 2020 individual countries’ reports refer to the year 2019.

## 6. Conclusions

Digital technologies provide tremendous growth opportunities. The corona crisis has changed the business of almost every entrepreneur. This crisis has shown how important it is to switch from analogue to digital business. The way of doing business had to change literally overnight. All business processes had to be organized differently in uncertain moments where instructions and notifications were received almost hour by hour. A quick adjustment and constant communication with all stakeholders are more important than ever.

Panel method used on 2015–2019 data for 11 CEE countries showed that use of Internet services (people who have never used the Internet; Internet users; news; music, videos and games; video on demand; video calls; social networks; doing an online course; banking; shopping; and selling online), Integration of digital technologies (electronic information sharing; social media; big data; cloud; SMEs selling online; e-Commerce turnover; and selling online cross-border) and Digital public services (e-Government users; pre-filled forms; online service completion; digital public services for businesses; and open data) have positive significant effect on GDP per capita. Other three components of the DESI index did not prove to be significant – DESI rank (positive sign), Connectivity (negative sign) and Human capital (positive sign). Among the significant variables, in the Use of internet

services Estonia Lithuania and Hungary have the most active internet users, while Romania, Bulgaria and Poland are the least active. The top performers in Integration of digital technology are Czech Republic, Lithuania, and Croatia. At the other end of the scale are Bulgaria, Romania, and Hungary. Estonia, Latvia, and Lithuania are top performers in Digital public services while on the other hand, Romania, Slovakia, and Croatia score the lowest.

It is extremely important to continuously implement the digital transformation of the economy. The digital transformation starts with the intention to introduce digital technologies in all parts of society, among the population, in companies, in government institutions, infrastructure and more. The introduction of digital transformation implies not only hardware and software adaptations, but also education of the population, business owners and employees in order to make the best use of the opportunities provided by new technologies such as Internet of Things, Big Data, blockchain, machine learning or artificial intelligence (AI).

Digitization is currently the most important economic reform. It remains for the Member States as well as the European Commission to adopt and implement digitization programs and to provide the financial capacity to support the digital transformation and building of the digital society.

## **Conflict of interest**

The author declares no conflict of interest.


## **Author details**

Mladen Turuk

Faculty of Economics and Business, University of Zagreb, Zagreb, Croatia

\*Address all correspondence to: [mturuk@efzg.hr](mailto:mturuk@efzg.hr)

## **IntechOpen**

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 



## References

- [1] Paunov, C. and Planes-Satorra, S. (2019) How are digital technologies changing innovation? Evidence from agriculture, the automotive industry and retail, *OECD Science, Technology and Industry Policy Papers*, No. 74, July 2019, OECD Publishing
- [2] Satalkina, L., Steiner, G. (2020) Digital Entrepreneurship and its Role in Innovation Systems: A Systematic Literature Review as a Basis for Future Research Avenues for Sustainable Transitions, *Sustainability*, 2020, 12, 2764; doi:10.3390/su12072764, pp. 1-27
- [3] Gohmann, S.F. (2012) Institutions, latent entrepreneurship, and self-employment: an international comparison. *Entrep. Theory Pract.*, 36(2), pp. 295–321
- [4] Giones, F., Brem, A. (2017) Digital technology entrepreneurship: a definition and research agenda. *Technol. Innov. Manag. Rev.*, 7(5), pp. 44–51
- [5] Antonizzi, J., Smuts, H. (2020) The Characteristics of Digital Entrepreneurship and Digital Transformation: A Systematic Literature Review, Conference on e-Business, e-Services and e-Society I3E 2020: Responsible Design, Implementation and Use of Information and Communication Technology, pp 239–251, [online]: [https://link.springer.com/chapter/10.1007/978-3-030-44999-5\\_20](https://link.springer.com/chapter/10.1007/978-3-030-44999-5_20) (Accessed: 25 November 2020)
- [6] van Welsum, D. (2016) Enabling Digital Entrepreneurs, *World Development Report 2016 Digital Dividends*, World Bank Group, pp. 1–12 [online]: <http://pubdocs.worldbank.org/en/354261452529895321/WDR16-BP-Enabling-digital-entrepreneurs-DWELSUM.pdf> (Accessed: 25 November 2020)
- [7] Davidson, E. & Vaast, E. 2010. Digital Entrepreneurship and its Sociomaterial Enactment. Paper presented at 43rd Hawaii International Conference on System Sciences (HICSS), 5–8 January 2010.
- [8] Hull, C., Hung, Y-T. & Hair, N. (2006): Digital Entrepreneurship, EDGE, Rochester Institute of Technology – RIT Scholar Works, pp. 1–29
- [9] European Commission (2015), Digital Transformation of European Industry and Enterprises: A report of the Strategic Policy Forum on Digital Entrepreneurship [online]: <http://ec.europa.eu/DocsRoom/documents/9462/attachments/1/translations/en/renditions/native> (Accessed: 15 May 2020)
- [10] Hair, N., Wetsch, L. R., Hull, C. E., Perotti, V., & Hung, Y-T. C. 2012. Market Orientation in Digital Entrepreneurship: Advantages and Challenges in A Web 2.0 Networked World. *International Journal of Innovation and Technology Management*, 9(6) DOI: 10.1142/S0219877012500459
- [11] Qu, J., Simes, R. and O'Mahony, J. (2016) How Do Digital Technologies Drive Economic Growth?, [online]: <https://esacentral.org.au/images/QuJSimesROMahonyJ.pdf> (Accessed: 3 September 2020)
- [12] Turuk, M. (2018) The Importance of Digital Entrepreneurship in Economic Development, 7th International Scientific Symposium: Economy of Eastern Croatia – Vision and Growth, Masek Tonkovic, Anka and Crnkovic, Boris (ed.), Osijek: Sveuciliste Josipa Jurja Strossmayera u Osijeku, Ekonomski fakultet u Osijeku, pp. 178–186

- [13] Hernandez, K., Faith, B., Prieto Martin, P. and Ramalingam, B. (2016) The Impact of Digital Technology on Economic Growth and Productivity, and its Implications for Employment and Equality: An Evidence Review, *IDS Evidence Report 207*, Brighton: IDS
- [14] Minges, M. (2015) Exploring the Relationship between Broadband and Economic Growth, *Background paper for the World Development Report 2016*, Washington DC: World Bank
- [15] Koutroumpis, P. (2009) The Economic Impact of Broadband on Growth: A Simultaneous Approach, *Telecommunications Policy*, 33.9, pp. 471–85
- [16] Czernich, N., Falck, O., Kretschmer, T. and Woessmann, L. (2009) Broadband Infrastructure and Economic Growth, *CESIFO Working Paper 2861*, Center for Economic Studies Information and Forschung (CESIFO), University of Munich
- [17] Garcia Zaballos, A. and Lopez-Rivas, R. (2012) Socioeconomic Impact of Broadband in Latin American and Caribbean Countries, Inter-American Development Bank Technical Note 471
- [18] Qiang, C.; Rossotto, C. and Kimura, K. (2009) ‘Economic Impacts of Broadband’, in *Information and Communications for Development 2009: Ex-tending Reach and Increasing Impact*, Washington DC: World Bank
- [19] Scott, C. (2012) ‘Does Broadband Internet Access Actually Spur Economic Growth?’ [online]: [https://colin-scott.github.io/personal\\_website/classes/ictd.pdf](https://colin-scott.github.io/personal_website/classes/ictd.pdf) (Accessed: 8 September 2020)
- [20] Eurostat (2020) [online], <https://ec.europa.eu/eurostat/data/database> (Accessed: 15 May 2020)
- [21] European Commission (2020) Digital Economy and Society Index (DESI) 2020 Thematic chapters, [online]: [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=67086](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=67086) (Accessed: 12 September, 2020)
- [22] European Commission (2016) DESI 2016 Country Profiles, [online]: <https://ec.europa.eu/digital-single-market/en/news/desi-2016-country-profiles> (Accessed: 12 September, 2020)
- [23] European Commission (2017) Digital Economy and Society Index (DESI) 2017, [online]: <https://ec.europa.eu/digital-single-market/en/news/digital-economy-and-society-index-desi-2017> (Accessed: 12 September, 2020)
- [24] European Commission (2018) Digital Economy and Society Index (DESI) 2018 Report, [online]: <https://ec.europa.eu/digital-single-market/en/news/digital-economy-and-society-index-2018-report> (Accessed: 12 September, 2020)
- [25] European Commission (2019) Digital Economy and Society Index (DESI) 2019, [online]: <https://ec.europa.eu/digital-single-market/en/news/digital-economy-and-society-index-desi-2019> (Accessed: 12 September, 2020)

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# The Paradox of Indonesian Digital Economy Development

*Vience Mutiara Rumata and Ashwin Sasongko Sastrosubroto*

## Abstract

In line with the rapid growth of the global e-commerce industry today, Indonesia has enormous digital economic potential in the future. The Indonesian government is focusing on developing the digital economy by increasing the connectivity infrastructures as well as the local market. Nevertheless, there are some paradoxes caused by the existing regulations. This paper elaborates on the paradox of digital economy development in Indonesia. By using a mainstream-approach policy analysis method, this study describes the problematic situation of Indonesian digital economy governance. This is a qualitative study where the primary data derive from mostly statutes, government official documents, as well as reports. The discussion consists of (1) e-commerce: the main driver of Indonesian Digital Economy; (2) Indonesian Digital Regulatory Framework and Challenges; and (3) The Paradoxes of Indonesian Digital Economy. Due to various sectors of the digital economy, the discussion focuses on the e-commerce sector.

**Keywords:** digital economy, e-commerce, paradoxes, Indonesia

## 1. Introduction

Indonesia's digital economy is under the spotlight. Many studies have confirmed its potential in the future. A study report launched by Google and Temasek/Bain in 2019 states that Indonesia's internet economy grows in high-speed which estimated at 40 billion USD in 2019 and vigorously on track to reach 130 billion USD by 2025 [1]. At the regional level, the internet economy value in the South East Asia region reaches 100 billion USD in 2019 and would increase tripled by 300 billion USD in 2025 [1].

In order to boost the digital economy, the Ministry of Communication and Informatics of the Republic of Indonesia (the MCI) has embarked a national information and communication technology (ICT) infrastructure development. The Telecommunication and Information Accessibility Agency (BAKTI), the MCI's public service agency, launches "Merdeka Sinyal 2020" meaning Independent Signal 2020, which is a program to provide telecommunication access in 5000 frontier, outermost, and underdeveloped areas or known as "3T areas" in 2020 [2]. In addition, the Palapa Ring project was officially inaugurated and operated in October 2019 [3]. The Palapa Ring project is a telecommunication network development project that connects 514 districts/cities in Indonesia, which consists of Marine Cable and Fiber Optic Communication system development. This project was initiated in 2005 but the construction was started in 2016. All these activities are part of the effort to fulfill the Indonesian agreement as a member of WSIS. Besides



the Palapa Ring project, the government also enhances the postal logistics infrastructure in the 3T areas through the Postal Service Obligation (PSO) program. This program is Indonesian government commitment as a member of the UN body—Universal Postal Union (UPU).

Nevertheless, the supporting programs for a sustainable digital economy remain in question. In 2018, the MCI launched a “1000 digital startup” national movement which mainly was a coaching program for future *technopreneurship* in 10 cities including Jakarta [4]. This coaching program consists of several phases: ignition (seminar to increase knowledge to become a *technopreneur*), workshop, hackathon (aims to develop a prototype or software or apps), and Bootcamp. Unfortunately, this program has missed the participants’ target. Instead of participants who have interest and idea to build digital applications, the participants who registered to the 1000 start-up digital web was regular young people who are “curious” and had less commitment about this program [4]. Immediate improvement is necessary or the idea to create new digital *technopreneurs* would be in peril.

McKinsey, in its 2016 report, states that “Indonesia has a long way to go in the digital age” [5]. There is a paradox that the country might not be able to embrace the benefits of modern technology. Although daily internet usage is considerably high, the level of literacy remains to lag behind compared to some other countries in the Southeast Asia region. The digital technology may drive the national economy in a country, but this should be critically assessed particularly in Indonesia’s case. The growing e-commerce apps usage in Indonesia has a direct impact on imported consumer goods [6]. In the meantime, Indonesia is still far behind in terms of digital competitiveness. According to the IMD World Digital Competitiveness Ranking 2019, Indonesia ranks 56 out of 63 countries (Knowledge ranks 56, Technology ranks 47, and Future Readiness ranks 58) [7].

This paper elaborates the paradox of digital economy development in Indonesia. The mainstream-approach policy analysis method is used in order to describe the problematic situation of Indonesian digital economy governance. This is a qualitative study where the primary data derive from mostly statutes, government official documents, as well as reports. The discussion consists of (1) e-commerce: the main driver of Indonesian Digital Economy; (2) Indonesian Digital Regulatory Framework and Challenges; and (3) The Paradoxes of Indonesian Digital Economy. Due to various sectors of the digital economy, the discussion focuses on the e-commerce sector.

## 2. Mainstream policy analysis

A policy analysis basically is about defining the issues, formulating and implementing the policies to address those issues. Policy-making is a complex process. It involves a wide range of elements of the State in the formulation process, as well as a wide range of impacts in the implementation. The complexity of the policy-making process may need effective policy analysis techniques. There is a suggestion that there are two primary domains of policy analysis: by looking at the process and the content [8]. The process may involve the network of involved actors at the local, national, and even international levels. The content may specifically depend on the issues, context, problems, scope, as well as regulative products and output. The common sense about policy analysis is that a policy process is a political process. In terms of the policy analysis method, it is decided to start by defining the orientation of the policy analysis. There are at least three policy research orientations: (1) mainstream, (2) traditional, and (3) interpretative [9].



This study is a mainstream policy research orientation that focuses on the policy process and also the interaction within the governmental networks involved in [9]. Nevertheless, there is a sliced section between mainstream and interpretative policy research orientations. It can be seen in the similarity of data sources and even the focus of the study [8]. There are at least 11 major mainstream methods that can be used in mainstream and also interpretative policy studies [8]. One of these methods is “frame reflexive policy analysis” which is rooted in the notion of “framing” which is generally understood as the way to define and understand reality according to own perspective. Framing, in the policy-making sense, is a way to examine the problematic situation and formulate normative actions to address it [10]. The policy controversies are common as it emerges due to multiple frames and perspectives of the government (i.e., the Executive, the Legislative, the Judicative), the general public, the community, or the social groups in viewing a problematic situation. Nevertheless, there is a two standpoint in judging someone’s frame: (1) positivism which argues that policy controversies can be solved by fact and logic and (2) relativism which argues that each of existing frames is equally valid [10].

The focus of frame reflexive policy analysis can be about the policy discourse, action frames, rhetorical frames, institutional frames, and even meta-cultural frames [8]. A policy discourse helps policy analysts to define the power behind the policy formulation process [11]; the emerging problematic situation and multidimensionality policy concerns from a media perspective [12].

This focus of policy analysis in this paper is defining the existing discourses about the Indonesian digital economy particularly the e-commerce industry. It aims to understand the complexity of digital economy governance and its impact on creating paradox situations. We conduct a document study as a data-gathering method. The documents mainly are statutes (i.e., the Presidential decree, the Ministerial Regulation, the Government Regulation, and other related regulative documents). This study does not describe the political condition or power that influences policy implementation.

### **3. E-commerce: the main driver of Indonesian digital economy**

#### **3.1 The digital economy in global trend**

There are some terminologies to describe today’s new economy: digital economy, attention economy, internet economy, knowledge economy, or network economy, which sometimes are used intertwined. In this paper, we use “the digital economy” terminology. Apparently, the digital economy is industry 4.0’s primary fuel. Industries, governments, and societies are adjusting themselves to this ever-changing business model which disrupts the old fashion one. Many companies have integrated digital technology to provide better products. Meanwhile, the government has integrated digital technology to provide better policies. Nevertheless, the effort to get the best benefit of the digital economy is still challenging.

The cores of the digital economy are the internet and digitization. The better utilization of these cores, the better the product produced and even the more profit gained. This can be seen from big technological companies particularly based in the United States. They are likely to control all of the digital business lines which at the end will inevitably monopolize the global market. The key element of the monopoly denomination is the company growth itself and its ability to make sure its customers continue to use or stick to its products [13]. Google, for example, spent billions of USD to conjure the company not only as of the leader of a search engine in the

world, but also to the leader of “one-stop online activity” kind of apps (email, communication apps, video-sharing, file storage, word-processing service, and so on). In addition, it aggressively reconstructs its position on the internet infrastructure to keep pace with technology.

The Digital Economy, in general meaning, is an economic activity by using digital and computing technologies. The Internet has evolved to provide basic infrastructure for the digital economy. Nevertheless, the impact of this digital economy is not merely just a business or economy, but also social, cultural, politics, and many other facets of human life. Tapscott argues that the digital economy is the economy of “the Age of Networked Intelligence.” He warns the dark side of this era that includes (1) dislocations (many old jobs will have perished); (2) privacy threat (the personal data breaches); (3) polarization of wealth (20% of household worth 80% of country’s wealth); (4) digital gap among society; and also (5) digital slave (technology invades every part of human time and space) [14]. Therefore, government policies should ensure that technology should not create these negative effects, but to serve people.

### **3.2 Indonesian e-commerce highlights**

It is internationally acknowledged that Indonesia has a great digital economy potency. In the 2018 Frost & Sullivan 2018 White Paper, it is mentioned that the digital service industry in Indonesia will increase significantly with a value up to 9528.4 million USD in 2022 [15]. In the region, Indonesia’s internet economy—along with Vietnam—will enjoy 40% growth rate annually which is bigger than Singapore, Malaysia, Thailand and the Philippines [1]. The MCI projects that in 2020, the digital economy in Indonesia can grow 130 billion US dollars or around IDR 1700 trillion, 20% of Indonesia’s total GDP [16].

The growth of digital start-ups in Indonesia can be traced back to 2010. The ride-hailing start-up Gojek was established in 2010. Some of the start-ups in that year have high involvement of foreign investors. Yahoo, for example, acquired Koprol, the Indonesian online social networking service, in May 2010 [17]. Today, some of Indonesia’s digital start-ups show expansion at the global level. Gojek is classified as “Decacorn” which has 10 billion USD valuations [18]. Gojek was the first local start-up that earned this classification along with other 21 companies globally. Following Gojek, the leading Indonesian e-tailing start-up Tokopedia has 7 billion USD valuations. It is predicted that it will get the “Decacorn” title within 2–3 years. These two start-ups have contributed significantly to national economic growth. Tokopedia contributed 58 trillion IDR or 4.1 billion USD to the Indonesian national economy in 2018. The contribution is predicted to grow up to 170 trillion IDR or around 12 billion USD in 2019 [19]. With over 90 million active users, Tokopedia has provided around 3 million new jobs in 2018, while Gojek contributed around 44.2 trillion IDR or 3.13 billion USD to the Indonesian national economy in 2018 [20].

E-commerce remains the star of the digital economy in Indonesia. The Morgan Stanley study finds that Indonesia’s e-commerce market size reaches 13 billion USD in 2018 or has grown 50% each year for the last 2 years [21]. This increasing market size may be driven by the increase in internet access and usage. According to APJII’s 2019 report, there are at least 171.17 million internet users or around 64.8% of total populations [22]. According to We are Social January 2019 report, the average of Indonesian internet users’ daily time spent online is 8 hours and 36 minutes while time spent on social media is 3 hours and 26 minutes. The same report shows there are at least 107 million people (40% of the total population) purchase consumer goods through e-commerce platforms. This number is predicted to grow

continuously due to the speed race of mobile gadget penetration. The annual sales revenue of consumer goods on e-commerce reaches 9.5 billion USD or 41 USD per capita [23]. These data show how lucrative the e-commerce market in Indonesia. There are some factors that influence the growth of e-commerce volume in Indonesia, such as (1) the increasing income per capita; (2) the increasing of various companies in e-commerce industry; (3) the expansion of telecommunication infrastructure and internet access particularly in rural areas; and (4) the changing of consumers' behavior from "offline" to online shops. Indonesia's economy tends to endure amid the uncertain global economic turbulence. The economic growth in 2018 reached 5.17% or increased from 5.07% from 2017 with GDP per capita reaches 3927 USD or 56 million IDR [24]. Even so, Indonesia is still considered as a "middle-income trap" country since the GDP per capita less than 4250 USD.

The variance of existing e-commerce business model in Indonesia is as follows: (1) Classified Ads/listing (e.g., olx.co.id, Berniaga, FJB-Kaskus); (2) Marketplace (e.g., Tokopedia, Bukalapak, Lamido); (3) Shopping mall (e.g., Matahari Mall); (4) B2C online shop (e.g., Berrybenka, Zalora, Lazada, Sociolla); and (5) Online shops on social media (e.g., Facebook, Instagram) [25]. These business models connect three sectors, which are the government, business, and costumers, indirect and interactive ways. But, to build e-commerce platforms requires exhaustive resources. It needs high-performance infrastructures, a huge amount of capital and investment, and even high skilled human resources. The availability of these resources is relatively rare in developing countries such as Indonesia and so this country is still dependent on developed countries. In addition, the advancement of the digital economy may lead to job replacement which requires more technology than human resources. The existing policies and regulations should not only ensure the growth of the digital economy industry but also to address these critical issues.

## **4. Indonesian digital regulatory framework and challenges**

### **4.1 The Indonesian digital governance**

Although Indonesia's digital economy is likely to grow in the future, there is no grand design or roadmap of digital economy development yet. Currently, however, the Indonesian government is drafting the national digital economy strategy [26]. This draft aims to address the upcoming challenges of the digital economy which has not been covered by the existing roadmap of e-commerce 2017–2019 through the enactment of the Presidential decree number 74 year 2017. The existing e-commerce roadmap determines the admission of e-commerce steering committee which consists of inter-sectoral government collaboration to implement at least eight primary programs, which are:

1. Funding which includes: Crowdfunding, SMEs business credits for the digital platform, Angel capital, Seed Capital, and Grants for start-ups.
2. Tax incentive for local investors as well as e-commerce start-ups especially with a turnover of IDR 4.8 billion per year. Also, the availability of equal tax regulation applied both domestic and foreign e-commerce entrepreneurs.
3. Consumer protection includes the development of national payment gateways as well as harmonization in regulatory level for electronic certification, accreditation process, payment mechanism policies, protection of consumers and e-commerce industry, and dispute resolution schemes.

4. Education and Human Resource that includes an incubator program, e-commerce awareness campaigns, and education.
5. Logistics includes the development of a national logistics system, revitalization of the state owned Post enterprise as well as the development of outsourcing of e-commerce logistics facilities.
6. The development of broadband networks throughout Indonesia regions.
7. Conducting a national supervision system model in e-commerce transactions.
8. The Establishment of collaborative and systematic management to accelerate the implementation of e-commerce roadmap [27].

The primary law of internet regulation in Indonesia is Law number 16 year 2019 (amendment of the law number 11 year 2008) on the electronic information and transactions (*Undang-Undang Informasi dan Transaksi Elektronik* or the ITE Law). The President will issue the Government Regulation (*Peraturan Pemerintah* or PP) to implement the Law. The PP to implement the UU ITE is the PP number 71 year 2019 (amendment of PP number 82 year 2012) on the Electronic System and Transaction Management. This PP regulates the global and local Electronic System and Transactions providers which operate in Indonesia, to:

1. Register their service to the Minister of Communication and Informatics.
2. Place the data center and data recovery center in Indonesian territory.

Nevertheless, the revised PP is relenting particularly on the global providers' requirement to place their data center and data recovery center in Indonesian territory [28].

The PP mandates that e-commerce is considered as "strategic electronic system which has a serious impact on public interest and service." Henceforth, the regulation on e-commerce should be carefully taken since this industry is open to global competition. Recently the Indonesian government has enacted PP number 80 year 2019 on electronic-based commerce (*Perdagangan Melalui Sistem Elektronik/PP PMSE*) after long-standing public debate and discussion. However, the PP has no significant difference with existing PP 71 year 2019 which obliges both local and foreign e-commerce business doers (B2B, B2C, C2C, G2B) to meet these requirements such as:

1. Using Indonesian ".id" Top Level Domain address for the website.
2. Using Internet Protocol Address according to the law.
3. Placing data center according to the law.
4. Registering the services to the authority according to the law.
5. Meeting the technical standard as well as having certificate of reliability that has been issued by the authority.
6. Complying other sectoral regulations that relate to the electronic based commerce [29].



In addition, the PP number 80 year 2019 mandates the local and global e-commerce platforms to have an Electronic based Reliability Certificate which issued by the Electronic Certification Provider (*Penyelenggara Sertifikat Elektronik* which commonly known as Certification Authority/CA). The CA is a legal subject that functions as a trustworthy third party that facilitates online transaction security systems with Digital Signature and Public Key Encryption, and also issues a quite range of digital certificate services that includes:

1. Examination of prospective Electronic Certificate holders.
2. Issuance of Electronic Certificates.
3. Validation and Extended Validation of Electronic Certificates.
4. Digital Certificate Revocation.

Based on the Ministerial of Communication and Informatics Regulation number 11 year 2018, this CA should get acknowledgment from the MCI based on three levels: registered, certified, and rooted [30]. By this digital certificate, the identity and legal status of the owner of the signature are cleared and ensured so that it may guarantee the online transactions. Nevertheless, whether this PP would be able to force global internet-based application and content services providers to comply with the Indonesian law remains unclear.

Some existing regulations are obsolete and seem unable to regulate the digital economy sector. Hence, the regulation to protect e-commerce customers remains unclear. The law number 8 year 1999 on Consumer Protection is insufficient to protect consumers in doing e-commerce transactions. For instance, the law mandates the consumers' rights to obtain comfort, security, and safety in using or consuming the goods and/or services [31]. In e-commerce, the provision of the right to obtain comfort may be impeded due to the absence of a physical place where consumers can see, touch, feel and even taste the products before buying. The provision of the right to obtain security on e-commerce transactions is another issue. Hence, the existence of security standards on e-commerce in Indonesia is also questionable.

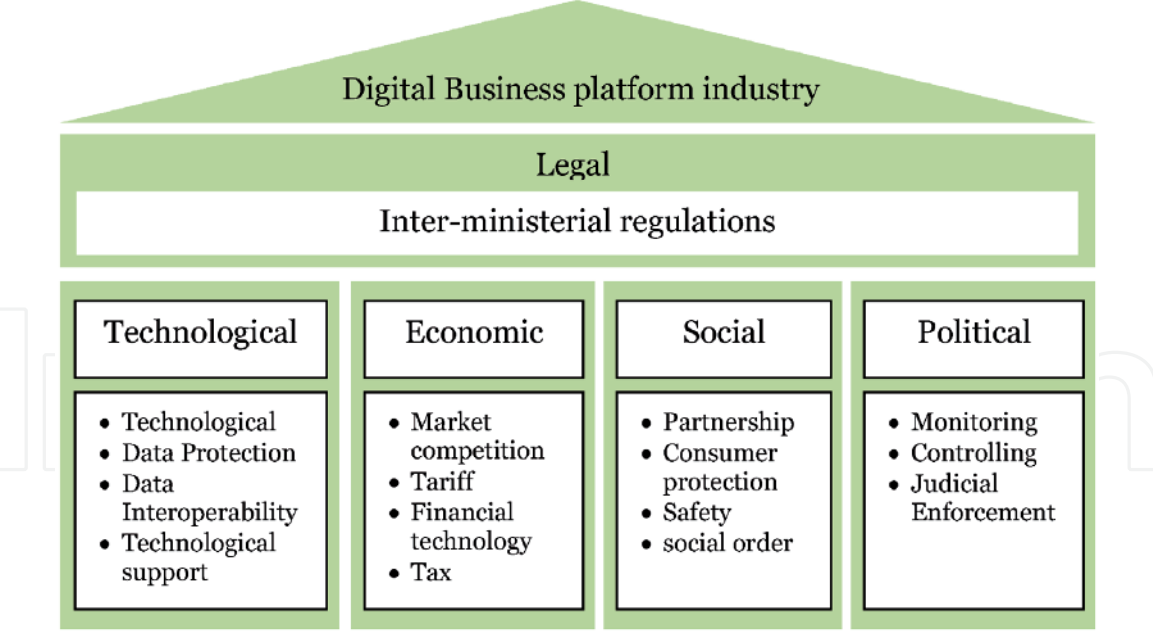
## **4.2 The challenges of the Indonesian digital governance**

The Research and Human Resource Development department of the Ministry Communication and Informatics (the MCI) proposes the digital platform based regulatory framework particularly in online transportation (including ride-hailing start-up). In **Figure 1**, the digital business platform industry involves several facets: technology, economic, social and politics. Hence, the legal aspect of this industry should embrace what extends the impact on these facets [32].

Nevertheless, the legal issue that emerged, regarding of digital business platform industry, is the inter-ministerial regulation that causes partial legal implementation and authoritarian. In the online transportation case, the MCI is authorized to regulate the digital platform including the company registration, while the Ministry of Transportation is authorized to regulate the safety and service aspects of public transportation. Therefore, it is suggested that the MCI should be the initiator in issuing comprehensive digital platform business regulations.

A similar issue also occurs in e-commerce industry. The practice of inter-ministerial regulation may be challenging particularly in the dynamic environment such as the digital economy. Some regulations concerning e-commerce: the Law number 7 year 2014 on Trade (authorization in the Ministry of Trade), the law number 10





**Figure 1.**  
*Digital business platform regulatory framework.*

year 1998 on Banking (authorization in the Central Bank), the Law number 25 year 2007 on Capital Investment (authorization in the Capital Investment Boarding Body), the Law number 20 year 2008 on Micro, Small and Medium Enterprises (authorization in the Ministry of Micro, Small and Medium Enterprises), the Minister of Finance Regulation number 112 year 2018 (authorization in the Ministry of Finance), the Law number 38 year 2009 on Postal and the Law number 16 year 2019 on ITE (authorization in the MCI). These laws have a different legal scope so that they might not be enforced comprehensively in the collision sector as digital economy. Currently, there is no single law on the digital economy. However, the President's new proposal on Omnibus Law for several activities a few months ago might be used to set up a single and more supportive law on e-commerce. It should be noted that laws can be initiated either by the Executive and/or the Legislative. Even so, the process to enact a law would take some time if there is a fierce debate between the Executive, Legislative as well as industry. The feasibility of one Omnibus Law on digital economy law needs further study.

Another challenge of digital economy regulation is absent in current regulations of upcoming digital economy issues such as personal data protection and cross border e-commerce transactions. The bill of personal data protection is still an ongoing discussion between the Legislative and the MCI. In the meantime, the regulation for cross border e-commerce transactions is quite challenging. The Central Bank (Bank Indonesia/BI) and The Ministry of Finance are developing the data integration system to monitor cross border e-commerce in Indonesia [33]. The question remains whether this system is sufficient to address cross border issues and needs further study.

The formulation of taxation particularly for global e-commerce providers is conflicting among the authorities. The MCI may loosen the obligation to place a data center in Indonesia territory while the Ministry of Finance will pursue the legal status of the global company as Indonesian taxpayers. This is one of the paradoxes that will be discussed more in the next subchapter.

Whether the Digital economy should or should not be regulated, the regulation policy of the digital economy remains challenging for regulators all over the world. The government may be facing a dilemma situation. The MCI explicitly will less

regulate the digital economy sector in order to create a business-friendly environment [34]. In order to do this, the Ministry has simplified 36 permitting regulations into five regulations for the industry. But the ultimate goal of regulation is to create a conducive environment for local e-commerce platform providers to grow and be able to compete at the global level which at the end will contribute more to the national economy.

## **5. The paradoxes of Indonesian digital economy**

The regulation on the digital economy may be influenced by both domestic and international regulatory frameworks. In e-commerce case, the existing regulations both national and international level would potentially create the paradox which furtherly discussed below.

### **5.1 The regulative paradox**

A good regulation is the one that focuses on the goal which may be addressing certain issues or problems. The paradox of regulation emerges when it does not have an appropriate level of enforcement by the government itself or other relevant stakeholders [35]. The law enforcement does not solely depend on the government, but more to the governance with the involvement of various actors outside of the state to exercise a certain level of control. By this governance paradigm, well-defined and focused goals regulation is needed. To achieve this kind of regulation is not simple since it involves many parties with different interests. The law enforcement remains the biggest challenge to regulate the application and content product providers, particularly to create an equal level of playing field between local and global electronic transaction and system providers especially in the e-commerce industry [36]. The local e-commerce business companies have to comply with domestic regulations, whereas these domestic digital laws seem to do not applicable to global e-commerce companies. Permanent Status registration is the salience issue.

Regulating the digital industry is challenging for the regulators in particular by defining who and how the regulation should be. There are three strategies to regulate the data-driven digital platform according to the European Commission: (1) command-and-control regulation; (2) self-regulation, and (3) co-regulation [37]. A first strategy is a top-down approach where regulation is legal legislation with sanction backup. This strategy, however, may not fit the digital platform industry due to three reasons: (1) it may potentially obstruct innovation and harm the platform provider; (2) the enforcement of the rules may not easily be borne; and (3) the regulation may add more drawbacks for the existing complex issues. Hence, the top-down approach legislation relies on well-informed, well-educated, specially trained regulatory officials. The second strategy is self-regulation which means that regulation lies in the hand of industry. The regulations are defined and enforced collaboratively among the players within the industry.

This strategy may be fit too since the digital platform providers need to be independent with less and relatively no bureaucratic interference for technological adoption and innovation. Nevertheless, the self-regulation mechanism can be mandated by the public authorities to set up a specific standard in the industry. The last strategy is co-regulation which collaboration between government and non-government (private) sectors with distinctive role and task to achieve public policy objectives. The last strategy is considered as the best regulatory approach to regulate the digital platform industry. The public authority set up the objectives, while the

mechanism to achieve these objectives lies on the hand of the private sectors [37]. Thus, co-regulation is also considered as “regulated self-regulation” which acquires reciprocal actions between the regulators and the regulated ones.

The Indonesian government seems to adopt co-regulation strategy to regulate the digital economy. The Indonesian e-commerce platform providers are committed to support the government’s digital economy sectors programs. *Tokopedia*, for instance, expands its services for tax payment gateway, e-government, as well as e-passport and e-ID by collaborating with several city governments in Indonesia [38]. In addition, *Tokopedia* drives local small medium enterprises to do business globally. The average of seller growth on *Tokopedia* reaches 150.4% annually where 86.5% of it is new sellers [19]. Even so, to what extent that the industry determines the formal regulation on digital economy is questionable. In other words, the co-regulation scheme between the government and the private sector in order to face the global competitive challenges remains unclear. In some sector, the co-regulation is clearer. PANDI, for example, as non-government organization is authorized by the government to regulate the Indonesian top level Domain (.id) except second level military (.mil) and governmental domain (.gov). The fact that the government tries to increase internet access as well as logistic infrastructures to support national digital economy, the growing of e-commerce marketplace in Indonesia, however, may potentially harm domestic industries particularly small and medium enterprises. The Ministry of trade at that time claimed that 90% of goods traded on e-commerce marketplace are imported goods [39]. This is contrary to the what is mandated in the PP number 80 year 2019 article 12 that both global and local to prioritize domestic products exchanged on the platform as well as to increase the competition level of domestic products [29]. The government has urged the marketplace providers such as *Tokopedia*, *Shopee*, *Bukalapak*, *Blibli.com* and *Blanja.com* to increase the local products proportion but without clearer regulation [39]. This poses potential threat and should be addressed in further study.

The diminishing of physical space, as the impact of e-commerce, has threatened the sustainability of physical shop, both in modern and traditional market, in the future. The regulations for physical shop are tighter than e-commerce shops. The Minister of Trade regulation number 70/M-DAG/PER/12/2013 on the traditional markets, shopping centers and modern shop guidance, mandates the physical shops to sell 80% of domestic products in their shops. However, the implementation of this regulation may be ineffective since the capabilities of stock management varies among retailers and also the absence of clear mechanism that determine the fulfillment of 80% domestic products [40].

The regulations should be made to make the industry grow properly and can compete optimally with foreign competitors. On the other hand, the existing regulations may hinder it by enforcing the law unequally between the local and global e-commerce players which operate in Indonesia. As an example, Facebook’s status in Indonesia will be discussed further. So far, its status is a service company instead of permanent establishment status (*Badan Usaha Tetap* or BUT) [41]. The Facebook’s status in Indonesia is highly questionable whether it is in accordance or not with PP 71/2019, and also three others regulations on Tax which are: the Directorate General of Tax circular letter number 62/PJ/2013 on e-commerce taxation provisions; the Directorate General of Tax circular letter number 04/PJ/2017 on the establishment of BUT for foreign over application and content services providers in Indonesia; as well as the Ministry of Finance Regulation number 210 year 2018 on the e-commerce taxation. It is worth to be noted that the number of Facebook users in Indonesia reach 64.6 million users in 2018 [42]. It means that Indonesia is a lucrative market for Facebook. In addition, Facebook is more than just a social network platform. It has expanded to marketplace platform where its website uses

Indonesian language. Other similar cases with other platforms and social media must also be taken into account. This is a forthcoming challenge for Indonesian e-commerce industry.

Unlike global companies, the local companies should face tight regulations in order to open its business that includes digital economy sector. There are two important legal aspects for local company to enter the digital sector business namely: the subject and the associated impacts. For legal subject, the regulations may include: the Law Number 40 year 2007 on Limited Liability Companies, the Law Number 17 year 2013 on Social Organizations, the Law Number 17 year 2012 on Cooperatives, and the Law number 28 year 2008 on Foundations.

## **5.2 The paradox of IT market growth**

Despite the debate between utopian and dystopian, the global discourse on ICT and its great impacts on national ICT governance and development should be critically assessed. Some studies found the contrast between global vision and the facts in the country [43].

As briefly mentioned above, the nature of e-commerce regulation in Indonesia is also influenced by international agreements, mainly regulations issued by global institutions such as the World Trade Organization (WTO). The negotiation on e-commerce has been initiated by the WTO E-commerce Working Party in 1998. At the beginning of 2019, there were 76 members of WTO, including Indonesia, which represents over 90% of global trade, urged WTO to update the existing rules particularly to address the changing technologies and issues related to e-commerce. Those existing rules such as the General Agreement on Trade in Services (GATS) and the Central Product Classification (CPC) system have not included internet-based services [44]. In the joint statement of 76 partners involved in 2019 talks on e-commerce, the new rules on e-commerce should reflect:

1. “Improve consumers’ trust in the on-line environment and combat spam
2. Tackle barriers that prevent cross-border sales
3. Guarantee validity of e-contracts and e-signatures
4. Permanently ban customs duties on electronic transmissions
5. Address forced data localization requirements and forced disclosure of source code” [45].

The free custom duties on electronic transmissions may be one of the critical topics of discussion at the international forums. The expansion of digital content beyond software and become an integral part of a wide array of distinctive products, goods, and services (game, movie, songs, and others) which then pose challenges particularly for developing countries such as Indonesia. In the 2017 WTO Ministerial agreement, it states that any digital product purchased and transmitted online should be free from custom duties, with no exception. This agreement may be advantageous for foreign producers and local importers. Although it seems to be rather unethical, although it might be legal, the foreign producers may even have the possibility to cut its hardware’s selling price and add it to the software’s selling price. Since the software can be transmitted through the internet, it can be exempted from customs duty [46]. It is to be noted that the Ministry of Finance has revised its regulation through the Minister of Finance



Regulation number 112/PMK.04/2018 on the Import Shipment Goods Provision. Through this revision, the government made adjustments to the minimum value of import duties and taxes in the context of import on shipments from the US \$ 100 to the US \$ 75 per person per day [47].

The 2017 WTO Ministerial agreement added the long-standing effort to liberalize ICT trade among countries. It may be initiated through the Information Technology Agreement (ITA) in late 1996 and entered in to force in April 1997. Indonesia was one of 29 original signatories as well as the only lower-middle-income country which agreed to this agreement and gradually reduced tariff import on IT ever since. The ITA membership expanded and in the 2015 Nairobi Ministerial Declaration of Trade in IT Products, there were 54 countries (EU counted as one country) agreed for ITA expansion with an additional of 201 IT products that should have zero custom duties [48].

The problem in applying IT, namely the occurrence of productivity paradoxes, occurs because IT investment still has not succeeded in providing the benefits expected by organizations [49–51]. So, the productivity paradox can arise when a company or organization has issued a large budget or investment for IT implementation but it is not followed by the increasing level of productivity. A similar way of thinking can be applied at the country level. If the government failed to balance the IT investment spending with productivity, then it may create a deficit. Today, Indonesia is perhaps one of the biggest net exporters of IT products. The import value of telecommunication equipment in 2017 was 7.426 billion USD increased twofold from the previous year [52]. China is the biggest exporter of telecommunication products in Indonesia lately. In 2017, the MCI issued at least 4053 certificates (out of 7308 certificates) imported telecommunication tools and devices from China [53].

The international agreements, particularly ITA and WTO Ministerial meetings, potentially hinder Indonesian local ICT industry [46]. The domestic IT production shows a deficit of 4.85 billion USD within 1996–2011 [54]. The local ICT producers could not be able to produce IT products competitively since ITA accommodates ICT products but not electronic components needed to produce ICT Products such as Passive and Active components such as Semiconductor, Printed Circuit Board, and many others. PT. INTI, for example, a state owned enterprise that used to produce telecommunication switching, telephone as well as other Telecommunication products, has changed its business core to the system-based solution which includes network management system and subscriber line maintenance system [55].

Nevertheless, these kinds of business models may slowly reduce the role of intermediary businesses such as physical shops and related logistics. The product exchange can be done directly from the seller (producers) to the consumers. Unfortunately, this potentially creates tax losses. The government's supervision over individual e-commerce business and transactions seems to be lag behind. The growing personal shopper of entrusted goods service exists, particularly on social media. This business is quite lucrative for frequent travelers, but the tax losses due to this emerging trend would damage the state's tax income. In 2019, for example, there were at least 422 cases that violate the free custom duty of 500 USD goods brought from abroad up until 25 September 2019 [56].

## **6. Conclusions**

To sum up, Indonesia has great potential for its digital economy in the future. E-commerce sector is the main star of its digital economy which is also lucrative for the global market. This sector is an open market and may still be dominated by



global players. There are two paradoxes of the Indonesian digital economy development: the paradox of regulation and the paradox of productivity.

The main contribution of these paradoxes is law enforcement by the Indonesian regulators that may create an uneven level of playing field between the local and global platform providers. The permanent establishment status of global digital platform service providers, as mandated by Indonesia's ITE law, remains the issue. As a result, this global digital platform service provider should not pay value-added tax. In contrary, it is different from local digital service providers where they should face highly tight regulation just to enter the market.

Also, the government's involvement in global governance particularly relating to e-commerce should be reviewed. The international agreements, particularly ITA and WTO, may cause more import ICT products both hardware and software. As a result, this may weaken the local ICT market and productivity. The government should initiate some programs that may increase local IT-driven productivity so that they can compete with import products. There should be a future study regarding this. These agreements may give an impact on the Indonesian e-commerce industry too.

Also, Indonesia is an active member of global institutions namely WSIS and UPU. As a member of WSIS, Indonesia should develop ICT infrastructure to delineate the digital gap within the regions. As a member of UPU, Indonesia should develop logistic infrastructure. These two infrastructure developments are e-commerce activities enablers. Instead of increasing the local product to be traded on e-commerce platform, the level of import goods is extremely higher which reaching 90%. To address this, the Indonesian government should take fierce action by forcing global e-commerce platform providers should obey the Indonesian regulations and have Permanent Establishment status in accordance with the Law. Another way is by regulating cyber shops or e-commerce platforms to have an obligation to sell local products by 80% as similar to physical shops.

## **Acknowledgements**

This work was supported by the Center of Research and Development on Informatics Application, Information and Public Communication, Research and Human Resources Development Agency of the Ministry of Communication and Information Technology of the Republic of Indonesia and also by the Research Center on ICT Business and Public Policy of Telkom University.

The views that expressed on this paper are those of the authors and do not reflect an official position of the Ministry of Communication and Informatics of the Republic of Indonesia.

In writing this paper, Vience M. Rumata is the lead author who conducting literature reviews as well as writing the article. Meanwhile, Ashwin S. Sastrosubroto contributes in terms of supervision and providing ideas. Both authors contributed to the final stage of the review.

## **Conflict of interest**

The authors whose names are listed certify that they have no financial interest with the subject matter or material discussed in this book chapter.

IntechOpen

## **Author details**

Vience Mutiara Rumata<sup>1\*</sup> and Ashwin Sasongko Sastrosubroto<sup>2</sup>

1 The Ministry of Communication and Informatics of the Republic of Indonesia,  
Indonesia

2 Telkom University, Indonesia

\*Address all correspondence to: vien001@kominfo.go.id

## **IntechOpen**

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Google and Temasek/Bain. e-Conomy SEA 2019: Swipe Up and to the Right: Southeast Asia's \$100 Billion Internet Economy [Internet]. 2019. Available from: [https://www.blog.google/documents/47/SEA\\_Internet\\_Economy\\_Report\\_2019.pdf](https://www.blog.google/documents/47/SEA_Internet_Economy_Report_2019.pdf) [Accessed: 28 November 2019]
- [2] BAKTI. Laporan Tahunan 2018: Indonesia Merdeka Sinyal 2020. Badan Aksesibilitas Telekomunikasi Indonesia, Kementerian Komunikasi dan Informatika: Jakarta (Indonesia); 2019. p. 335
- [3] Kemaritiman dan Investasi. Presiden Resmikan Proyek Palapa Ring, Misi Pemerintah Satukan Indonesia Lewat Internet Tercapai [Internet]. 2019. Available from: <https://maritim.go.id/presiden-resmikan-proyek-palapa-ring-misi-pemerintah-satukan/> [Accessed: 28 November 2019]
- [4] The Ministry of Communication and Informatics. Laporan Tahunan Direktorat Jenderal Aplikasi Informatika Kementerian Komunikasi dan Informatika. Jakarta: Direktorat Jenderal Aplikasi Informatika; 2019. p. 75
- [5] Das K, Michael G, Sudhir P, Tan KT. Unlocking Indonesia's digital opportunity. Jakarta: McKinsey Indonesia Office; 2016. p. 25
- [6] Isna T. E-commerce Pengaruhi Naiknya Impor Barang Konsumsi, Pemerintah Disarankan Lakukan Hal Ini [Internet]. 2019. Available from: <https://republika.co.id/berita/pvrzdv17000/e-commerce-pengaruhinaiknya-impor-barang-konsumsi-pemerintah-disarankan-lakukan-hal-ini> [Accessed: 28 November 2019]
- [7] IMD. IMD World Digital Competitiveness Ranking 2019 [Internet]. 2019. Available from: <https://www.imd.org/wcr/world-digital-competitiveness-rankings-2019/> [Accessed: 02 December 2019]
- [8] Collins T. Health policy analysis: A simple tool for policy makers. *Public Health: Journal of the Royal Institute of Public Health*. 2004;**119**:192-196
- [9] Browne J, Coffey B, Cook K, Meiklejohn S, Palermo C. A guide to policy analysis as a research method. *Health Promotion International*. 2018;(5):1032
- [10] Rein M, Schön D. Reframing policy discourse. In: Fischer F, Forester J, editors. *The Argumentative Turn in Policy Analysis and Planning*. London: UCL Press Limited; 1993. p. 323
- [11] Fischer F. Policy discourse and the politics of Washington think tanks. In: Fischer F, Forester J, editors. *The Argumentative Turn in Policy Analysis and Planning*. London: UCL Press Limited; 1993. p. 323
- [12] Yrjänä L, Rashidfarokhi A, Toivonen S, Viitanen K. Looking at retail planning policy through a sustainability lens: Evidence from policy discourse in Finland. *Land Use Policy*. 2018;**79**:190-198. DOI: 10.1016/j.landusepol.2018.08.013
- [13] Hindman M. *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. New Jersey: Princeton University Press; 2018. p. 240
- [14] Tapscott D. *The Digital Economy: Rethinking Promise and Peril in the Age of Networked Intelligence*. Anniversary Edition—After 20 Years, a New Look Forward. New York: McGraw-Hill Education; 2015. p. 413
- [15] Frost & Sullivan. Digital Market Overview: Indonesia [Internet]. 2018.

Available from: [https://ww2.frost.com/files/3115/2878/4354/Digital\\_Market\\_Overview\\_FCO\\_Indonesia\\_25May18.pdf](https://ww2.frost.com/files/3115/2878/4354/Digital_Market_Overview_FCO_Indonesia_25May18.pdf) [Accessed: 27 February 2019]

[16] MTH. Pergeseran Komoditas ke Layanan Dorong Pertumbuhan Ekonomi Digital [Internet]. 2017. Available from: [https://kominfo.go.id/content/detail/10498/pergeseran-komoditas-ke-layanan-dorong-pertumbuhan-ekonomi-digital/0/berita\\_satker](https://kominfo.go.id/content/detail/10498/pergeseran-komoditas-ke-layanan-dorong-pertumbuhan-ekonomi-digital/0/berita_satker) [Accessed: 27 November 2019]

[17] Kompas.com. Wow... KoprolDibeli Yahoo [Internet]. 2010. Available from: <https://tekno.kompas.com/read/2010/05/25/14182317/Wow....Koprol.Dibeli.Yahoo> [Accessed: 17 December 2019]

[18] CBInsights. The Global Unicorn Club: Current Private Companies Valued At \$1B+ [Internet]. 2019. Available from: <https://www.cbinsights.com/research-unicorn-companies> [Accessed: 28 November 2019]

[19] LPEM FEB UI. Dampak Tokopedia terhadap Perekonomian Indonesia. Infografis. Jakarta (Indonesia): Universitas Indonesia, Lembaga Penyelidikan Ekonomi & Masyarakat; 2019. p. 10

[20] LD FEB UI. Hasil Riset LD FEB UI Tahun 2018: GOJEK Sumbang Rp 44,2 Triliun ke Perekonomian Indonesia. Jakarta (Indonesia): Lembaga Demografi, Universitas Indonesia. Report; 2019. Available from: <http://ldfebui.org/wp-content/uploads/2019/03/Berita-Pers-Lembar-Fakta-LD-UI-Dampak-GOJEK.pdf> [Accessed: 28 November 2019]

[21] Bisara D. Indonesia's E-commerce Market Larger than Estimated; Consumer Habits Changing: Study [Internet]. 2019. Available from: <https://jakartaglobe.id/context/indonesias-e-commerce-market-larger-than-estimated-consumer-habits-changing-study> [Accessed: 02 December 2019]

[22] APJII. Laporan Survei: Penetrasi & Profil Perilaku Pengguna Internet Indonesia. Jakarta (Indonesia): Asosiasi Penyelenggara Jasa Internet Indonesia. Report; 2018

[23] We Are Social. Digital 2019: Indonesia [Internet]. 2019. Available from: <https://datareportal.com/reports/digital-2019-indonesia> [Accessed: 27 November 2019]

[24] BPS. Ekonomi Indonesia 2018 Tumbuh 5,17 Persen [Internet]. 2019. Available from: <https://www.bps.go.id/pressrelease/2019/02/06/1619/ekonomi-indonesia-2018-tumbuh-5-17-persen.html> [Accessed: 17 December 2019]

[25] The Ministry of Communication and Informatics. Perubahan Model Bisnis di Era Ekonomi Digital. Jakarta: Pusat Penelitian dan Pengembangan Aplikasi Informatika dan Informasi dan Komunikasi Publik; 2016. p. 63

[26] MTH. Tuntaskan Peta Jalan E-Commerce, Pemerintah Kini Siapkan Strategi Nasional Ekonomi Digital [Internet]. 2019. Available from: <https://kominfo.go.id/content/detail/20820/tuntaskan-peta-jalan-e-commerce-pemerintah-kini-siapkan-strategi-nasional-ekonomi-digital/0/berita> [Accessed: 29 November 2019]

[27] The Roadmap of E-Commerce 2017-2019 the Presidential Decree Number 74 Year 2017 (Indonesia)s 2(3)

[28] The Implementation of Electronic Systems and Transactions the Government Regulation Number 71 Year 2019 (Indonesia)s 99(3)

[29] The Electronic Based Commerce the Government Regulation Number 80 Year 2019 (Indonesia)s 21(1)

[30] The Ministerial of Communication and Informatics Regulation Number 11 Year 2018 (Indonesia)s 5



- [31] The Consumer Protection Law Number 8 Year 1999 (Indonesia)s 2
- [32] The Ministry of Communication and Informatics. Model Pengaturan Transportasi Online di Indonesia. Jakarta: Pusat Penelitian dan Pengembangan Aplikasi Informatika dan Informasi dan Komunikasi Publik; 2018. p. 52
- [33] Alaydrus H. Pendataan Barang E-Commerce, BI dan Kemenkeu Berambisi Kembangkan SiMoDis [Internet]. 2019. Available from: <https://ekonomi.bisnis.com/read/20190107/9/876229/pendataan-barang-e-commerce-bi-dan-kemenkeu-berambisi-kembangkan-simodis> [Accessed: 17 December 2019]
- [34] The Ministry of Communication and Informatics. Regulasi Tepat Untuk Tingkatkan Ekonomi Digital [Internet]. Jakarta: Kementerian Komunikasi dan Informatika. Press Release Number 272/HM/KOMINFO/10/2018; 2018. Available from: [https://kominfo.go.id/content/detail/14999/siaran-pers-no-272hmkominfo102018-tentang-regulasi-tepat-untuk-tingkatkan-ekonomi-digital/0/siaran\\_pers](https://kominfo.go.id/content/detail/14999/siaran-pers-no-272hmkominfo102018-tentang-regulasi-tepat-untuk-tingkatkan-ekonomi-digital/0/siaran_pers) [cited: 17 December 2019]
- [35] Haines F. The Paradox of Regulation: What Regulation Can Achieve and What It Cannot. Cheltenham: Edward Elgar Publishing Limited; 2011
- [36] Rumata VM, Sastrosubroto AS. The Indonesian E-commerce governance challenges in addressing the penetration of global user generated commerce platforms. International Conference on Computer, Control, Informatics and Its Applications: IC3INA 2017: Emerging Trends in Computational Science and Engineering. Indonesia: IEEE; 23-26 October 2017. p. 7-11. DOI: 978-1-5386-3978-8/17
- [37] Finck M. Digital co-regulation: Designing a supranational legal framework for the platform economy. *European Law Review*. 2018;43:47-68
- [38] Putri VM. Tahun Ini Bukalapak Punya Banyak Kerja Sama dengan Pemerintah [Internet]. 2019. Available from: <https://inet.detik.com/business/d-4379481/tahun-ini-bukalapak-punya-banyak-kerja-sama-dengan-pemerintah> [Accessed: 06 December 2019]
- [39] Tempo.co. Kemenperin: 90 Persen Produk E-commerce Indonesia Barang Impor [Internet]. 2018. Available from: <https://bisnis.tempo.co/read/1123536/kemenperin-90-persen-produk-ecommerce-indonesia-barang-impor/full&view=ok> [Accessed: 17 December 2019]
- [40] The Ministry of Trade. Laporan Akhir: Analisis Produk dalam Negeri di Toko Modern [Internet]. Jakarta: Pusat Kebijakan Perdagangan Dalam Negeri, Badan Pengkajian dan Pengembangan Kebijakan Perdagangan Kementerian Perdagangan; 2014; 47 p. Available from: [http://bppp.kemendag.go.id/media\\_content/2017/08/laporan\\_akhir\\_kajian\\_produk\\_dalam\\_negeri\\_di\\_toko\\_modern.pdf](http://bppp.kemendag.go.id/media_content/2017/08/laporan_akhir_kajian_produk_dalam_negeri_di_toko_modern.pdf) [cited: 17 December 2017]
- [41] Haryanto AT. Status Perusahaan Facebook di Indonesia Terungkap! [Internet]. 2018. Available from: <https://inet.detik.com/law-and-policy/d-4009700/status-perusahaan-facebook-di-indonesia-terungkap> [Accessed: 05 December 2019]
- [42] Statista. Indonesia: Number of Facebook Users 2017-2023 [Internet]. 2019. Available from: <https://www.statista.com/statistics/304829/number-of-facebook-users-in-indonesia/> [Accessed: 05 December 2019]
- [43] Hanafizades P, Khosravi B, Badie K. Global discourse on ICT and the shaping of ICT policy in developing countries. *Telecommunications Policy*. 2019;43:324-338. DOI: 10.1016/j.telpol.2018.09.004
- [44] Shwab J. E-Commerce at the WTO: The Need for a New Agreement



- [Internet]. 2019. Available from: <http://www.mjilonline.org/e-commerce-at-the-wto-the-need-for-a-new-agreement/> [Accessed: 09 December 2019]
- [45] European Commission. 76 Partners Launch WTO Talks on E-Commerce [Internet]. 2019. Available from: [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_19\\_684](https://ec.europa.eu/commission/presscorner/detail/en/IP_19_684) [Accessed: 09 December 2019]
- [46] Rumata VM, Sastrosubroto AS. The impact of ICT international agreements on local Indonesian ICT industry: A policy research. *Asia Pacific Journal of Contemporary Education and Communication Technology*. 2019;5(2):65-73
- [47] The Import Shipment Goods Provision the Ministry of Finance Regulation Number 112/PMK.04/2018 (Indonesia)s 13
- [48] European Commission. The Expansion of the Information Technology Agreement: An Economic Assessment [Internet]. Luxembourg: Publications Office of the European Union; 2016. 26 p. DOI: 10.2781/11765. Available from: [https://trade.ec.europa.eu/doclib/docs/2016/april/tradoc\\_154430.pdf](https://trade.ec.europa.eu/doclib/docs/2016/april/tradoc_154430.pdf) [cited: 09 December 2019]
- [49] Roach S. America's Technology Dilemma: A Profile of the Information Economy. New York: Morgan Stanley. Economics Newsletter Series; 1987
- [50] Brynjolfsson E, Yan S. Information technology and productivity: A review of the literature. *Advances in Computers Academic Press*. 1996;43:179-214
- [51] Ward J, Peppard J. Strategic Planning for Information System. 3rd ed. England: John Wiley & Sons; 2002. p. 624
- [52] BPS. Nilai Impor Perlengkapan Telekomunikasi Menurut Negara Asal Utama (Nilai CIF: juta US\$), 2000-2017 [Internet]. Jakarta: Badan Pusat Statistik; 2019. Available from: <https://www.bps.go.id/statictable/2014/09/08/1049/nilai-impor-perlengkapan-telekomunikasi-menurut-negara-asal-utama-nilai-cif-juta-us-2000-2017.html> [cited: 09 December 2019]
- [53] Ditjen SDPPI. Laporan Tahun 2017 SDPPI: Optimalisasi Spektrum Frekuensi Radio untuk Negeri [Internet]. Jakarta: Direktorat Jenderal Sumber Daya dan Perangkat Pos dan Informatika; 2017. 144 p. Available from: [https://www.postel.go.id/downloads/45/20180716115840-Final\\_Laptah\\_2017\\_v12.pdf](https://www.postel.go.id/downloads/45/20180716115840-Final_Laptah_2017_v12.pdf) [cited: 09 December 2019]
- [54] Ningsih R. Implikasi Perjanjian Teknologi Informasi terhadap Kinerja Perdagangan Produk Teknologi Informasi Indonesia. *Buletin Ilmiah Litbang Perdagangan*. 2013:19-36
- [55] Detik Finance. PT INTI, Penyokong Industri Telekomunikasi dari Bandung [Internet]. 2014. Available from: <https://finance.detik.com/industri/d-2588422/pt-inti-penyokong-industri-telekomunikasi-dari-bandung> [Accessed: 08 March 2019]
- [56] Kusuma H. Begini Modus Jastip yang Sengaja Hindari Pajak [Internet]. 2019. Available from: <https://finance.detik.com/berita-ekonomi-bisnis/d-4724933/begini-modus-jastip-yang-sengaja-hindari-pajak> [Accessed: 17 December 2019]

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Natural Language Processing Applications in Business

*Mohammed Bahja*

## Abstract

Increasing dependency of humans on computer-assisted systems has led to researchers focusing on more effective communication technologies that can mimic human interactions as well as understand natural languages and human emotions. The problem of information overload in every sector, including business, health-care, education etc., has led to an increase in unstructured data, which is considered not to be useful. Natural language processing (NLP) in this context is one of the effective technologies that can be integrated with advanced technologies, such as machine learning, artificial intelligence, and deep learning, to improve the process of understanding and processing the natural language. This can enable human-computer interaction in a more effective way as well as allow for the analysis and formatting of large volumes of unusable and unstructured data/text in various industries. This will deliver meaningful outcomes that can enhance decision-making and thus improve operational efficiency. Focusing on this aspect, this chapter explains the concept of NLP, its history and development, while also reviewing its application in various industrial sectors.

**Keywords:** NLP, ML, AI, deep learning, chatbots

## 1. Introduction

The aim of this chapter is to review the applications of NLP in business. Advancement in the application of technologies in daily lives has revolutionised communication and interaction between humans and computers. It has also had the same effect on humans throughout the world, for they can now use communicative applications (enabling translation and transliteration). Interaction/communication is an important factor in daily lives, as it is the basis for humans to become involved in activities such as business, education, commerce, healthcare management, politics and socialising. The way in which humans communicate can be referred to as natural language, which includes speech, text and emotions [1]. Humans are more involved in speaking as a means of communication than in writing (text). Text, rather than speech, is important for developing any AI application that facilitates the communication process, as machines, as yet, are not as effective in learning languages as humans, as a result of which, they have to depend on the text/data [2]. A child spends years learning any language, which involves emotions, fluency, grammar and speaking norms, among other things. Hence, it is not an easy task for a computer scientist to build applications that deal with natural language input/output.

The study of language, with regard to its grammar, rules, semantics and phonetics, is referred to as linguistics. Devising rules of language, methods for syntax and semantics are included as part of linguistics study. The methods and rules proposed by theoretical linguists can be processed by computer systems, which generate natural language which cope with such matters as grammar, semantic norms etc. This approach has been referred to as computational linguistics [3], for which statistical approaches are utilised in the process of analysing text/data. However, the approaches for processing natural language may not be limited to statistics, for they could also involve the application of advanced inference methods, like machine learning (ML) and deep learning, which are popular artificial intelligence (AI) techniques. Application of advanced techniques could address various challenges associated with the processing of natural language, such as breaking sentences, tagging the parts of speech (POS), generating dependency graphs, building an appropriate vocabulary, linking different components of vocabulary, setting the context, extracting semantic meanings or transforming unstructured data into a structured format [4–6].

With the development of technologies and applications of various methods for processing natural language, the understanding and definition of NLP has changed over time. Statistical NLP uses the data generated according to some probability distribution and then makes some inferences about this [7]. In a wider sense, NLP should be able to cover any kind of computer manipulation of natural language: it could be as simple as counting word frequencies to compare different writing styles; or as complex as understanding complete human utterances for delivering responses to an input [8]. Given the application of effective technologies, such as ML, in recent years, for NLP, the most appropriate definition is proposed as:

‘NLP is a collective term referring to automatic computational processing of human languages. This includes both algorithms that take human-produced text as input and algorithms that produce natural looking text as outputs’ [9].

Focusing on the brief background discussed, it can be identified that there is a growing demand and scope for larger application of NLP in various sectors. Therefore, reviewing the applications of NLP in different sectors contributes valuable knowledge to academicians, researchers and industry practitioners working in the areas of NLP applications. Accordingly, the purpose of this chapter is to review the applications of NLP in various business service sectors and provide future directions of NLP applications.

## **2. Methods**

Review chapters could be an important sources of information for academicians and practitioners in guiding their decision-making and work practices [10]. High-quality reviews are cited more frequently [11]; and are found to be downloaded more often than any published article, as they offer a high-quality information from various articles in an effective way [12]. In addition, reviews enable one to have an overview, if not a detailed knowledge of the area in question, as well as references to the most useful primary sources [13].

Literature review, in this aspect, can be very useful in collecting information from various sources such as systematic reviews, theoretical reviews etc. Literature review may be the best methodological tool for gathering evidence in a certain area [14]. As this study focuses on gathering information related to application of NLP in business, it can be considered as explorative study. The approach in literature reviews can be narrow, such as investigating the effect of relationship, or it can be broader, such as exploring collective evidence in relation to a specific area or



concept. Literature reviews are identified to be very useful, when the aim of the study is to provide an overview about certain issue. In addition, it can be useful for creating research agendas, identifying gaps in research or simply discussing a particular matter [15]. It can be very beneficial in mapping the development of a particular research field in different contexts over time [16]. Considering the nature of this study, which focuses on investigating the development of NLP and its applications in business, literature review methodology is considered as it is identified as an effective methodological tool. Therefore, this chapter adopted a literature review methodology by collecting the relevant information from various sources of information including academic journals, online articles, conferences, white papers and industry reports. A brief literature review about the development of NLP over time is presented in Section 3. In addition, the information collected regarding NLP applications is categorised into major areas/industries including commerce, E-Governance, healthcare, and education, and other relevant areas as explained in Section 4.

### **3. Brief history of NLP**

The roots of NLP development can be traced back to 1950s, when the idea ‘can machines think’ was put forward and examined by Alan Turing, which is popularly known as the Turing Test for the creation of intelligence [17]. This study paved the way for research related to the understanding of natural languages and machine translation. However, there was not much progress, for after many experiments, as identified in a 1966 ALPAC report, no one had succeeded in meeting his specified criteria [18, 19]. Some promising NLP systems were developed in the 1960s, which have been updated since through various techniques and new conceptualisations. Some of the major applications in the 1960s included ‘SHRDLU’, a natural language understanding program that allowed user interaction in English terms [20]; and ‘ELIZA’, a NLP program that stimulated human-machine conversation using a pattern matching and substitution methodology [21].

The research during the 1970s was focused on formatting and structuring real-world data into computer understandable data, which was based on conceptual ontologies. Some of the major developed applications included MARGIE, which could read an English sentence, using Riesbeck’s expectation-based parser as well as building a conceptual dependency form, which represented the meaning of the sentence [22]; and QUALM, which used semantic modelling for framing questions and associating them with different strategies to find answers [23]. In the 1980s, most of the applications relied on complex sets of hand-written rules for NLP. By the late 1980s and 1990s, the research was focused on the application of statistical models for making soft and probabilistic decisions based on the input data along with ML algorithms, such as decision-trees [18, 24]. Accordingly, various techniques of ML became popular approaches to NLP, as they could achieve effective results for various NLP tasks, such as modelling and parsing [25–28]. Other popular techniques utilised in recent years are text embedding to capture the semantic properties of words, deep learning, machine translation and neural machine translation [29] (**Figure 1**).

Deep learning has gained momentum in NLP over ML in recent years for various reasons, including the ability to handle larger amounts of training data, the availability of faster machines and multicore CPU/GPUs. Moreover, new models and algorithms with advanced capabilities that have improved performance have stimulated deep learning research [29]. The manually designed features in ML are often over-specified, incomplete and take a long time to design and validate,



Time Period	Development
1950s	Turing test for creation of intelligence
1960s	SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies ELIZA, simulated conversation by using a "pattern matching" and substitution methodology
1970s	Used "conceptual ontologies", for structuring real-world information into computer-understandable data. Examples: MARGIE (1975), SAM (1978), PAM (1978), QUALM (1977)
1980s-1990s	Used Machine Learning algorithms such as decision trees Used statistical models for making soft and probabilistic decisions Development of Multilingual textual Corpora
2000s	Use of supervised and unsupervised learning algorithms
2010s -	Use of representation learning, deep learning, neural network style learning Understanding semantic properties of words

**Figure 1.**  
*Brief history of NLP evolution.*

whereas deep learning’s underpinning parameters are quick to adapt and learn fast. Machine translation on various platforms is one of the classic representations of NLP implementation, the use of which is increasing in various industries and the market is projected to be \$56 billion by 2021 [30]. Applications, such as ‘Google Translate’, one of the widely used applications on online platforms; Facebook, using machine translation for translating texts in posts and comments; and eBay, similarly using machine translation for enabling cross-border trade by connecting the users across the world. Neural Machine Translation, which manages the entire process of NLP through an artificial neural network known as a Recurrent Neural Network, has been found to be very effective [31]. Another important breakthrough for NLP application is the delivery of dialogue and conversations through chatbots or virtual assistants, such as Apple Siri, Amazon Alexa and Microsoft Cortana. These are increasingly being used, being continuously updated through ML. In sum, the growth of NLP can be observed with the changing needs and increasing adoption of technologies, being transformed through various techniques, such as rules-based algorithms, statistical models, ML, deep learning and RRN.

## 4. Applications of NLP in business

Communication is an important factor that builds relationship among individuals, and also between organisations and individuals. Interactions among the people from different countries have been facilitated using translators. However, with the development of NLP, communications between such entities have become easier, with language no longer being a barrier to human and business interactions. Businesses can internationalise their operations, and trade relations can be improved, thereby enhancing global commerce. Application of NLP can be found in a range of business contexts, some of which are described in what follows.

### 4.1 NLP in commerce

Sentiment analysis is one NLP technique that is widely used in the finance market/trading. It is the process of understanding an opinion about a given subject

through written or spoken language and accurate predictions are considered to be a game changer in achieving success on the stock market. Financial analysts, business analysts and trade analysts are employed in the organisations for this purpose, to monitor and analyse the impact of various happenings to stock prices. Their work can be simplified by using NLP with ML and AI techniques, are effective in analysing the data from internet, news, blogs and social networking sites (analysis of huge chunks of data across various channels). They can predict trade fluctuations, which enable investors to take the right decisions at the right time. For instance, when using sentiment analysis, words such as 'good', 'profit', 'benefit', 'positive' and 'growth' can be tagged with positive value, while those, such as 'risk', 'fall', 'bankruptcy', and 'loss' are tagged with negative values. Making meaning from the text identified using NLP and ML allows for accurate predictions to be made [32].

Similarly, NLP's application in business can be observed in other forms, such as: filtering emails and identifying spam (by analysing text); enhancing security through voice recognition; extracting information from large datasets; using online assistants/chatbots (questioning-answering) for providing customer service; promoting business intelligence; placing relevant ads online by using keyword matching and sense disambiguation; analysing competitors and the market using event extraction techniques, this list being inconclusive [33–37]. NLP can be applied in E-Commerce/S-Commerce applications for analysing the consumer behaviour based on the feedback, posts, reviews, ratings etc. (text analysis). It can also provide effective customer support through virtual assistants that can respond to customer queries in a range of formats, including: text, speech, audio or video [38].

#### **4.2 NLP in “E-Governance”**

NLP can enhance E-Governance, which is completely reliant on an information and communication technologies infrastructure [39]. It can facilitate interaction between the citizens and government using an E-Governance framework. For example, citizens who are illiterate can share their opinions and interact with government through audio/video conversation, which can be translated into text for documentation. Similarly, voice-enabled mobile applications can read the text and convey messages to the people. In addition, government can use NLP for monitoring various channels through which people interact to identify their concerns. Moreover, in this context, it can be used for enhanced security, by preventing any breaches of this.

Interactions between citizens and government involve enormous volumes of information exchange. Filtering and formatting such large amounts of data is a complex task, which can be effectively managed by the use of NLP with AI and ML techniques [40]. For instance, sentiment analysis can be used for mining opinions from huge datasets, including feedback, complaints and reviews about a particular policy, thereby ascertaining what the general consensus on it is [41]. That is, it can be used to gauge opinions from feedback, complaints and suggestions made by the residents of a city in relation to various urban issues, such as carbon emissions, lack of a solid waste management system and lack of sanitation [42]. This would help the relevant government agencies in selecting appropriate strategies to address the issues identified. In sum, as a result of adopting these approaches, the communication between government and citizens could be markedly improved, in particular, the former could use them to identify the concerns of the latter and thus, address these in a timely manner. In addition, it encourages E-Participation from both sides, thereby greatly contributing to the effective implementation of government policies.

### **4.3 NLP in healthcare**

The healthcare industry is one of the fastest growing. It has been integrating technology applications with healthcare practices, such as diagnosis and treatment for delivering effective and efficient services, which can be accessed by the majority of people, thereby overcoming any regional disadvantages (access in rural areas through E-Health services). However, the amount of data being collected through this process, in the form of such as EHRs (Electronic Health Records), sensor information, diagnoses, monitoring data, healthcare operations and management data is vast. It has been ascertained that about 80% of the healthcare data is unstructured, of poor quality, and considering the current format, it is effectively unusable [40–68]. NLP, in this context, can be considered as one of the effective approaches for addressing these problems by using it to parse information and extract critical strings of data, thereby offering an opportunity to leverage unstructured data. It has been already in use in various healthcare operations across the world. The market for NLP in healthcare and life sciences is expected to reach \$2650.2 million by 2021 [44]. The major drivers of NLP applications in healthcare [40–69] include the following:

- the rising need for handling the surge in clinical data;
- support for value-based care and population health management;
- improving clinical documentation and making computer-assisted coding more efficient;
- improving patient and healthcare provider interactions;
- empowering patients with health literacy;
- increasing need for higher quality healthcare.

So far, the applications of NLP in healthcare have achieved remarkable results, as a result of which its deployment in healthcare technologies has been increasing in recent years. For instance, NLP techniques were used for reviewing more than 2 billion EHR documents for indications of PTSD, depression and potential self-harm in veteran patients, in 2013, by the Department of Veterans Affairs (USA) [45]. Similarly, in another study [46], an analytics algorithm that leverages NLP was able to predict the onset of psychosis in high-risk youths with 100% accuracy. Similarly, a study conducted by researchers from the University of Alabama found that by applying NLP techniques the reportable cancer cases identified was 22.6% more accurate than a manual review of medical records [47]. Further, various other studies have identified the potential benefits, including effective decision-making by physicians; reducing physician burnout (disillusionment among doctors tired of repetitive data entry tasks and administrative duties as well as excessive time spent combing through patient data); addressing accelerating demand for healthcare services; and dealing with increasing numbers of healthcare claims [48–51], through the application of NLP techniques in the healthcare context.

### **4.4 NLP in education**

NLP can enhance education technology in various ways, in particular, being effective in analysing data and text, thereby delivering meaningful outcomes. For

learning of language-related subjects, it can enhance teaching by focusing on reading, writing and speaking. In general, it can support the needs of students, teachers and researchers [40–42, 52–71]. Language assessment (formative and summative) is one of the upcoming major areas of NLP application, being used for assessing student's proficiency in reading, writing and speaking (e.g., automated scoring system for assessing/grading essays using NLP techniques). Syntactics analysis, for instance, can be used for correcting writing errors, which can provide instant feedback for correction and this process can be very useful for students, especially with the disabled, such as the deaf/ blind [41]. Other areas of application include proofreading, detecting errors of machine translation, assessing the meaning and content of student works, such as essays, short/long answers, training non-native speakers, and assessing reading and speaking [42, 52–55].

Moreover, NLP can be used in the teaching the use of language, that is, any subject using any language as a medium. One of its current uses pertains to the chatbot/virtual assistant applications that are used for various purposes, including teaching, training, tutoring and assessment in the education sector. Chatbots can also be integrated with gaming technologies, which promotes an active learner environment in the E-Learning context, with virtual characters or avatars [56–71]. The chatbot application, similar to E-Learning resources, also fosters students' engagement in learning activities by motivating them and gaining their attention by using innovative media features [57]. They can also be used in self-guided learning through learning modes, such as quizzes, which is one of the important principles of E-Learning [58]. However, many extant E-Learning approaches (online educational videos, online classes, portals) have been focused on one-way communication using rich and real-time media, thus failing to address pedagogical and social aspects. With the introduction of discussion forums and messaging applications, this issue was addressed to an extent, but increased the complexity and the amount of information to be processed, which can be a burden to the learner [59]. Chatbots in such instances can be very effective in providing the accurate information in real time, where the learners need not depend on different entities to seek answers. The personalisation and additivity features can be integrated into the chatbot applications, which can increase the accuracy of the learner model obtained by the student. He/she can negotiate jointly with the model using a natural conversational flow of interaction [59]. This aspect reflects the difference in designing educational chatbot applications and E-Learning resources. That is, for the former, voice user interface and conversational flow techniques need to be designed closely with the users of the system, this allows for the inclusion of culture and social issues in the stages of design, unlike E-Learning where the developers greatly influence the design process [70].

In addition to assessing students' linguistic inputs, through such matters as feedback, assessments/ coursework, essays, short/long answers and serving as a medium of instruction (tutoring, teaching, training), NLP can also be used in processing text and speech in other ways that can support teachers, students, system developers and researchers. As the amount of electronically available text in education is increasing rapidly, NLP can be effective in organising the relevant text for teaching. For, it can be a very onerous task for teachers to identify appropriate materials for effective input during lectures. That is, teaching with the most up to date course materials is beneficial for both students in learning and teachers in teaching the subjects more effectively and efficiently [41]. In addition, NLP can be effective in research, especially in formulating meaningful extractions from bodies of literature (systematic reviews, scoping reviews, meta-analyses etc.). In addition, NLP can be effective in processing qualitative data, such as those collected from interviews, in various formats including audio, video and text. Moreover, it can



be used for converting audio translations from one language to other or into text, thereby assisting with data analysis [60, 61].

#### **4.5 NLP in other sectors**

Sectors such as hospitality and travel, manufacturing and logistics, media and entertainment etc. can benefit from the application of NLP. For instance, the service engineers who repair electronics and other products at homes and offices need to provide the fault reports to the companies, which will be analysed and the identified issues would be resolved in products manufactured in future. However, compiling and interpreting information provided by the service engineers (age of devices, parts replaced, fault description etc.) could be a complex task for quality engineers as the text can be full of abbreviations, spelling mistakes and can range from verbose to abrupt in content length [72]. NLP can effectively refine such data and provide appropriate information for quality engineers. NLP can be implemented in all those areas where AI is applicable either by simplifying communication process or by refining and analysing information. Similarly, it is applied in hospitality and travel industry in providing excellent customer satisfaction through chatbots and virtual assistants in selecting hotels, booking flights, providing information about places to visit and experience etc., which is customised according to the travellers' preferences [73]. Similar applications of NLP can be identified across all the sectors where the services are provided to the customers, such as streamlining the programs according to the customers' viewing history on televisions (Media & entertainment) [74]; or communicating with travel chatbots about the travel routes while driving such as Google Maps(travel). Therefore, NLP can be applied across various sectors which are not only service oriented but also manufacturing sectors.

#### **5. Future directions**

NLP has made a major leap in the past few years, both in theory and practical integration into various industry solutions. Interacting with a virtual chatbot when booking a hotel or flight to extracting insights from call-centre interactions to analyse customer behaviour, such as their attitudes and satisfaction levels, has made NLP's presence felt in almost every industry-based solution. With deep learning being progressively developed, it has finally come into its own with RNN, outperforming the traditional ML methods in a wide range of NLP tasks. It can deliver better performance than humans in complex tasks such as questioning-answering and machine translation [62]. In addition, the recent advances in the technology and practices are promising for improving scalability and robustness. That is, they herald a shift in how business organisations consume computing resources and deploy NLP applications. From the technical perspective, various new approaches have been emerging for such learning (NLP and deep learning using training data), such as the Universal Language Model Fine-tuning (ULMFit); and Bidirectional Encoder Representations from Transformers (BERT), developed by Google for contextual pre-training, which has now advanced to XLNet. This avoids the issues that BERT suffers from by using a technique called 'permutation language modelling' [63]. These methods have improved NLP tasks, outperforming state-of-the-art performance as well as obtaining high accuracy. For example, application of the ULMFit method on text classification tasks has reduced the error by 18–24% for the majority of datasets. In addition, with only 100 labelled examples, it matched the performance of training from scratch on a hundred times more data [62].

The market for NLP in future is considered to be very promising, for various reports [30–64] have projected a rapid increase in its application. The NLP market, including hardware, applications and service is projected to reach \$22.3 billion by 2025, while the market for AI enabled NLP software has been forecast to reach \$5.4 billion by 2025 [64]. With the integration of deep learning techniques and NLP, with platforms like iDS Cloud, businesses in future will reap benefits in the areas of customer services, marketing, business intelligence and operational management. This will reduce the dependency on the data scientists, thereby reducing the costs and improving process efficiency [65]. Humanoids/robotics is one of the most promising advancements in technology, which can be experienced in near future in industrial sectors and also in social lives for managing various tasks. However, enabling humanoids to function like humans in every aspect and assisting them in performing most complex activities would need the ability to comprehend entirely accurately human speech. Given the versatile nature of robots, NLP will become more important than ever, because a single misinterpreted command by one could lead to it performing undesirable actions and may even cause harm to humans. Research on designing socially intelligent robots by using NLP [66, 67] has been increasing in recent years, which could result in the development of fully functional and safe humanoids in the near future.

## **6. Conclusion**

The purpose of this paper is to review the applications of NLP in business. NLP potential lies in effectively learning and understanding the natural language. It can handle the issues associated with translation and transliteration by effectively improving the communication process between humans and machines across various formats. Based on the review, it can be concluded that its presence in various industry-based solutions has been increasing in the recent years. Complex processes in commerce, such as predictions and decision-making in stock-market trading; simplifying customer interactions using chatbots on commerce platforms, thus making the interaction more enjoyable; analysing citizens issues from large volumes of data in E-governance; effectively managing healthcare operations, such as diagnoses, service delivery and records management; and improving the learning and teaching approaches in education sector, are some of the benefits that NLP techniques can engender. In addition, NLP's integration with advanced technologies, such as ML, AI and deep learning, can deliver more accurate outcomes compared to the use of traditional methods. However, its application in the areas of AI, robotics, and other advanced systems is under-researched. In sum, given the efficiency of NLP techniques in improving the accuracy in data analysis and processing of natural language, it has immense scope for use in the areas of robotics and business intelligence in future.

IntechOpen

IntechOpen

### **Author details**

Mohammed Bahja

School of Computer Science, University of Birmingham, Birmingham, UK

\*Address all correspondence to: [m.bahja@cs.bham.ac.uk](mailto:m.bahja@cs.bham.ac.uk)

### **IntechOpen**

---

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Eisenstein J. Introduction to Natural Language Processing. Cambridge, MA: The MIT Press; 2019
- [2] Ghosh S, Gunning D. Natural Language Processing Fundamentals. Birmingham: Packt Publishing Ltd; 2019
- [3] Grishman R. Computational Linguistics. 4th ed. New York: Cambridge University Press; 1999
- [4] Mitkov R. The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press; 2019
- [5] Barnden J. Challenges in natural language processing: The case of metaphor (commentary). International Journal of Speech Technology. 2008;**11**(3-4):121-123. DOI: 10.1007/s10772-009-9047-3
- [6] Weischedel R, Bates M. Challenges in Natural Language Processing. Cambridge: Cambridge University Press; 2006
- [7] Manning C, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge (Massachusetts): MIT Press; 1999
- [8] Bird S, Klein E, Loper E. Natural Language Processing with Python. Beijing: O'Reilly; 2011
- [9] Goldberg Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research. 2016;**57**:345-420
- [10] Paré G, Trudel M-C, Jaana M, Kitsiou S. Synthesizing information systems knowledge: A typology of literature reviews. Information & Management. 2015;**52**(2):183-199
- [11] Rowe F. What literature review is not: Diversity, boundaries and recommendations. European Journal of Information Systems. 2014;**23**(3):241-255
- [12] Montori V, Wilczynski N, Morgan D, Haynes R. Systematic reviews: A cross-sectional study of location and citation counts. BMC Medicine. 2003;**1**(1). DOI: 10.1186/1741-7015-1-2
- [13] Cronin P, Ryan F, Coughlan M. Undertaking a literature review: A step-by-step approach. The British Journal of Nursing. 2008;**17**(1):38-43
- [14] Tranfield D, Denyer D, Smart P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management. 2003;**14**(3):207-222
- [15] Torraco R. Writing integrative literature reviews: Guidelines and examples. Human Resource Development Review. 2005;**4**(3):356-367
- [16] Snyder H. Literature review as a research methodology: An overview and guidelines. Journal of Business Research. 2019;**104**:333-339
- [17] Stevan H. The annotation game: On turing on computing, machinery, and intelligence, in Epstein, Robert. In: Peters G, editor. The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer. New York: Springer; 2008
- [18] Hutchins J. The History of Machine Translation in a Nutshell. 2005. Available from: <http://www.hutchinsweb.me.uk/Nutshell-2005.pdf> [Accessed: 23 January 2020]
- [19] Pierce JR, Carroll JB, Hamp EP, Hays DG, Hockett CF, Oettinger AG, et al. Language and machines — Computers in translation and linguistics.



In: ALPAC Report. Washington, DC: National Academy of Sciences, National Research Council; 1966

[20] Winograd T. SHRDLU. 2020. Available from: <http://hci.stanford.edu/winograd/shrdlu/> [Accessed: 22 January 2020]

[21] Weizenbaum J. Computer Power and Human Reason: From Judgment to Calculation. San Francisco: W. H. Freeman and Company; 1976. ISBN 0-7167-0463-3

[22] Schank R. Conceptual Information Processing. Amsterdam: Elsevier Science; 2015

[23] Lehnert WG. A conceptual theory of question answering. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence, Massachusetts Institute of Technology. Cambridge, Massachusetts, USA; 1977

[24] Barman B. The linguistic philosophy of Noam Chomsky. *Philosophy and Progress*. 2014;**LI-LII**(1-2):103-122. DOI: 10.3329/pp.v51i1-2.17681

[25] Deyringer V, Fraser A, Schmid H, Okita T. Parallelization of neural network training for NLP with Hogwild! *The Prague Bulletin of Mathematical Linguistics*. 2017;**109**(1):29-38. DOI: 10.1515/pralin-2017-0036

[26] Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. Cornell University. 2016. Available from: <https://arxiv.org/abs/1602.02410> [Accessed: 11 February 2020]

[27] Vinyals O, Kaiser L, Koo T, Petrov S, Sutskever I, Hinton G. Grammar as a foreign language. *Conference on Neural Information Processing Systems*. 2015. Available from: <https://papers.nips.cc/paper/5635-grammar-as-a-foreign-language.pdf> [Accessed: 23 January 2020]

[28] Collins M. Three generative, lexicalised models for statistical parsing. In: *Proceedings of the 35th Annual Meeting of the ACL*. USA; 1997

[29] Le J. The 7 NLP Techniques that Will Change how you Communicate in the Future (Part I). 2018. Available from: <https://heartbeat.fritz.ai/the-7-nlp-techniques-that-will-change-how-you-communicate-in-the-future-part-i-f0114b2f0497> [Accessed: 23 January 2020]

[30] GALA. Translation and Localization Industry Facts and Data. Available from: <https://www.gala-global.org/industry/industry-facts-and-data> [Accessed: 22 January 2020]

[31] Zhaopeng T, Zhengdong L, Yang L, Xiaohua L, Hang L. Modeling Coverage for Neural Machine Translation. Cornell University. Aug 2016; arXiv:1601.04811

[32] Lagi M. Natural Language Processing – Business Applications. 2019. Available from: <https://emerj.com/ai-sector-overviews/natural-language-processing-business-applications/> [Accessed: 24 January 2020]

[33] Kass A, Cowell-Shah C. Business event advisor: Mining the net for business insight with semantic models, lightweight NLP, and conceptual inference. *Geography Compass*. 2019;**10**(13):54-61

[34] Kass A, Cowell-Shah C. Using lightweight NLP and semantic modeling to realize the Internet's potential as a corporate radar. *American Association for Artificial Intelligence*. 2006. Available from: <https://www.aaai.org/Papers/Symposia/Fall/2006/FS-06-02/FS06-02-014.pdf> [Accessed: 24 January 2020]

[35] Sintoris K, Vergidis K. Extracting business process models using natural language processing (NLP) techniques. In: *IEEE 19th Conference on Business*

Informatics (CBI). 2017. DOI: 10.1109/cbi.2017.41

[36] Manuel P, Demirel O, Gorener R. Application of eCommerce for SMEs by using NLP principles. In: IEMC 2003 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change. 2003. DOI: 10.1109/iemc.2003.1252317

[37] Davda A, Mittal P. NLP and Sentiment Driven Automated Trading. 2008. Available from: [https://www.seas.upenn.edu/~cse400/CSE400\\_2007\\_2008/DavdaMittal/NLPSentimentTrading.pdf](https://www.seas.upenn.edu/~cse400/CSE400_2007_2008/DavdaMittal/NLPSentimentTrading.pdf) [Accessed: 24 January 2020]

[38] Zhang Z. Weighing stars: Aggregating online product reviews for intelligent E-commerce applications. *IEEE Intelligent Systems*. 2008;23(5): 42-49. DOI: 10.1109/mis.2008.95

[39] Ghosh S. Application of natural language processing (NLP) techniques in E-governance. In: *E-Government Development and Diffusion*. 2009. pp. 122-132. DOI: 10.4018/978-1-60566-713-3.ch008

[40] Litman D. Natural language processing for enhancing teaching and learning. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Arizona, USA: Phoenix; 2016

[41] Tetreault J, Chodorow M. The ups and downs of preposition error detection in ESL writing. In: *Proceedings of COLING*. 2008. pp. 865-872. Available from: <https://www.aclweb.org/anthology/C08-1109>

[42] Ward NG, Escalante R, Bayyari YA, Solorio T. Learning to show you're listening. *Computer Assisted Language Learning*. 2007;20:385-407

[43] Maruti T. 6 Driving Factors behind NLP in Healthcare. 2020. Available

from: <https://marutitech.com/nlp-in-healthcare/> [Accessed: 22 January 2020]

[44] Marketsandmarkets. NLP Healthcare & Life Sciences Market by Technology & Services| MarketsandMarkets. 2020. Available from: <https://www.marketsandmarkets.com/Market-Reports/healthcare-lifesciences-nlp-market-131821021.html> [Accessed: 22 January 2020]

[45] Bresnick J. VA Uses EHRs, Natural Language Processing to Spot Suicide Risks. 2020. Available from: <https://ehrintelligence.com/news/va-uses-ehrs-natural-language-processing-to-spot-suicide-risks> [Accessed: 22 January 2020]

[46] Bresnick J. Predictive Analytics, NLP Flag Psychosis with 100% Accuracy. 2020. Available from: <https://healthitanalytics.com/news/predictive-analytics-nlp-flag-psychosis-with-100-accuracy> [Accessed: 22 January 2020]

[47] Osborne J, Wyatt M, Westfall A, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association*. 2016;23(6):1077-1084. DOI: 10.1093/jamia/ocw006

[48] Popowich F. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*. 2005;7(1):59-66. DOI: 10.1145/1089815.1089824

[49] Wen A, Fu S, Moon S, El-Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digital Medicine*. 2019;2:130. DOI: 10.1038/s41746-019-0208-8

- [50] Aaron M. Using AI and NLP to alleviate physician burnout. Books, Presentations, Posters, Etc. 2019. Available from: [https://digitalcommons.psjhealth.org/other\\_pubs/39](https://digitalcommons.psjhealth.org/other_pubs/39) [Accessed: 22 January 2020]
- [51] Sandeep Kumar E, Satya Jayadev P. Deep learning for clinical decision support systems: A review from the panorama of Smart healthcare. In: *Studies In Big Data*. 2019. pp. 79-99. DOI: 10.1007/978-3-030-33966-1\_5
- [52] Xue H, Hwa R. Syntax-driven machine translation as a model of ESL revision. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. 2010. pp. 1373-1381. Available from: <https://www.aclweb.org/anthology/C10-2157>
- [53] Mitchell CM, Evanini K, Zechner K. A trialogue-based spoken dialogue system for assessment of English language learners. In: *Proceedings of the International Workshop on Spoken Dialogue Systems*. CA: Napa; 2014
- [54] Leacock C, Chodorow M, Gamon M, Tetreault J. Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*. 2010;3(1):1-134
- [55] Dzikovska M, Nielsen R, Brew C, Leacock C, Giampiccolo D, Bentivogli L, et al. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In: *Proceedings 6th International Workshop on Semantic Evaluation*. 2013. pp. 263-274. Available from: <https://www.aclweb.org/anthology/S13-2045>
- [56] Van Rosmalen P, Eikelboom J, Bloemers E, Van Winzum K, Spronck P. Towards a game-Chatbot: Extending the interaction in serious games. In: Felicia P, editor. *Proceedings of the 6th European Conference on Games Based Learning*. 2012. pp. 525-532
- [57] Benotti L, Martínez M, Schapachnik F. Engaging high school students using chatbots. In: *Proceedings of the 2014 Conference on Innovation & Technology In Computer Science Education - Iticse '14*. 2014. pp. 63-68. DOI: 10.1145/2591708.2591728
- [58] Pereira J. Leveraging chatbots to improve self-guided learning through conversational quizzes. In: *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM '16*. 2016. DOI: 10.1145/3012430.3012625
- [59] Kerlyl A, Hall P, Bull S. Bringing Chatbots into education: Towards natural language negotiation of open learner models. *Applications and Innovations in Intelligent Systems XIV*. 2007;20(2):179-192. DOI: 10.1007/978-1-84628-666-7\_14
- [60] Gharehchopogh F, Khalifelu Z. Analysis and evaluation of unstructured data: Text mining versus natural language processing. In: *5Th International Conference on Application of Information and Communication Technologies (AICT)*. 2011. DOI: 10.1109/icaict.2011.6111017
- [61] Burstein J. Opportunities for natural language processing research in education. *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer; 2009. pp. 6-27. DOI: 10.1007/978-3-642-00382-0\_2
- [62] Wasserblat M, Pereg O, Singer G. Future Directions for NLP in Commercial Environments. 2019. Available from: <https://www.intel.ai/future-directions-nlp/#gs.tnq0fe> [Accessed: 23 January 2020]



- [63] Bell P. The future of natural language processing. Flatiron School. 2019. Available from: <https://flatironschool.com/blog/the-future-of-natural-language-processing> [Accessed: 22 January 2020]
- [64] Tractica. Natural Language Processing Market to Reach \$22.3 Billion by 2025. 2019. Available from: <https://www.tractica.com/newsroom/press-releases/natural-language-processing-market-to-reach-22-3-billion-by-2025/> [Accessed: 22 January 2020]
- [65] Ghosh P. The future of NLP in data science. Dataversity. 2018. Available from: <https://www.dataversity.net/future-nlp-data-science/> [Accessed: 22 January 2020]
- [66] Cynthia M. Grounded language learning: Where robotics and NLP meet. Invited talk. In: Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, Sweden; 2018
- [67] Shruthi J, Swamy S, Sarika C. Survey on - socially intelligent robots by using NLP. International Journal of Computer Applications. 2017;171(1):19-21. DOI: 10.5120/ijca2017914731
- [68] Bahja M. Identifying Patient Experience from Online Resources Via Sentiment Analysis and Topic Modelling Approaches, ICIS of the Association for Information Systems (AIS); 2018
- [69] Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. In: IEEE/ACM International Conference on Big Data Computing. USA: Applications and Technologies; 2016
- [70] Bahja M, Hammad R. A user-centric design and pedagogical-based approach for mobile learning. In: International Conference on Mobile, Hybrid, and on-Line Learning. 2018. pp. 100-105
- [71] Bahja M, Hammad R. Talk2Learn: A framework for Chatbot learning. In: European Conference on Technology-Enhanced Learning. 2019
- [72] May P. My First Usage of Natural Language Processing (NLP) in Industry. 2019. Available from: <https://towardsdatascience.com/my-first-usage-of-natural-language-processing-nlp-in-industry-c20842b37cad> [Accessed: 14 February 2020]
- [73] Maruti T. Re-Modelling the Hospitality Industry with Artificial Intelligence, Predictive Analytics & NLP. 2018. Available from: <https://chatbotsmagazine.com/re-modelling-the-hospitality-industry-with-artificial-intelligence-predictive-analytics-nlp-e875fff604b8> [Accessed: 14 February 2020]
- [74] Verma A. How AI Is Transforming the Media & Entertainment Industry. 2019. Available from: <https://www.wipro.com/en-IN/holmes/how-ai-is-transforming-the-media-and-entertainment-industry/> [Accessed: 14 February 2020]



# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Using Multi-Agent System to Govern the IT Needs of Stakeholders

*Chakir Aziza and Chergui Meriyem*

## Abstract

Many organizations spread and integrate the practices of the Information Technology Governance, Risk and Compliance (IT GRC). The problem that arises is how to choose the best practice to satisfy a precise need. This chapter concerns the study and the conception of decision-making architecture with the multi-agent system (MAS). So, the objective of this research is to build a decision-making model to satisfy a precise IT need. The proposed approach rests on four main stages to set up the decision-making model, which takes as input the strategic needs. The realized work has as objective to minimize the incoherence between the decisions taken by the stakeholders of an organization compared with the defined strategic objectives. The decision-making would contribute to legitimize the taken decision. This work is based on modeling a MAS, which rests on the idea that it is possible to represent directly the behavior and the interactions of a set of autonomous individuals evolving in a common environment. Finally, the proposed solution is part of a global platform for IT Governance, Risk and IT Compliance (EAS-IT GRC) (“EAS is the name of our team”).

**Keywords:** IT mapping, best IT framework, decision-making system, IT governance

## 1. Introduction

The development of a strategic vision of an organization is considered as a requirement for the information systems, direction, especially with regard to its contribution to the global performance of the organization. The smooth running of the information system of an organization, their evolution and their effective improvement of the quality of services of information technology, is reasoned by the multiplicity and the diversity of the best practice and of the different methods used [1, 2]. The main actors of an organization use a set of the IT directives which can be COBIT (Control Objectives for Information and related Technology) [3], for the executive management, ITIL (Information Technology Infrastructure Library) [4, 5], for the management of information systems and the series of the standards ISO (International Organization for Standardization) 27000 for the security of information systems [6].

Many companies are trying to integrate or implement an IT framework to meet a strategic need defined by stakeholders, with regard to the global conditions and options available. The challenge is how to select good practices given the diversity of IT GRC methods, frameworks, and best practices that exist in the IT market. The managers face a strategic difficulty of choosing the adequate IT framework, to meet the stakeholders’ needs [7].

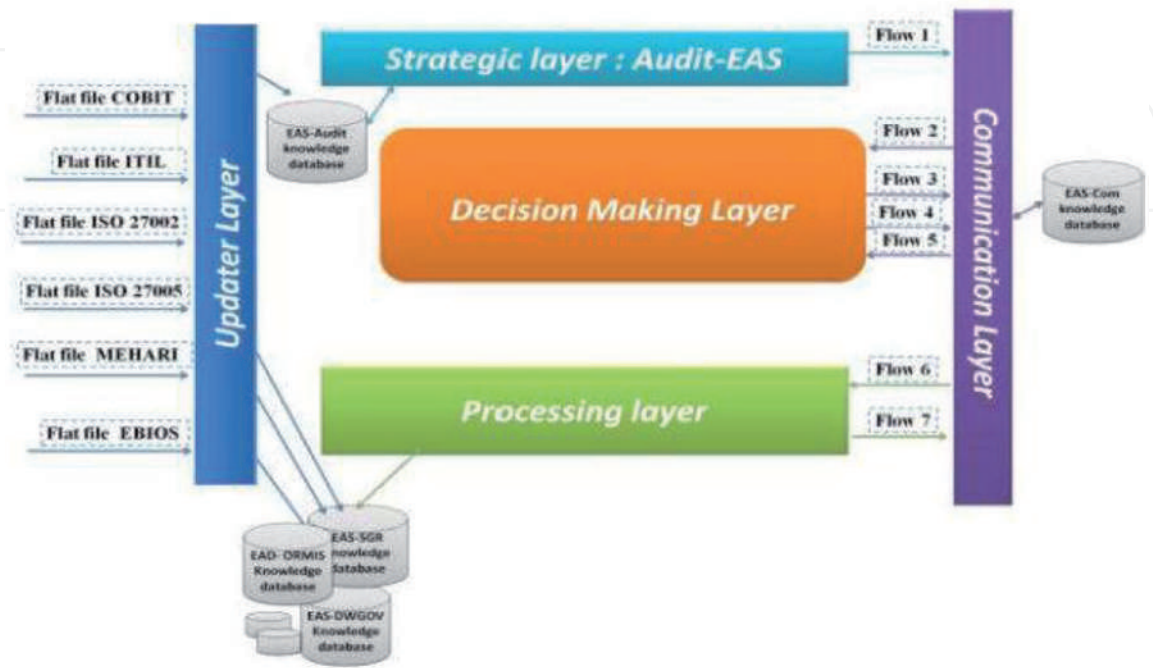
In this scientific work, we describe a correspondence between the strategic objectives of an organization and the processes of ITIL [8], PMBOK, ISO 27001, and ISO 27002 [9] by basing itself on a decision-making system to select the best framework by report a strategic objective. Furthermore, given the importance of interaction, coordination, and collaboration in information systems, we have to propose a solution that answers these essential requirements for the appropriate functioning of an organization.

## 2. Global architecture: EAS-IT GRC

The global platform of the IT GRC (EAS-IT GRC) ensures the alignment of the objectives of the organization as regards the needs defined by the stakeholders. This is illustrated by the progress strategic to a given organization and also by the taken decisions. The proposed solution supplies a high-level model for the IT GRC, which will allow the implementation of the IT GRC in an intelligent way. We give a brief description of every layer of the platform EAS-IT GRC (**Figure 1**) for a good understanding of the global architecture.

The architecture consists of five layers:

- **Strategic layer:** This layer ensures a permanent IT strategic alignment with needs defined by the stakeholders in an interactive intelligent way.
- **Decision-making layer:** It makes a study of the strategic request to release the best adequate framework to handle the request by the processing layer.
- **Processing layer:** It sets up the various reference tables, and it makes the treatment of the strategic request by the system that corresponds to the chosen platform. This layer arranges intelligent systems that translate the handled reference tables.



**Figure 1.**  
EAS-IT GRC architecture.

- **Communication layer:** All communications between the various layers is supported by this layer. It prepares messages exchanged according to formats required by the other layers. This layer spreads numerous mechanisms to make an exchange in real time.
- **Updater layer:** This layer ensures the adaptation of the new practices in the platform EAS-IT GRC by defining policies of change to integrate these new practices regarding information system.

### 3. Decision-making layer (DML)

The IT GRC market has expanded from a tactical focus on statutory compliance to a strategic focus on enterprise risk management. Many companies are trying to integrate or implement best practices to address a strategic need identified by stakeholders.

The challenge is how to select good practices given the diversity of IT GRC methods, standards, frameworks, and best practices that exist in the IT market. So the diversity of best practices poses a strategic challenge for companies to choose the right IT GRC best practice. As a result, our approach aims at choosing a good IT practice in an effective way by estimating IT frameworks with regard to the IT objectives.

The model of decision-making, DML, receives as input the IT service prepared and sent by the communication layer; this layer makes a good treatment of the IT service in an intelligent way. It has two levels, the first level generates the choice of the best reference table according to the strategic need and the second level is going to allow us to ensure the satisfaction of the chosen solution by basing itself on performance indicators communicated after every treatment made.

The following plan illustrates the proposed model of decision (Figure 2):

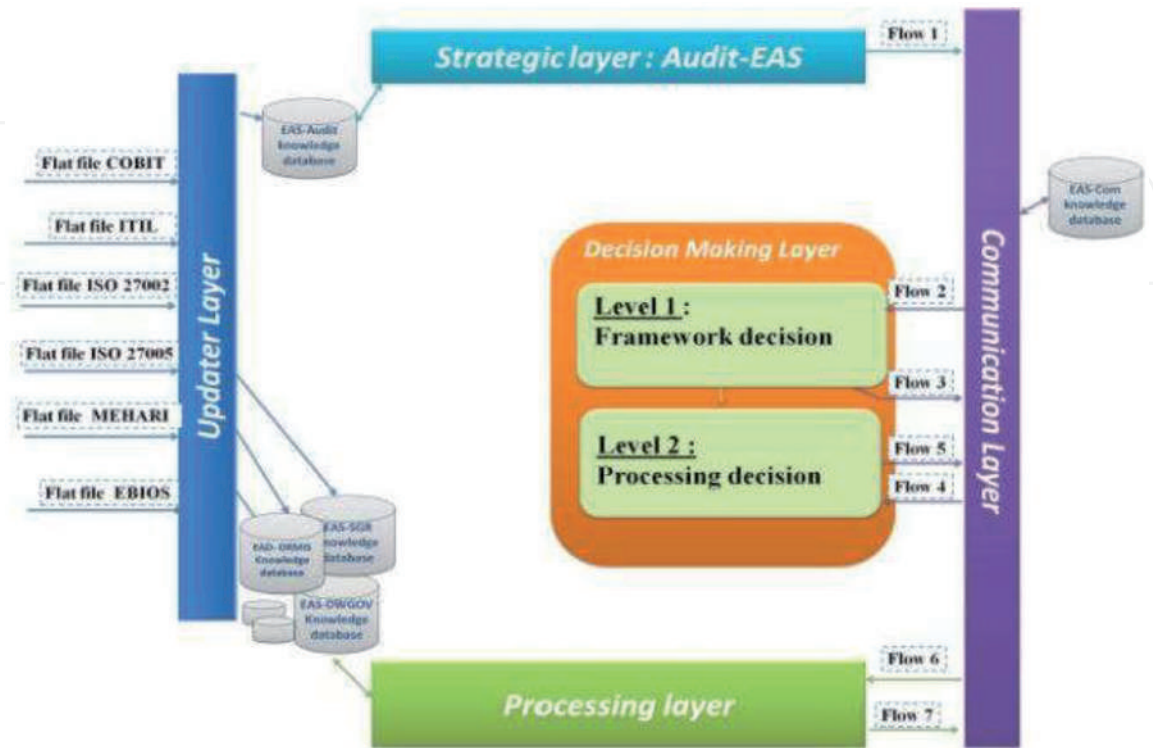


Figure 2.  
The model of decision-making DML.



3.1 Level 1: framework IT decision

The first level set up two layers, every layer has a precise function to achieve.

- The first layer is “MAS sequencing” based on MAS; it ensures that these sub-problems are scheduled according to the environment variables (the type of organization, the priority of one IT requirement over another, etc.), and also it has three sub layers “Categorization decision 1.1”, and “Sequencing.”
- The second layer is “MAS evaluation collective” based on MAS; it formalizes each sub-problem by taking into consideration the versions of the references, the certification or not of the employers of the organization, and it also takes other performance indicators into consideration.

The following plan schematizes the details of level 1 (Figure 3):

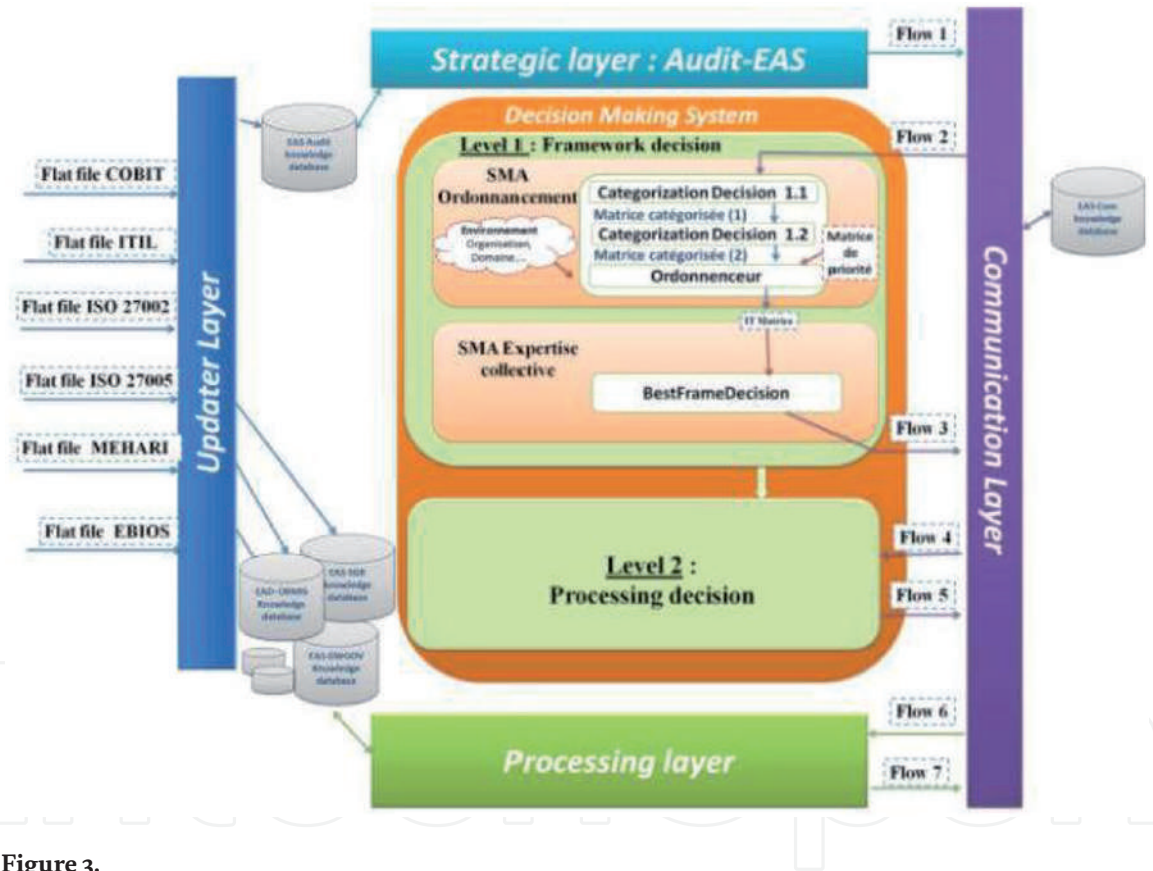


Figure 3.  
Level 1 “Framework IT decision.”

3.1.1 MAS sequencing

3.1.1.1 Algorithm: categorization Decision 1.1

The first sub layer “Categorization Decision 1.1” has as objective to make the connection between the strategic needs expressed by COBIT and the IT matrix to produce a reduced matrix, which will be handled by the second sub layer of the model of decision-making (Figure 4).

The categorization was made according to a strategy of selection of IT framework. The following graph shows the sequence of the stages of the generation of the categorized IT request.

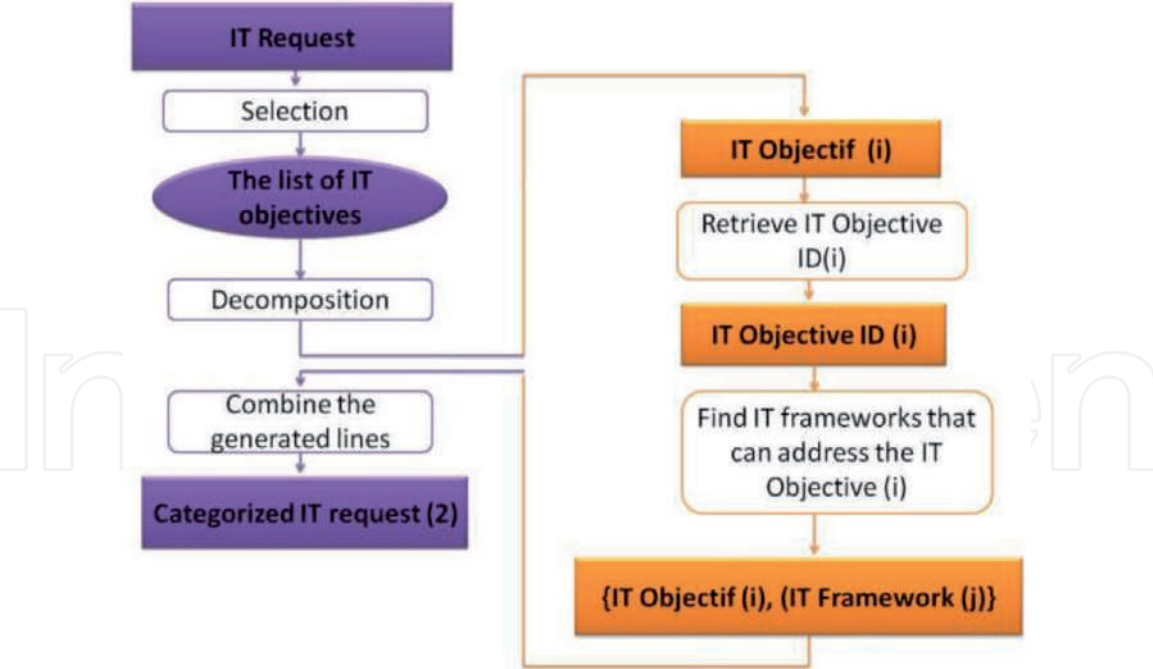


Figure 4.  
Graph of resolution of the algorithm “Categorization Decision 1.1.”

3.1.1.2 Algorithm: categorization Decision 1.2

The second sub layer “Categorization Decision 1.2” takes as input the matrix produced by the first one under layer, the type of activity of the organization, and the IT matrix we allocate a weight to every reference table which can answer a strategic need. The second categorization was made according to a strategy of evaluation of the IT objectives with regard to every IT framework. The following graph shows the chain of the stages of the regeneration of the categorized IT request (Figure 5).

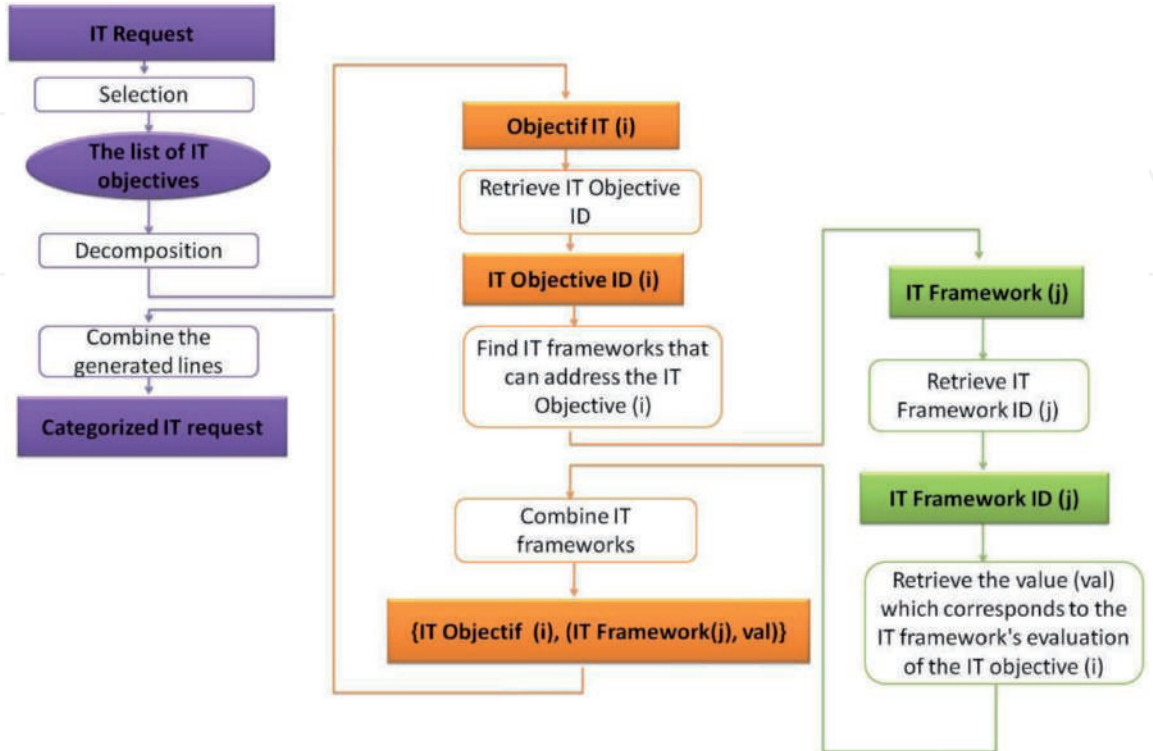


Figure 5.  
Graph of resolution of the algorithm “Categorization Decision 1.2.”

3.1.1.3 Algorithm: sequencing

The strategic need expressed by the stakeholders is seen as a set of IT objectives, the selection of which is managed by a particular algorithm. This algorithm makes the sequencing of the IT objectives to be treated (Figure 6).

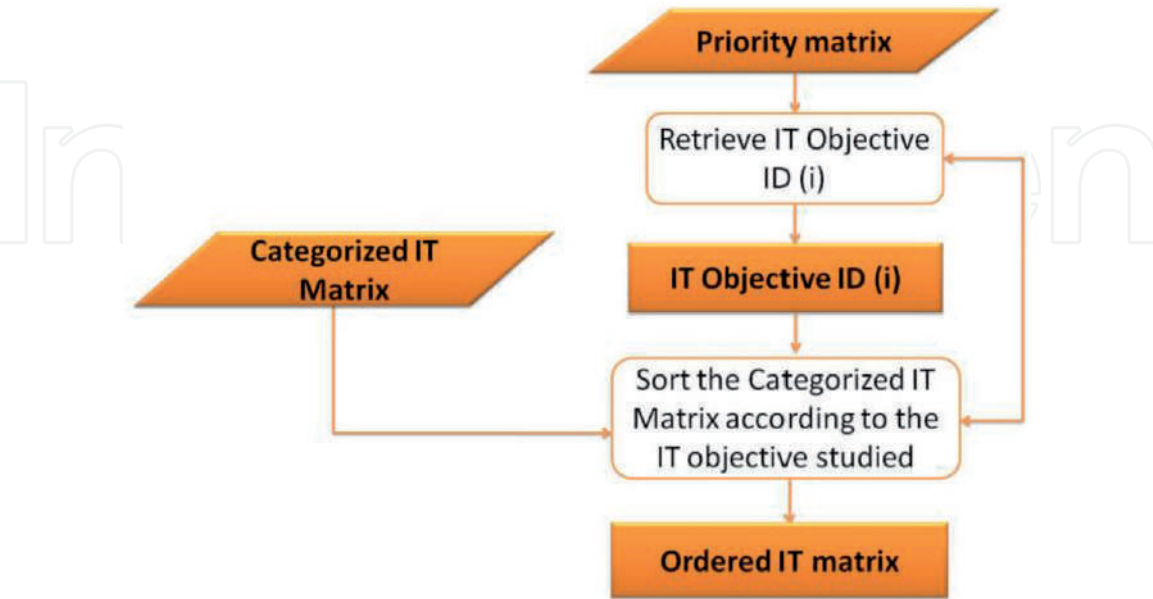


Figure 6.  
Graph of resolution of the algorithm “Sequencing.”

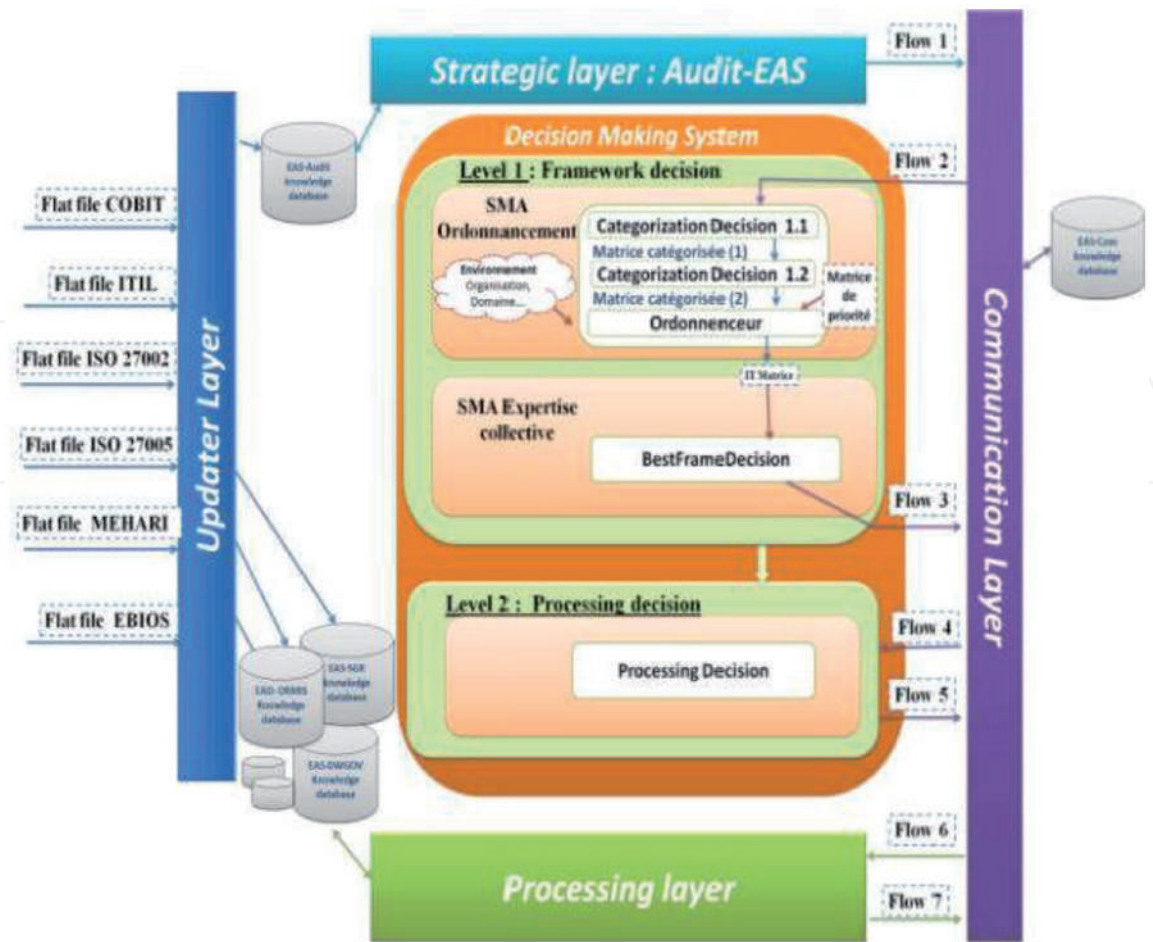


Figure 7.  
Both levels of DML.

3.1.2 MAS evaluation collective

The second layer “MAS evaluation collective” takes as input the matrix produced by the first layer, that present the list of the IT objectives, after it treats each IT objective as an under IT problem. In an intelligent way, this layer generates the best choice of the IT GRC practice to treat the IT need expressed as input. It takes as input the IT matrix, the versions of reference tables, the certification of organization’s employers, and the performance indicators based on the logging of treatments made by all the layer of the platform EAS-ITGRC.

3.2 Level 2: decision-making treatment

The second level ensures the satisfaction or the dissatisfaction of the choice by basing itself on a set of performance indicators, for example on the success rate—thus, if the success rate is greater than a threshold, the choice is good and if it is not the case, we have to reformulate the strategic needs by the user, or send back the second choice to the communication layer.

The following plan illustrates both levels in detail (Figure 7).

4. Contribution: framework IT decision

4.1 Method proposed in version 1

The IT GRC remains an emergent subject in the world of information technology (IT). However, to this day, there is a lack of research on a selective approach of IT framework with regard to the IT objectives.

4.1.1 Principle of functioning

The first version of our approach describes the correspondence between the processes IT of COBIT and the processes of ITIL, PMBOK, ISO27001, and ISO27002 to investigate into the coexistence of the links between the processes IT of COBIT and the processes IT by basing itself on the synthesis of the positioning of surrounding areas of the information technologies.

The following table illustrates the link between the IT objectives and the IT matrix (Table 1).

In this scientific work, we make the link between the IT processes of COBIT and the handled frameworks (ITIL, ISO 27001, ISO 27002 and PMBOK) by affecting

IT Objectives	Reference of the IT GRC IT			
	ITIL	PMBOK	ISO 27002	ISO 27001
PO 1 : Define a strategic IT plan	IT Evolution, 2	IT Evolution, 1 Business Evolution, 1	Null	Null
SE 3 : Ensure compliance with external obligations	Null	All, 1	All, 2	All, 3
DS5 : To ensure system security	Network management, 1	Null	All, 2	All, 1

Table 1.  
IT matrix.



two values (the first value corresponds to the key points that the IT framework will can treaties, and the second value corresponds to the classification of an IT framework by report the others) if there is a link between both and if it is not the case we affect the value “Null”.

4.1.2 Example

For handling the objective “Define a strategic IT plan,” we have two IT frameworks: ITIL and PMBOK.

- For ITIL, we have for weights: IT evolution, 2
- For PMBOK, we have for weights: IT evolution, 1 and business evolution, 1
- PMBOK covers more aspects than ITIL, thus we choose PMBOK.

The best framework is PMBOK.

4.1.3 Limitations

According to the proposed approach, the best IT framework for the IT objective “Define a strategic IT plan” is PMBOK and according to the IT expert the best reference table to handle the objective, thus we notice well that this approach does not present the best proposal.

4.2 Method proposed in version 2

To make the generated decision more efficient, we have to integrate the five pillars of IT governance in our decision-making model. The IT governance’s pillars are the value creation, the strategic alignment, the measure of performance of the processes, the resource management and skills and the management of the IT risk.

This method generates the decision of the adequate IT framework IT on two levels.

4.2.1 Principle of functioning

4.2.1.1 Stage 1: the IT objectives and the axes of the IT governance

For each of the COBIT IT processes (the 34 processes), a description is provided, with objectives and key indicators in the form of a cascade [3, 10].

This scientific research is based on the coexistence of links between IT objectives (COBIT processes) and the axes of the IT governance [11] by indicating the type of relation as “primary” (P or 2) or “secondary” (S or 1) [12, 13].

We apply the same treatment for the set of the IT processes and obtain the matrix below (**Table 2**) [12].

4.2.1.2 Stage 2: the IT objectives and the IT frameworks

The second phase consists in estimating the link between the IT objectives (the processes of COBIT) and IT framework by basing itself on aspects approached by every IT framework (**Table 3**) [12].

IT governance axe IT Objectives	Strategic Alignment	Value Deli- very	Risk Man- agement	Resource Manage- ment	Performance measurement
PO 1: Define a strategic IT plan	2	0	1	1	0
PO 2: Define the information architecture	2	1	1	2	0
PO 3: Determine the technological direction	1	1	1	2	0
PO 4: Define the processes, organization and labor relations	1	0	2	2	0
PO5: Managing IT investments	1	2	0	1	1

Table 2.  
The evaluation of IT objectives to IT axes.

IT Objectives	ITIL	PMBOK	ISO 27001	ISO 27002
PO 1: Define a strategic IT plan	3	0	0	0
PO 2: Define the information architecture	1	0	2	2
PO 3: Determine the technological direction	1	0	2	2
PO 4: Define the processes, organization and labor relations	3	0	3	5
PO5: Managing IT investments	3	2	1	1

Table 3.  
The relation between the framework COBIT and the other IT framework.

4.2.2 Objective function

In our research, we do not give a favorable opinion of one of the IT methods; we define an objective function of maximization based on five selection criteria, to select the best IT methods to treat the strategic need.

The proposed function is defined as follows:

$$f(\text{ObjectifIT}) = \text{Max}(\sum_{k=0}^5 \text{val}(x^k) * g(\text{framework}))$$
$$g(\text{framework}) = \text{val}(\text{framework})$$

- $x$ : It corresponds to the axes of the IT governance, it belongs to {Strategic Alignment, Value Delivery, Risk Management, Resource Management, Performance measurement}
- $\text{val}(x^i)$ : It is the value of criteria “ $x^i$ ” compared with the IT objective.
- $g(\text{framework})$ : It is the function that turns the value of one IT method by reporting an IT objective.

- $\text{val}(\text{framework})$ : It is the value of framework “framework<sub>i</sub>” compared with the IT Objective.

4.2.3 Example

Let us suppose that an organization wishes to manage its third services and she wants to set up an IT framework to realize this IT objective.  
To select the best IT framework, we apply the function  $f$ . The IT objective is “Manage third-party services” (DS5)

$$\sum_{k=0}^5 \text{val}(x^k) = \text{Val}(\text{Strategic Alignment}) + \text{Val}(\text{Value Delivery}) + \text{Val}(\text{Risk Management}) + \text{Val}(\text{Resource Management}) + \text{Val}(\text{Performance measurement})$$
$$\sum_{k=0}^5 \text{val}(x^k) = 0 + 2 + 2 + 1 + 1 = 6$$

and  
 $g(\text{ITIL}) = 3$ ;  $g(\text{PMBOK}) = 2$ ;  $g(\text{ISO27001}) = 2$ ;  $g(\text{ISO27002}) = 3$ . Then,  $f(\text{Manage third-party services}) = \text{Max}(18, 12, 12, 18) = 18$ . Then the best reference table to realize this objective is ITIL or ISO 27002.

4.2.4 Limitations

According to the proposed approach, the best IT framework to handle the objective “Manage third-party services” can be ITIL or ISO 27002 so we have two generated decision. Thus we notice that the taken decision is not the best one.

4.3 Method proposed in version 3

The improvement of the taken decision requires the addition of performance indicators at the end to consider the environment of the organization that wishes to implement an objective IT.  
This proposal is enriched by an expert system that is going to allow us to reproduce the reasoning and to deduce new knowledge, exploiting the performance of data warehousing.

4.3.1 Principle of functioning

4.3.1.1 Data warehouse

The basic function of using a decision-making system is to manage the journaling of data. So for our approach we integrate the datawarehousing as a decision-making system to evaluate the process of choosing the best practice by defining a set of performance indicators, which are [14]:

- Estimated rate: it is the time considered for treating an IT request.
- Use rate corresponds to the execution of an IT framework.
- Resolution rate corresponds to the time when an IT framework answered well to an IT request.

We proposed the following dimensions: Organization, Decision, IT Objective, IT service, Reference table, and Time.

#### 4.3.1.2 Expert system

For good exploitation of data stored in the multidimensional base, we have set up a system with knowledge, expert system. This system has as purpose the modeling of the knowledge and the reasoning of an expert in the field of IT GRC.

For that purpose, three main actors should present their contribution to the development of expert system, which are worth knowing: the end user, the IT expert, and the engineer of knowledge.

The interaction between these three actors is used to develop the expert system, which consists of a knowledge base, a facto base, and an interference engine.

The inference rules that we proposed are:

Let “Ai” be the IT framework and “O” the IT Objective

- **Rule 1:**

R1:  $\text{Weight (Ai)} > \text{Weight (Aj)} \rightarrow \text{Ai}$

- **Rule 2:**

R2:  $\text{Certification (organization, Ai)} = \text{true and Ai} \in \text{selected IT frameworks} \rightarrow \text{Ai}$

- **Rule 3:**

R3:  $\text{Estimated rate (Ai)} < \text{Estimated rate (Aj)} \rightarrow \text{Ai}$

- **Rule 4:**

A4:  $\text{Use rate (Ai)} > \text{Use rate (Aj)} \rightarrow \text{Ai}$

- **Rule 5:**

R5:  $\text{Resolution rate (Ai)} > \text{Resolution rate (Aj)} \rightarrow \text{Ai}$

- **Rule 6:**

A6:  $\text{Data warehouse (O)} = \text{true and IT Matrix (O)} = \text{Aj and Data warehouse (O)} = \text{Ai} \rightarrow \text{Ai}$  (Data warehouse (O) = true equivalent the objective (O) exists in the data warehouse)

- **Rule 7:**

R7:  $\text{IT Matrix (O)} = \text{Aj and Data warehouse (O)} = \{\text{Ai, Ak}\} \text{ and Version-Organization-IT framework (Ak)} = \text{V1 and Data warehouse-versioning (O, Ak)} = \text{V1} \rightarrow \text{Ak}$

- **Rule 8:**

R8:  $\text{Data warehouse (O)} = \text{false and IT Matrix (O)} = \text{Ai} \rightarrow \text{Ai}$ .

The latest version of our architecture decision-making model is as follows (Figure 8).



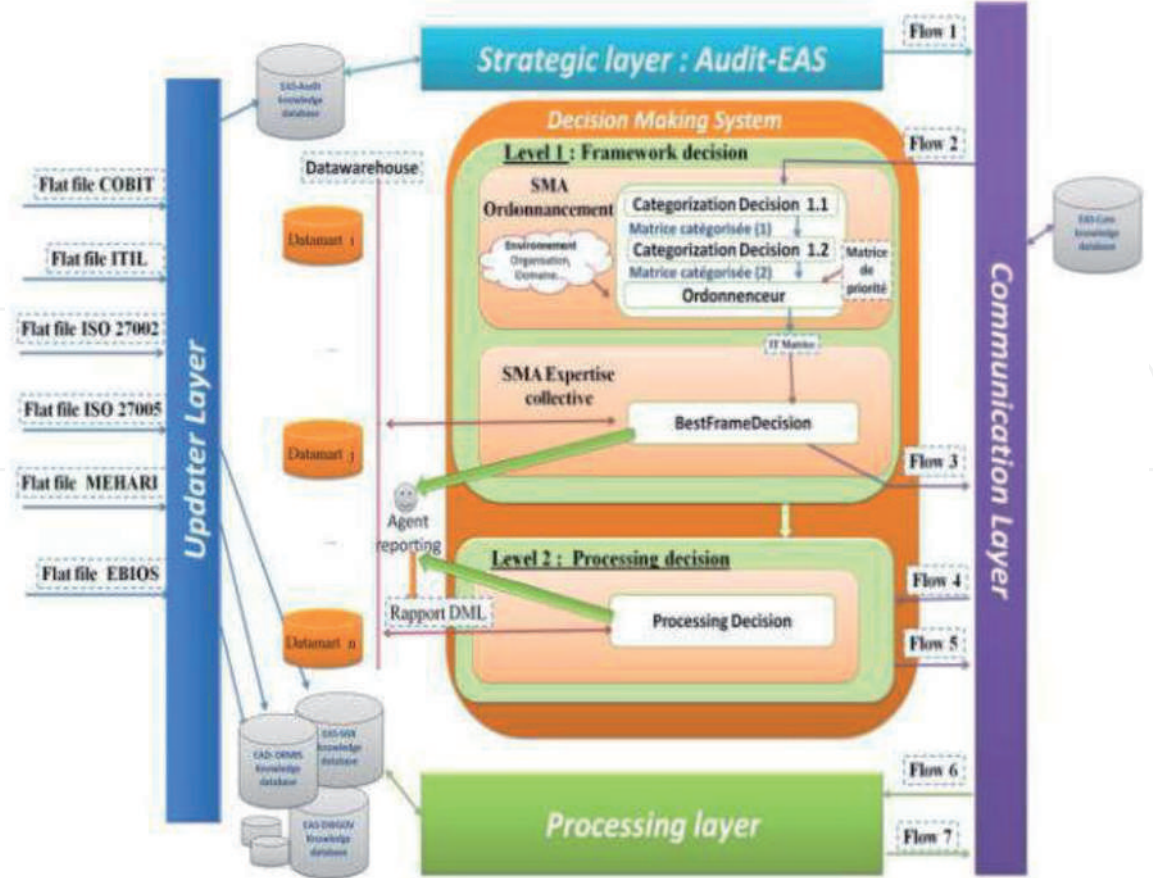


Figure 8.  
The detailed model of decision-making DML.

4.4 Case study

The following table shows a comparative study between the proposed approaches and the improvement of our model of decision-making, which is going to receive as input the following IT objectives (**Table 4**):

- Objective 1: Define a strategic IT plan
- Objective 2: Ensure compliance with external obligations
- Objective 3: Manage third-party services

IT Objective	O1: Define a strategic IT plan	O2: Ensure compliance with external obligations	O3: Manage third-party services
IT expert (Best IT Framework)	ITIL	ISO 27001	ISO 27002
Method proposed in version 1	PMBOK	ISO 27001	ITIL
Method proposed in version 2	ITIL	ISO 27001	ITIL OR ISO27002
Method proposed in version 3	ITIL	ISO 27001	ISO27002

Table 4.  
Comparative study.

The second line shows the opinion of an IT expert who gives the best IT framework by IT objective. For objective 1 (Define a strategic IT plan), the best IT framework is ITIL; for objective 2 (Ensure compliance with external obligations), the best IT framework is ISO 27001; and for objective 3 (Manage third-party services), the best IT framework is ISO 27002.

The third line shows the limitations of the first version of the proposed approach, taking the case of objective 1; it gave as results the IT framework PMBOK and it is not the best IT framework. The fourth line shows the limitations of the second version of the proposed approach, taking the case of objective 3; it gave as results the two IT frameworks ITIL and ISO27002 and they are not the best IT framework.

For the third version of our decision-making model, which corresponds to the fifth line, the results for the first objective give ITIL, for the second objective give ISO27001, and for the third objective give ISO27002.

We noticed that there is a correspondence between the results generated by our decision-making model and IT frameworks chosen by the IT expert.

## 5. Conclusion

The objective of our research is to build a model of decision-making to satisfy a precise IT need. The proposed approach integrates two disciplinary aspects, which are the data warehousing and the practices of the IT GRC to make the best decision.

We plan to add an extra layer that keeps the efficiency of the taken decisions by handling the factors influencing the quality of data [15, 16], or rather the moderating factors influencing the quality of basic data processed by a data warehouse [17].

We also plan to propose a generic approach that can integrate any IT methods and incorporate more performance indicators to make the appropriate choice for every organization.

### Author details

Chakir Aziza<sup>1\*</sup> and Chergui Meriyem<sup>2</sup>

<sup>1</sup> Faculty of Law Economics and Social Sciences, Hassan II University, Casablanca, Morocco

<sup>2</sup> Higher National School of Electricity and Mechanics, Hassan II University, Casablanca, Morocco

\*Address all correspondence to: [aziza1chakir@gmail.com](mailto:aziza1chakir@gmail.com)

### IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Kooper MN, Maes R, Roos Lindgreen EEO. On the governance of information: Introducing a new concept of governance to support the management of information. *The International Journal of Information Management*. 2011;31(3):195-200. <http://www.ibimapublishing.com/journals/CIBIMA/c>
- [2] Pereira R, da Silva MM. Governance implementation: The determinant factors. *Communications of the IBIMA*. 2012;2012:970363. 16p
- [3] COBIT® 4.1. IT Governance Institute®. Available from: <http://www.isaca.org/>
- [4] Väyrynen H. Improving IT service of delivering the field service management service and device for users in Company X [thesis]; 2019
- [5] Alastair B, Kevin J, Salah K, Maureen T, Adrie S. ITIL adoption in South African: A capability maturity view. *The e-Skills for Knowledge Production and Innovation Conference*; 2014
- [6] Wilai S, Vasin C. A benchmarking study of standard frameworks for information technology governance. *The Second Asian Conference on Information Systems*. ACIS; 2013
- [7] Nicolas R, Edgar Weippl, Andreas Seufert "A process model for integrated IT governance, risk, and compliance management". *Proceedings of the Ninth International Baltic Conference on Databases and Information Systems (DB&IS 2010)*. Riga: University of Latvia Press; 2010. pp. 155-170
- [8] Günther LC et al. Data quality assessment for improved decision-making: A methodology for small and medium-sized enterprises. *Procedia Manufacturing*. 2019;29:583-591
- [9] Maico G. Combining ITIL, COBIT and ISO/IEC 27002 for structuring comprehensive information technology for management in organizations. *Navus-Revista de Gestão e Tecnologia*. 2012;2(2):66-77
- [10] IT Governance Institute, editor. *COBIT Mapping: Overview of International IT Guidance*. 2nd ed. IT Governance Institute COBIT®; 2006. <https://www.sox-expert.com/uploads/files/COBIT%20Mapping%202nd%20Edition.pdf>
- [11] Chadi A, Savanid V, Yang C. IT Governance Framework Adoption: Establishing success factors. *International Working Conference on Transfer and Diffusion of IT*. Australia: University of Technology Sydney; 2011. Available from: [https://link.springer.com/chapter/10.1007/978-3-642-24148-2\\_15](https://link.springer.com/chapter/10.1007/978-3-642-24148-2_15)
- [12] Aziza C, Chergui M, Medromi H, Sayouti A. An approach to select effectively the best IT framework according to the IT governance axes, to handle and to set up an IT objective. *Third World Conference on Complex Systems (WCCS)*. IEEE. Marrakech: WCCS15; Novembre 2015
- [13] Ramlaoui S, Semma A. Comparative study of COBIT with other IT governance frameworks. *International Journal of Computer Science Issues*. 2014;11(6):1
- [14] Silva J et al. Intelligent and distributed data warehouse for Student's academic performance analysis. In: *International Symposium on Neural Networks*. Cham.: Springer; 2019. pp. 190-199
- [15] Azeroual O, Saake G, Abuosba M. Data quality measures and data cleansing for research information systems. *arXiv preprint arXiv:1901.06208*; 2019

[16] Cécile F, Fadila B, Omar B.  
Maintenance de charge pour  
l'optimisation des entrepôts de  
données évolutifs: aide à  
l'administrateur. In: EDA. 2008.  
p. 115-122

[17] Accerboni F, Sartor M. ISO/IEC  
27001. In: Quality Management: Tools,  
Methods, and Standards. Emerald  
Publishing Limited; 2019. pp. 245-264



# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Recent Advancements in Commercial Integer Optimization Solvers for Business Intelligence Applications

*Cheng Seong Khor*

## Abstract

The chapter focuses on the recent advancements in commercial integer optimization solvers as exemplified by the CPLEX software package particularly but not limited to mixed-integer linear programming (MILP) models applied to business intelligence applications. We provide background on the main underlying algorithmic method of branch-and-cut, which is based on the established optimization solution methods of branch-and-bound and cutting planes. The chapter also covers heuristic-based algorithms, which include preprocessing and probing strategies as well as the more advanced methods of local or neighborhood search for polishing solutions toward enhanced use in practical settings. Emphasis is given to both theory and implementation of the methods available. Other considerations are offered on parallelization, solution pools, and tuning tools, culminating with some concluding remarks on computational performance vis-à-vis business intelligence applications with a view toward perspective for future work in this area.

**Keywords:** integer programming, valid inequalities, local branching, relaxation induced neighborhood search (RINS), evolutionary algorithms, solution polishing

## 1. Introduction

The ongoing drive on Industrial Revolution 4.0 particularly to take advantage of big data analytics has impacted business intelligence applications significantly spanning various areas including resource assessment, corporate development, and advanced technology R&D research and development [1]. A key enabler supporting the transformation to digitalization is optimization technology which encompasses the established methodologies of linear and nonlinear programming with extensions to discrete or integer programming. This chapter focuses on recent advancements in commercial optimization solvers notably the industry-leading software package of IBM ILOG CPLEX [2] as applied to variants of integer programming problems particularly mixed-integer linear programming (MILP) models.

This chapter aims to contribute towards highlighting the growing and maturing capability of integer optimization especially in the last decade or so towards addressing, solving, analyzing, and eliciting insights from practical business intelligence applications. With rapid developments in the realm of big data analytics

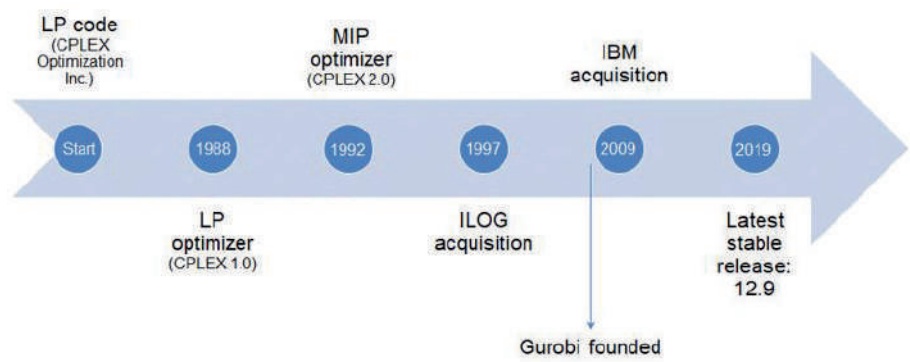
as spurred by Industry Revolution 4.0, advancement in optimization technology including integer optimization is imperative to support if not spearhead the changes at the forefront of the transformation taking place. The rest of the chapter is organized as follows. Section 2 gives an overview of the present role of integer optimization in business intelligence applications. Major solution methods and algorithms with certain enhanced features typically available in standard integer optimization solvers are detailed in Section 3 including those intended to exploit model formulations. Section 4 describes and discusses several real-world use cases on practical business intelligence applications that illustrate the applicability and strengths of integer optimization solvers. Finally, concluding remarks on the salient features of standard integer optimization solvers for business intelligence applications are offered including perspectives for future research directions.

## 2. Overview of integer optimization in business intelligence applications

Numerous business intelligence applications can be posed as mathematical programming problems that can be handled by commercial optimization solvers such as CPLEX, Gurobi [3], or KNITRO [4]. The problems can be formulated as models that include linear programming (LP), mixed-integer linear programming (MILP), quadratic programming (QP), mixed-integer quadratic programming (MIQP), quadratically-constrained programming, and mixed-integer quadratically-constrained programming. Such solvers are also used in tandem with other appropriate optimization solvers to handle other mainly nonlinear problems such as mixed-integer nonlinear programming (MINLP) models or in general, mixed-integer programs (MIP) [5].

### 2.1 Computational performance of commercial integer optimization solvers

The actual computational performance of a commercial optimizer (or optimization package) such as CPLEX results from a combination of improvement in several aspects. They include LP solvers with capability and features including preprocessing, algebra for sparse systems, solution methods (primal or dual simplex and barrier), and techniques to overcome degeneracy and numerical difficulties [6]. Equally important is the use of cutting planes as valid inequalities in solving problems that bridges the gap from theory to practice [7]. Further improvement involves applying heuristics including node heuristics (e.g., local branching, guided dives) and relaxation-induced neighborhood search, invoking evolutionary algorithms for solution polishing; and implementing parallelization for efficient computations [8].



**Figure 1.**  
*Historical background of IBM ILOG CPLEX integer optimization solver.*

Year	Activity/accomplishment
1988	Develops LP solver (CPLEX 1.0)
1992	Offers simple branch and bound with limited cuts (CPLEX 2.0)
1998	Incorporates simple heuristic; provides faster dual simplex (CPLEX 6.0)
1999	Introduces five node heuristics and six cutting plane types (CPLEX 6.5)
2000	Caters for semi-continuous and semi-integer variables; stipulates dual simplex as default LP solution method; introduces preprocessing; improved cuts (CPLEX 7.0)
2002	Introduces new LP method of sifting, concurrent optimization, new QP capabilities, and 9 cutting plane types (CPLEX 8.0)
2003	Introduces quadratic constraint programming (QCP) and relaxation induced neighborhood search (RINS) (CPLEX 9.0)
2006	Improved MIQP, changes in MIP start, feasible relaxation; introduces indicators and solution polishing features
2007	CPLEX 11.0 incorporates solution pool, tuning pool, and parallel mode
2010	Offers faster MILP solution; introduces multicommodity flow cuts; enhanced heuristics and dynamic search (in CPLEX 12.2)
2017	Enables faster MILP solution; enhanced CP Optimizer; new callback framework (in CPLEX 12.8)
2019	Includes handling of multiobjective problems; provides modeling assistance (in CPLEX 12.9)

**Table 1.**  
*Software release history of IBM ILOG CPLEX integer optimization solver.*

**2.2 A commercial success story: CPLEX integer optimization solver**

CPLEX is a state-of-the-art commercial integer optimization solver currently marketed by IBM. It represents an early commercial success story of an optimization package with various acquisitions and a spin-off solver (called Gurobi) which is now a success story of its own. **Figure 1** presents brief historical facts of CPLEX while **Table 1** summarizes the software release history.

**3. Solution methods and algorithms**

**3.1 Integer optimization algorithms**

A suite of algorithms is available in various integer optimization solver to exploit the underlying problem structure of a business intelligence application towards achieving efficiency and accuracy. **Table 2** summarizes the typical main algorithms employed by CPLEX according to the problem type identified together with remarks on the enhancement provided to increase computational performance [9].

**3.2 Branch-and-bound**

A general structure of mixed-integer program is given by:

minimize

$$c^T x$$

(1)

subject to

$$Ax = b$$

(2)



$$l \leq x \leq u \tag{3}$$

some  $x$  are integers. (4)

Branch-and-bound is a base algorithm to solve MIP which uses LP as a subroutine [10]. The key strategies of a branch and bound procedure involve splitting (i.e., branching) the solution space into disjoint subspaces, bounding the objective function values for all solutions in the subspaces, and pruning or fathoming nodes of branches that cannot yield better solutions. Although it is provably exponential in time, tricks are available to accelerate its search which mostly apply to a subset of models with a suite of algorithms available.

The branching strategies are performed on the integer variables and comprise two main steps: (1) Choose an integer variable as a branching variable  $x_j$ , (2) Split the problem into two submodels:  $x_j \leq i$  or  $x_j \geq i + 1$  where for the special case of binary variables, the problem becomes  $x_j = 0$  or  $x_j = 1$ .

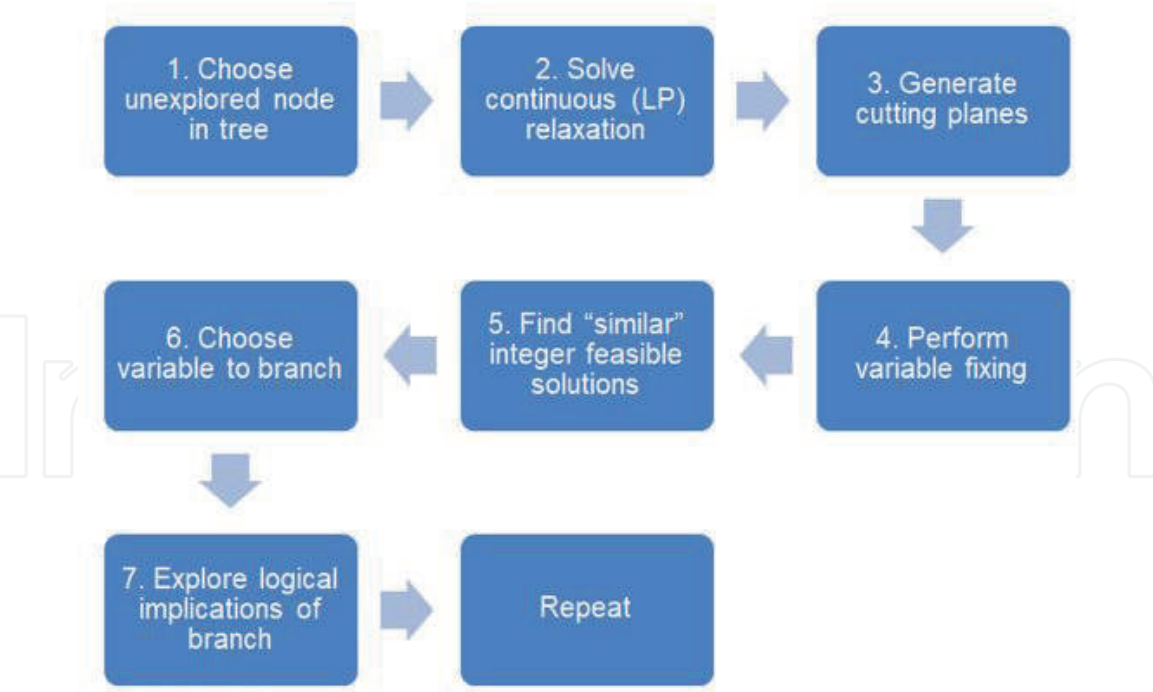
The bounding problem given by the continuous (LP) relaxation to determine a lower bound  $z_{IP}^L$  on the objective function value of the original MIP problem can be described as follows: minimize  $c^T x (= z_{IP}^L)$  subject to  $Ax = b, l \leq x \leq u$  (simple bounds), and some  $x_j$  are integers. The continuous relaxation problem gives solution of an optimal objective value of  $z_{IP}^L$ , which is a lower bound on the objective function value of the original MIP problem by relaxing the integrality restriction. There are two useful properties of continuous relaxation: (1) If its solution satisfies integrality restrictions, there is no need to further explore the subspace; (2) It offers natural branching candidates as the integer variables with fractional values in a relaxation solution.

Key steps in the branch-and-bound procedure are summarized in **Figure 2**. As described in **Figure 2**, node selection in step 1 involves a tradeoff between achieving feasibility and optimality. The options available for node selection include depth first, breadth first, best first, limited discrepancy, and best estimate. When exploring nodes deep in a search tree, one is more likely to find integer feasible solutions and explore nodes that would be pruned by later feasible solutions. The method called plunging (as combined with those aforementioned) always choose a child node of previously explored node.

In step 2, the node relaxation step is ideally suited to dual simplex method. It involves only a small change from the parent relaxation solution (at the root node) and gives a new bound on the branching variable while maintaining dual feasibility of the previous basis. Thus, the solution is likely to be close to the previous basis.

Solver/optimizer	Algorithm	Model type	Remark
Simplex	Primal, dual, network	LP, QP	Reoptimization with simplex algorithms is faster when starting from a previous basis
Barrier	Interior-point	LP, QP, QCP	Explore multiple threads presence Barrier optimizer cannot start from advanced basis—limited application in B&B for MILP
Mixed-integer optimizers	Branch and cut, dynamic search	MILP, MIQP, MIQCP	IBM proprietary/trade secret methodology to solve MIP (some user callbacks cannot be used)

**Table 2.**  
*Algorithms available in IBM ILOG CPLEX integer optimization solver.*



**Figure 2.**  
 Key steps in the branch-and-bound procedure [9].

Typically, a few dual simplex iterations are sufficient to restore optimality, and the cost per node is quite small. The subsequent step 3 entails generating cutting planes as needed to obtain a continuous (LP) relaxation solution.

Step 4 involves variables fixing using reduced cost. If the following condition as given by Eq. (5) holds at a branch-and-bound node:

$$z_{LP} + |D_j| \geq z^* \tag{5}$$

where  $z_{LP}$  = objective value of LP relaxation solution at the root node,  $z^*$  = objective value of an incumbent (i.e., best known integer feasible solution), and  $D_j$  = reduced cost (marginal cost of releasing a variable from its bound), then we apply the strategy of fixing  $x_j$  to its current value in this subtree of the search. The goal here as described by step 5 is to obtain integer feasible solutions which are similar to the relaxation solution.

Selecting an appropriate branching variable can significantly affect the search tree size, which is emphasized in the subsequent step 6. In this regard, the guiding principles are to make the important decisions early (as modeled by the integral branching variables) by being aware of the impact of both branching directions. To illustrate by using a factory building problem, such a decision involves whether to build a factory first while the decision on the number of lines to be placed in the factory can be made later. In general, we can predict the impact of a branch by considering variables that are furthest from their bounds which indicate maximum infeasibility. Thus, the impact for each branching candidate can be measured to allow for strong branching to be performed, e.g., by using historical information such as pseudo-costs.

Finally, in step 7, the main idea in propagating implications logically is to fix the binary variables to possible values during tree exploration and determine the binary variable values. Bound strengthening is used to tighten variable bounds.

Practical considerations render implementing branch-and-bound to be unsuitable for large scale problems chiefly because the number of iterations grows

exponentially with number of variables. Therefore in practice, a commercial business intelligence solver such as CPLEX uses a branch-and-cut procedure as a modification which applies model reformulation by using presolve strategies and adding cutting planes (or cuts) as shown in **Figure 3** with possible enhancement in practice around the root node computations [11].

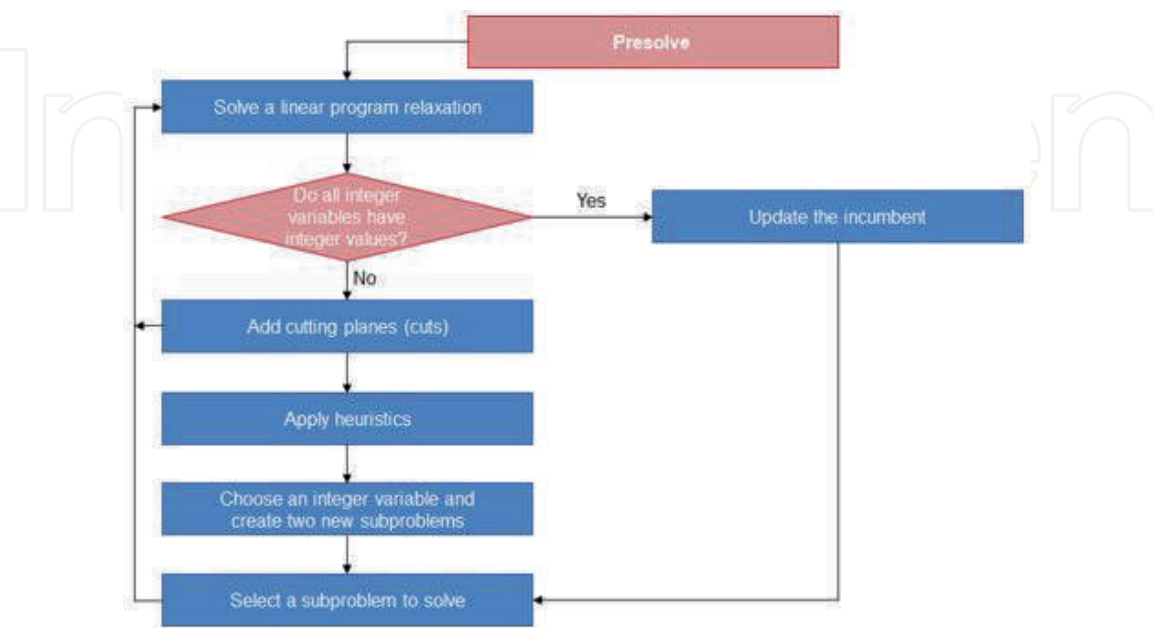
3.3 Presolve and cutting planes

The original MIP formulation can be improved by tightening it with fewer constraints and variables thus entailing less data handling requirement (yet with the same solution quality). A tighter formulation also leads to a smaller difference between the space of the feasible continuous and feasible integer solutions, hence relying less on branching to refine the continuous relaxation computation. Two techniques are used: (1) presolve which combines preprocessing and probing strategies [12, 13]; and (2) cutting planes [14].

Presolve generates a new tighter improved model without size increase that is independent of the relaxation solution. Preprocessing aims to identify feasibility and redundancy while improving bounds (e.g., through rounding) while that of probing improves coefficients by fixing the binary variable values while checking for their logical implications. In both cases, we achieve a tighter model reformulation using similar steps of adding or replacing constraints that maintain the same integer solutions but with fewer continuous relaxation solutions. Adding a single constraint can produce an exponential number of tighter constraints. Such tighter constraints dominate the existing constraints without creating a larger problem. Note that reformulation solution is different from that of relaxation.

In contrast, we add a cutting plane (or valid inequality) to an existing model (typically the presolve-reformulated model) to remove a relaxation solution—this feature constitutes an important difference between the two techniques. Therefore, cutting planes introduce tighter constraints that cut off a particular relaxation solution and in so doing, achieves focused growth in model size.

In summary, presolve is vital in solving MIP as there is significant scope to improve most model formulations through reducing problem sizes (by more than



**Figure 3.**  
*Branch and cut algorithm (CPLEX).*

5 times is not uncommon) or runtimes (similarly by up to 10 times). On the other hand, cutting planes are available in numerous varieties with many valid types applicable for a particular model. Thus we need to identify relevant ones which serve to cut off appealing relaxation solutions. There is a need to strike a balance in terms of how many cuts to generate for a relaxation solution. Since we need to cut off relaxation solution only once, and it is expensive to resolve in obtaining a new relaxation solution for each cut added, we conduct multiple rounds of cutting plane generation while limiting the number of cuts per round in view of the increased model size [15].

### 3.4 Heuristics

Heuristics for solving MIP aims to produce good and possibly feasible solutions quickly without relying on branching in satisfying user demands for a problem. Thus, heuristics avoid exploring unproductive subtrees (in a branch-and-cut scheme) while exploring parts of tree that a solver typically will not. In doing so, heuristics help to prove optimality explicitly by pruning nodes more efficiently as well as implicitly by giving integer solutions [16].

Heuristics can be classified into two classes as available in a solver like CPLEX: (1) plunging (diving) heuristics, and (2) local improvement heuristics which explore interesting neighborhoods around potential solutions using search strategies such as local branching, relaxation induced neighborhood search (RINS), guided dives, and evolutionary algorithms for solution polishing. Plunging heuristics maintains linear feasibility in trying to achieve integer feasibility while local improvement heuristics operate conversely [17]. A typical strategy for heuristics applied at the root node involves the sequence shown in **Figure 4**.

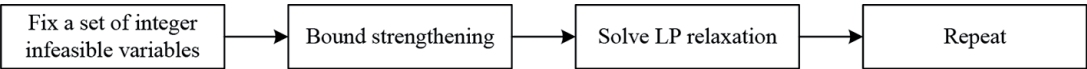
Some considerations in applying plunging heuristics include tradeoffs of how many variables to fix per computation round and in what order. While it is computationally inexpensive to fix all variables rather than a few variables, LP relaxation solutions in the latter (not needed in the former) can guide later choices (e.g., on variable values and reduced costs). Variations in variable fixing order can be useful for diversification. On the other hand, a high-level structure of local improvement heuristics involves choosing integer values for all the integer variables, which produces linear infeasibility; iterating over the integer variables; and applying infeasibility metrics [16].

The effectiveness of heuristics is evidenced in that feasible solutions are found for most models before branch-and-bound is performed. Approximately 10% improvement in computational time to proven optimality has been reported [16]. Furthermore, heuristics often get solutions not obtained by branching.

### 3.5 Combined local search and heuristics

A combination of local search and heuristics offers a powerful optimization framework to solve difficult MIP or combinatorial optimization problems. Examples of local search methods include simulated annealing, tabu search, and genetic algorithms. Local search methods consist of the key strategies of neighborhood (i.e., considers a set of solutions in the vicinity of current solution); intensification (i.e., temporary focus on part of solution space), and diversification (i.e., mechanism to change focus occasionally). In applying local search to MIP, generally neighborhoods are based on the problem structure, e.g., nodes and edges in graphs with no high level structural information available in arbitrary MIP models [16]. A question that arises is how we can generate and explore an interesting





**Figure 4.**  
*Heuristics at root node.*

neighborhood given an incumbent solution. In this regard, two methods are available, namely local branching [18] and relaxation induced neighborhood search (RINS) [19].

### 3.6 Parallelization

Parallelization is available in an integer optimization solver such as CPLEX, which encompasses the MIP solution engine, barrier algorithm, and concurrent optimization techniques for solving LP and QP problems. In the instance of CPLEX, parallelization involves launching several optimizers to solve the same problem—the process stops when the first solver reaches a solution. Within a branch and bound scheme, parallelization involves solution of the root node and nodes as well as strong branching in parallel [20].

### 3.7 Solution pools

The motivation to consider solution pools lies in the value of having more than one solution due to inaccurate data, approximations in model formulations, or inability of a model to capture the full essence of a problem. Thus, solution pools aim to generate and keep multiple solutions by using various options and tools that involve collecting solutions within a given percentage of optimal solution or those with diverse solutions and properties. However, difficulty is noted in implementing solution pools with the strategy of rolling horizon decompositions [17].

### 3.8 Tuning tools

As MIP solvers have multiple algorithm parameters which dictate their performance, the objective of tuning tool is to identify solver parameters that improve the performance for a given problem set. While default parameter values of MIP solvers are defined to work well for a large collection of problems, there is no such guarantee for a specific user problem [21].

## 4. Use cases

This section presents three use cases of applying commercial integer optimization solvers to implement and improve or enhance business intelligence applications. The model formulations for the use cases are implemented on GAMS modeling platform (and available in GAMS Model Library) from which the CPLEX solver is accessed.

### 4.1 Use case 1: energy optimization

The first use case presents a practical application of CPLEX as a standard solver for an energy business portfolio optimization problem for an electric utility company. For such electricity distribution public service, the problem involves to determine the amount to produce internally (i.e., in one's own power plant) and that to purchase

externally (i.e., from the spot market or load following contracts). The problem formulation leads to a medium-to-large scale MILP model with size and computational statistics as described in **Table 3**. To accelerate solution convergence, several computational options are invoked including priority branching within a branch-and-bound procedure and multiple processing through parallelization (i.e., techniques introduced in the foregoing section). The computational results and implications as discussed in the cited reference demonstrates the applicability of the solver as an effective tool for 1-day ahead planning within a real-world electricity market in Germany.

4.2 Use case 2: financial optimization

The second use case involves financial optimization of risk management with commercial implications . The problem is amenable to be posed as an integer optimization model to capture an extensive set of rules and regulations that governs the delivery and settlement of mortgage-backed securities. The availability of reliable, robust, and efficient commercial integer optimization solvers alongside computing technology developments have facilitated the deployment and validation of such models with the computational statistics summarized in **Table 4**. The advancement achieved has led to optimization models including (if not particularly) integer programs to become essential omnipresent tools in current financial operations, which is comparable to the application of operations research and management science models in the domains of manufacturing, transportation, and logistics.

Computing platform	GAMS 24.2.3 on laptop with Intel Core i7-8550U 1.80 (up to 1.99) GHz, 8 GB of RAM
No. of continuous variables	1260
No. of discrete variables	773
No. of constraints	2178
No. of iterations	2,799,216
CPU time	408.234 second
Objective function value	EUR266,793 (for optimality gap = 0%)

**Table 3.**  
*Model size and computational statistics for use case 1.*

Computing platform	GAMS 24.2.3 on laptop with Intel Core i7-8550U 1.80 (up to 1.99) GHz, 8 GB of RAM
No. of continuous variables	255
No. of discrete variables	199
No. of constraints	487
No. of iterations	238
CPU time	0.157 second
Objective function value	36.96 (for optimality gap = $1 \times 10^{-4}\%$ )

**Table 4.**  
*Model size and computational statistics for use case 2.*

Computing platform	GAMS 24.2.3 on laptop with Intel Core i7-8550 U 1.80 (up to 1.99) GHz, 8 GB of RAM
No. of continuous variables	81
No. of discrete variables	8
No. of constraints	73
No. of iterations	15
CPU time	0.016 second
Objective function value	36.96 (for optimality gap = 0 <sup>4</sup> %)

**Table 5.**  
*Model size and computational statistics for use case 3.*

4.3 Use case 3: manufacturing optimization

The third use case concerns production planning for a manufacturing facility . The application can be formulated as a standard integer optimization model of an uncapacitated lot-sizing problem. The objective function seeks to minimize production cost in meeting market demand constraints with cost components on production, stocking, and machine setups. **Table 5** gives the model size and computational statistics for the largest problem instance solved for this use case.

5. Conclusions

Performance variability across commercial integer optimization solvers applied to business intelligence applications (such as that for the use case in Section 4) occurs due to opportunistic parallelization, use of heuristics particularly by invoking polishing option (which involves random seed), or simply numerical reasons. Variability may be observed in computational time, performance in terms of number of nodes and iterations, or solution quality. A main limitation of the applicability of integer optimization solvers typically pertains to the number of integer variables that can be handled within acceptable computational load or solution time. Therefore, it is worthwhile for future research in this area to consider further improvement in the mentioned areas [22, 23] towards achieving acceptable performance levels that are requisite and crucial for business intelligence applications.

Acknowledgements

This work is completed partly under support from UTP-UCTS private grant no. 015MD0-037.

IntechOpen

## Author details

Cheng Seong Khor<sup>1,2</sup>

1 Chemical Engineering Department, Universiti Teknologi PETRONAS,  
Perak Darul Ridzuan, Malaysia

2 Centre for Process Systems Engineering, Institute of Autonomous Systems,  
Universiti Teknologi PETRONAS, Perak Darul Ridzuan, Malaysia

\*Address all correspondence to: [chengseong.khor@utp.edu.my](mailto:chengseong.khor@utp.edu.my);  
[khorchengseong@gmail.com](mailto:khorchengseong@gmail.com)

## IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 



## References

- [1] Tsay C, Baldea M. 110th anniversary: Using data to bridge the time and length scales of process systems. *Industrial & Engineering Chemistry Research*. 2019;**58**(36):16696-16708
- [2] IBM. IBM ILOG CPLEX Optimization Studio V12.9.0. 2020. Available from: [https://www.ibm.com/support/knowledgecenter/SSSA5P\\_12.9.0/ilog.odms.studio.help/Optimization\\_Studio/topics/COS\\_home.html](https://www.ibm.com/support/knowledgecenter/SSSA5P_12.9.0/ilog.odms.studio.help/Optimization_Studio/topics/COS_home.html)
- [3] Gurobi Optimization. Gurobi Optimizer Reference Manual. Beaverton, Oregon: Gurobi Inc.; 2020
- [4] Byrd RH, Nocedal J, Waltz RA. KNITRO: An integrated package for nonlinear optimization. In: Pillo GD, Roma M, editors. *Large-Scale Nonlinear Optimization*. Springer; 2006. pp. 35-59
- [5] Jünger M et al. 50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art. Berlin Heidelberg: Springer-Verlag; 2010. p. 804
- [6] Rardin RL. *Optimization in Operations Research*. New Jersey: Prentice-Hall; 1998
- [7] Williams HP. *Model Building in Mathematical Programming*. 4th ed. Chichester, West Sussex, England: John Wiley & Sons; 1999
- [8] Danna E. Performance variability in mixed integer programming. In: *Workshop on Mixed Integer Programming 2008 (MIP 2008)*. New York City, NY: Columbia University; 2008
- [9] Rothberg E. The CPLEX library: Mixed integer programming. In: 4th Max-Planck Advanced Course on the Foundations of Computer Science (ADFOCS 2003). Saarbrücken, Germany: Max-Planck-Institut für Informatik; 2003
- [10] Land AH, Doig AG. An automatic method of solving discrete programming problems. *Econometrica*. 1960;**28**(3):497-520
- [11] Lima RM, Grossmann IE. On the solution of nonconvex cardinality Boolean quadratic programming problems: A computational study. *Computational Optimization and Applications*. 2017;**66**(1):1-37
- [12] Savelsbergh MWP. Preprocessing and probing techniques for mixed integer programming problems. *ORSA Journal on Computing*. 1994;**6**(4):445-454
- [13] Wolsey LA. *Integer Programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Chichester: Hoboken, NJ: Wiley; 1998. pp. 203-258
- [14] Nemhauser G, Wolsey L. The theory of valid inequalities. In: *Integer and Combinatorial Optimization*. Hoboken, NJ: Wiley; 1988. pp. 203-258
- [15] Rothberg E. The CPLEX library: Presolve and cutting planes. In: 4th Max-Planck Advanced Course on the Foundations of Computer Science (ADFOCS 2003). Saarbrücken, Germany: Max-Planck-Institut für Informatik; 2003
- [16] Rothberg E. The CPLEX library: MIP heuristics. In: 4th Max-Planck Advanced Course on the Foundations of Computer Science (ADFOCS 2003). Saarbrücken, Germany: Max-Planck-Institut für Informatik; 2003
- [17] Rothberg E. An evolutionary algorithm for polishing mixed integer programming solutions. *INFORMS Journal on Computing*. 2007;**19**(4):534-541

[18] Fischetti M, Lodi A. Local branching. *Mathematical Programming.* 2003;**98**(1):23-47

[19] Danna E, Rothberg E, Pape CL. Exploring relaxation induced neighborhoods to improve MIP solutions. *Mathematical Programming.* 2005;**102**(1):71-90

[20] Lima R. IBM ILOG CPLEX: What is inside of the box? In: *Enterprise-Wide Optimization (EWO) Seminar*. Pittsburgh, PA: Carnegie Mellon University; 2010

[21] IBM. CPLEX Performance Tuning for Mixed Integer Programs. 2019. Available from: <https://www.ibm.com/support/pages/cplex-performance-tuning-mixed-integer-programs>

[22] Bixby R et al. MIP: Theory and practice—Closing the gap. In: *System Modelling and Optimization: Methods, Theory, and Applications*. Boston, MA: Kluwer Academic Publishers; 2000. pp. 19-49

[23] Bixby R, Rothberg E. Progress in computational mixed integer programming—A look back from the other side of the tipping point. *Annals of Operations Research.* 2007;**149**(1):37-41

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Recent Advances in Stock Market Prediction Using Text Mining: A Survey

*Faten Subhi Alzazah and Xiaochun Cheng*

## Abstract

Market prediction offers great profit avenues and is a fundamental stimulus for most researchers in this area. To predict the market, most researchers use either technical or fundamental analysis. Technical analysis focuses on analyzing the direction of prices to predict future prices, while fundamental analysis depends on analyzing unstructured textual information like financial news and earning reports. More and more valuable market information has now become publicly available online. This draws a picture of the significance of text mining strategies to extract significant information to analyze market behavior. While many papers reviewed the prediction techniques based on technical analysis methods, the papers that concentrate on the use of text mining methods were scarce. In contrast to the other current review articles that concentrate on discussing many methods used for forecasting the stock market, this study aims to compare many machine learning (ML) and deep learning (DL) methods used for sentiment analysis to find which method could be more effective in prediction and for which types and amount of data. The study also clarifies the recent research findings and its potential future directions by giving a detailed analysis of the textual data processing and future research opportunity for each reviewed study.

**Keywords:** machine learning, deep learning, natural language processing, sentiment analysis, stock market prediction

## 1. Introduction

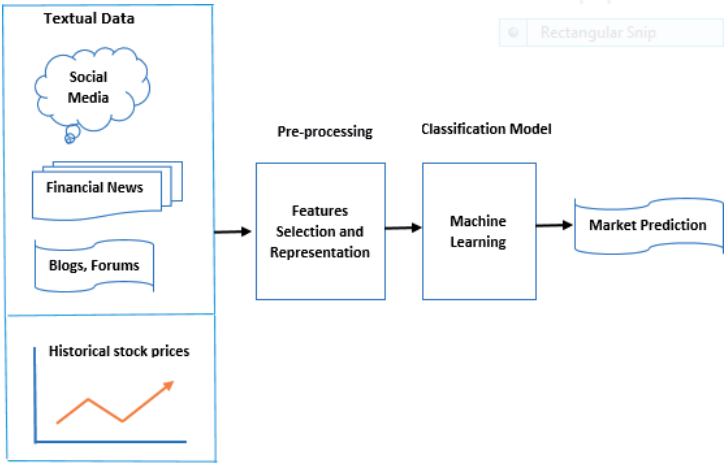
Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. Predicting how the stock market will move is one of the most challenging issues due to many factors that involved in the stock prediction, such as interest rates, politics, and economic growth that make the stock market volatile and very hard to predict accurately. The prediction of shares offers huge chances for profit and is a major motivation for research in this area; knowledge of stock movements by a fraction of a second can lead to high profits [1]. Since stock investment is a major financial market activity, a lack of accurate knowledge and detailed information would lead to an inevitable loss of investment. The prediction of the stock market is a difficult task as market movements are always subject to uncertainties [2]. Stock market prediction methods are divided into two



main categories: technical and fundamental analysis. Technical analysis focuses on analyzing historical stock prices to predict future stock values (i.e. it focuses on the direction of prices). On the other hand, fundamental analysis relies mostly on analyzing unstructured textual information like financial news and earning reports. Many researchers believe that technical analysis approaches can predict the stock market movement [3–5]. In general, these researches did not get high prediction results as they depend heavily on structured data neglecting an important source of information that is the online financial news and social media sentiments. These days more and more critical information about the stock market has become available on the Web. Examples include BBC, Bloomberg, and Yahoo Finance. It is hard to manually extract useful information out of these resources. This draws a picture of the significance of text mining techniques to automatically extract meaningful information for analyzing the stock market. In this research, the most crucial past literature was reviewed, and a major contribution was made to the subject of using text mining and NLP for market prediction.

We revealed the finding of the selected studies to show the significantly improved performance of stock market forecasting via many machine learning methods. This study also clarifies the recent innovation researches and its potential future contribution. Comparisons and analyses of different researches are made on the financial domain of market prediction that can help to establish potential opportunities for future work. In this research, we also focused on the promising results accomplished by machine learning methods for analyzing the stock market using text mining and natural language processing (NLP) techniques.

In contrast to the other current survey articles that concentrate on summarizing many methods used for forecasting the stock market, we aim to compare many machine learning (ML) and deep learning (DL) methods used for sentiment analysis task of social media and financial news articles to find which method could be more effective in prediction. **Figure 1** represents the reviewed study framework. The rest of this work is organized as follows. Section 2 provides a review of background concepts that are needed to be known before the detailed analysis of the literature. Section 3 illustrates the relationship between stock market prediction and text mining. Section 4 includes a review of the machine learning main methods used for stock market prediction based on textual resources. Section 5 explained the least frequently used algorithm for stock prediction based on text mining. Section 6 describes the reviewed work text sources and period and number of collected items. Section 7 contains the reviewed works finding, limitation, the measurement used, and future work. Finally, Section 8 concludes this paper.



**Figure 1.**  
*The reviewed study framework.*

## 2. A review of background concepts

Our work defined the following concepts as important to understand this research topic.

### 2.1 Sentiment analysis

Sentiment analysis uses text mining, natural language processing, and computational techniques to automatically extract sentiments from a text [6]. It aims to classify the polarity of a given text at the sentence level or class level, whether it reflects a positive, negative, or neutral view [7]. In stock market prediction task, two important sources of the text are used either social media mainly using Twitter data or online financial news article.

#### 2.1.1 Twitter sentiment

Twitter is a significant source of data, and many researchers have examined its relationship with stock market movements [8]. While each tweet is restricted to 140 characters, it is believed that the information can accurately reflect public mood [9].

#### 2.1.2 Online financial news sentiment

Financial news articles are perceived to be a more consistent and reliable source of information. Many researchers suggested that the financial news articles have a strong relationship with stock market fluctuation; therefore, analyzing financial news reports can help in predicting the stock market movements [10]. In [11], the author used a unified latent space model to examine the relationship between stock prices and news article releases. The result indicates a good return accuracy, which proves that news article analysis has an important impact on stock market movement.

### 2.2 Textual data preprocessing

Textual data need to be prepared before used by the machine learning algorithm for sentiment analysis task using these methods.

#### 2.2.1 Feature extraction

Feature extraction or sometimes called attribute selection aim to select features, attributes, or piece of text that is more relevant to the prediction task. Many methods have been used for feature selection. The commonly used feature selection procedure for document or sentence classification task is the bag-of-words (BOW) approach, which was recently used for market prediction by many authors [12–14]. In the mentioned model, each word in a text or document will be treated as a feature neglecting the grammar or word order and only preserving the abundance. The second most popular method used recently for the feature selection process is Word2vec [12]. In this technique, the aim is to learn word embedding using a two-layer neural network. The input to that neural network is a text, and the output is a group of vectors (i.e. the input is a corpus and the output is a vector of words).

Another important feature selection method is the latent Dirichlet allocation (LDA) technique used recently for market prediction in [13]. In the LDA model, the text is viewed as probabilistic collections of terms or words, and the collections are then treated as selected features. Other researches [12, 14] used a Skip-Gram model

that aims to predict the context word (surrounding words) for a given target word. However, feature selection is a crucial step in the textual data preprocessing, and many other strategies may also be used for text analysis.

### *2.2.2 Feature representation*

After feature selection, every feature must be illustrated by a numeric value so that it can be analyzed by machine learning techniques. The most common technique of feature representation is a binary representation (BR), which is a number system that uses two values such as 0 and 1 exclusively to represent the information. This technique has been exploited for market prediction researches by many authors [15–17]. The second most popular method used in text mining for financial application is the term frequency-inverse document frequency (TF-IDF), which is a numeric value that represents the significance of a word for a document or corpus that is used recently by many authors [12, 18]. Other feature representation methods can also be used successfully in text preprocessing, and we will discuss those with more details in the following sections.

## **3. The relationship between stock market prediction and text mining**

Many papers study the relationship between stock price movements and the market sentiments, and the most relevant studies will be discussed in this section.

Ref. [19] examined the ability to use sentiment polarity (positive and negative) and sentiment emotions selected from financial news or tweets to predict the market movements. For sentiment analysis, they have collected a large dataset of the top 25 historical financial news headlines in addition to a large set of financial tweets collected from Twitter. Furthermore, they collected stock historical price data for many S&P 500 companies and used the close price as an indicator of the stock movements. For evaluation, they used the Granger causality test [20] that is a statistical test technique commonly used to reveal causality in time series data and explore if one-time series data can predict the other. For sentiment analysis, the authors examined two machine learning methods SVM and LSTM. The experiment result illustrated that in some cases sentiment emotions contribute to Granger-cause stock price fluctuation, but the finding was not inclusive and must be examined for each case. Also, it has been revealed that for some stocks, adding sentiment emotions to the machine learning market prediction model will increase the prediction accuracy. Comparing the two machine learning methods, SVM achieved better and more balanced results, and that's because the size of the dataset is quite small to be sufficiently used with SVM.

Another paper [21] examined the efficiency of using sentiment analysis of microblogging sites to forecast the stock price returns, volatility, and trading volume. The extracted intraday data from the two sources of information, Twitter and StockTwits, were collected for 2 years. For the evaluation, the authors used five famous stocks, namely, Amazon, Apple, Goldman Sachs, Google, and IBM. Prices were represented every 2 min, and the sentiment data were collected for the same period span of each trading day. To find the links between stock price outcomes and tweet sentiment, they applied Granger causality analysis. The experiments indicate that there is a causal link between Twitter sentiments and stock market returns, volatility, and volume. Among all five stocks, market volatility and volume seem to be more predictable than market direction or return.

In [22], the author exploited a multiplex network approach to study the correlation between market movements and social media sentiments. The proposed

model merges information from two sources of data: Twitter posts and market price data. The authors selected 100 of the biggest capitalized companies of the S&P 500 index for a 5-year period from May 2012 to August 2017. In their model, they suggested that financial network correlation was established by the integration of the two techniques. The first one suggests that two stocks tend to be associated if they share joint neighbors. The other techniques suggest that two connected stocks usually remain connected in the future. The findings demonstrated that a multiplex network approach incorporating information from both social media and financial data can be used to forecast a causal relationship framework with high accuracy.

The authors in [23] investigated the ability of economic news to predict Taiwan stock market returns. The proposed model used text mining techniques throughout many steps. Firstly, they converted the textual news into numerical values. Secondly, they append the resulting numerical variable to regression models with macroeconomic attributes to examine the role of news articles in predicting stock price returns. The model also defines specific keywords and calculates the number of positive, negative, and neutral words in each news text and then converts them into three news attributes, which are then fed to the regression model. The experiments find that adding news articles was able to reduce the root mean square error (RMSE) that proves that the economic news has crucial impacts on market returns. The experiments also indicate that negative news has more influence on the stock market returns than positive news articles.

The study proposed in [24] aims to analyze whether tweet messages could be used to predict future trends of stocks for particular companies listed on the Dow Jones stock market, focusing on 12 companies related to 3 distinct and crucial economic branches in technology, services, and health care. The authors gathered the company's market data and Twitter posts for a 70-day period for analysis. The companies of each category were chosen based on the volume of messages that mention the company names on the StockTwits website. The study illustrates that some of the proposed ad hoc forecasting models well predict the next day direction of the stock movements for some companies with 82% of success and there is no unified method to be used with all cases. The results also indicate that more volume of a tweet will yield better prediction results. Moreover, the study proved the robust correlation between tweet's posts and the trend movements for some companies.

Overall, past studies indicate that there is a strong relationship between market movements and information published in news and social media. The information on social media contributes to enhancing the prediction models with all of the discussed papers. The evaluation of event sentiment may affect the market returns further and boost the outcome of forecasting.

## **4. Machine learning for market prediction**

Recently, many research studies used machine learning via text mining innovation methods to successfully predict the stock market changes, and the most significant ones are going to be discussed in this section.

### **4.1 Support vector machines**

Support vector machines (SVMs) are a supervised machine learning model used extensively in classification and regression tasks. SVM is a hyperplane that divides a collection of documents into two or more classes with a maximum margin [25].

SVM was first applied to the text classification task by Joachims [26]. In his approach, the author used a limited vocabulary as the feature collection by using



a list of the most occurred words and discard of uncommon words from the feature set. Utilizing 12,902 documents from the Reuters-21578 document group and 20,000 medical summaries, the author compared the effectiveness of many machine learning techniques such as SVM and Naive Bayes (NB). For both document groups, the experiments demonstrated that the SVM achieve better classification result compared to NB classifier.

For stock market prediction, many research papers used the SVM for text classification and sentiment analysis. Combining both textual information and historical stock prices for stock market prediction [27] research applied the SVM to forecast the Chinese stock direction and stock prices between the years 2008 and 2015. For text mining, the authors formed a stop word and sentiment dictionary based on a specific domain. In the study, there were two kinds of input. The first one includes 2,302,692 news items, whereas the other contains only stock data of the largest 20 Chinese stocks based on trading volume. Support vector regression (SVR) is used to predict stock price, and support vector classification (SVC) is exploited to predict stock direction. The result indicates that both audience numbers and news quality have a crucial impact on the stock market. Moreover, for SVC, the direction accuracy was 59.1734%, which illustrates better progress than other works. The result also indicates that news articles have an important effect on the stock market fluctuations.

Another research [28] introduced a stock market prediction framework. For sentiment analysis, the researchers used two financial sentiment dictionary, namely, the Harvard IV-4 sentiment dictionary (HVD) and Loughran and McDonald (LMD) [29] financial dictionary. The dataset consists of 5 years of historical Hong Kong Stock Exchange prices and financial news collected from January 2003 to March 2008. For text classification SVM was used for training. Experiments indicate that the techniques with sentiment analysis perform better than a bag-of-words model in accuracy measures. It also revealed the small difference between the two models LMD and HVD. For LMD the accuracy was 0.5527, whereas HVD accuracy was 0.5460, which indicates that the two dictionaries can be used effectively for the market prediction task.

Another paper [30] developed a model to predict three stock price directions with 1-day, 2-day and 3-day lag. The dataset contains financial news of SZ002424 stock from September of 2012 to March of 2017. In order to analyze the structure of news and get the hiding information inside the contents, the authors proposed a semantic and structural kernel (S&S kernel). The kernel was based on SVM and uses medical industry news for evaluation. Experiments find that the proposed kernel can reach up to 73% accuracy when predicting the price trend with 2-day lag, which proves that content structure hidden in daily financial news can predict the stock market movements. The result also reveals that financial news has an important influence on stock movements that typically last for 2–3 days.

In the work of [31], the authors used a lexicon-based approach to predict the stock market based on Twitter user feelings. The authors used historical stock data in addition to Twitter messages to predict DJIA and S&P 500 indices movements. Twitter data were obtained to train support vector machine and neural networks (NN) for 7 days. The dataset was created by adding a normalized set of tweets that contains 8 categories of emotions in about 755 million tweets. The collected tweets were downloaded from the period of February 13, 2013, to September 29, 2013. For sentiment analysis, a dictionary approach has been created manually by an expert in the field. The best average accuracy was obtained by using the SVM algorithm to forecast the DJIA indicator with an accuracy equal to 64.10%. However, using NN to predict S&P 500 achieves only a 62.03% in accuracy measure, which proves that

SVM performs better than the NN algorithm for market prediction. Moreover, the results achieved by the model indicate that it is possible to increase the prediction accuracy using human sentiment analysis and a lexicon-based approach.

In the paper of [15], the authors proposed a model with the user interface to predict the market movement for 1 day ahead. The proposed model consists of historical stock prices, technical indicators, Wikipedia company pages, and Google news. The model employs three machine learning methods to compare and select from, namely, ANN, SVM and decision tree (DT). The model concentrates on forecasting the AAPL (Apple NASDAQ) stock movement for a period from May 1, 2012, to June 1, 2015. For the APPL prediction case study, the authors used SVM recursive feature elimination (RFE) to choose the most important features. RFE is applied via backward choosing of predictors relying on feature importance ranking. Combining many data sources, the financial expert system achieves 85% accuracy in prediction. The result indicates that incorporating data from multiple sources will improve the efficiency of market prediction.

In [32], the author introduced a method to predict the stock movement for 1 day ahead. The proposed technique used a manually labeled corpus. The dataset contains 16 randomly selected stocks that are commonly discussed by StockTwits users collected from the period of March 13, 2012, to May 25, 2012. The collected tweets were about 100,000 posts. For text analysis, the model used SVM to analyze sentiment in StockTwits. The results prove the outstanding performance of SVM for sentiment classification tasks with accuracy that can reach up to 74.3%, whereas the overall accuracy for predicting the market up and down change based on the suggested model was 58.9%.

From the findings recorded in **Table 1**, it can be noted that SVM efficiency surpasses the effectiveness of approaches that used neural network models as we discussed earlier.

## 4.2 Deep learning

A deep learning concept is derived from machine learning methods that utilize many layers of data processing for the extraction of features, patterns, and classification. Recently, deep learning techniques are launched to sentiment analysis tasks, and they are considered effective in most cases [33].

In [34] the authors investigated whether deep learning methods can be modified to improve the accuracy of StockTwits sentiment analysis. Several neural network variants such as LSTM, doc2vec, and CNN were examined to discover stock market sentiments posted on StockTwits. The results prove that the convolutional neural network is one of the best deep learning methods for predicting authors' sentiment in the StockTwits dataset. Many other types of research discussed the successful use of deep learning for sentiment analysis and natural language processing tasks. On the survey research in [35], some of the different methods used in sentiment analysis tasks are compared. The main result showed the excellent performance of deep learning methods for sentiment analysis, in particular, CNN and LSTM methods.

Another paper [36] proposed a method to predict the French stock market based on sentiment and subjectivity analysis of Twitter data. The author applied a simple feedforward neural network to analyze tweets and predict CAC40 index movements for the next day. The Twitter collected data for the period of February 27, 2013, to June 16, 2013, was about 25,930 tweets. In addition to Twitter data, Martin also used historical stock market prices for the CAC40 index and other stocks. The results yield a direction accuracy of 80%, which indicates that using a neural network can be used successfully to predict the stock market movements.

Reference	Data type	Methods	Feature selection of textual data	Feature representation	Measure used	Results
Porshnev et al. [31]	Twitter, historical stock data of DJIA and S&P 500	SVM, NN, and sentiment dictionary	Emotion lexicon	Sentiment score of 8 scales	Accuracy	SVM ACC = 64.10%. NN ACC = 62.03%
Li et al. [28]	Five years historical Hong Kong Stock Exchange prices and financial news	SVM and two financial sentiment dictionary HVD and LMD	Polarity asymmetry of the news	Sentiment score	Accuracy	LMD ACC = 0.5527, HVD ACC = 0.5460
Xu and Keelj [32]	StockTwits	Manually labeled corpus and SVM	Unigram, bigram, line length, and punctuation	Sentiment score	Accuracy	ACC 58.9%
Weng et al. [15]	Historical stock prices, technical indicator, Wikipedia company pages, and Google news Apple NASDAQ stock	ANN, SVM, and DT	RFE	Binary	Accuracy for including all data sources	Approximately 85%
Xie and Jiang [27]	Financial news, 20 Chinese stock prices	SVM and specific sentiment dictionary	BOW	Sentiment score from -5 to +5, where -5 represents the most negative impact, +5 represents the most positive impact, and 0 represents for stop word	Accuracy	ACC 59.1734%
Long et al. [30]	Financial news of SZ002424 stock	SVM and S&S kernel	BOW	Keyword frequency	Accuracy	ACC = 73% with 2-day lag

**Table 1.**

*Support vector machine for stock market prediction based on text mining studies.*

#### *4.2.1 Artificial neural networks*

Artificial neural networks are a subset of deep learning technology that falls within the large artificial intelligence domain, and it mimics the human brain and its nervous system work. The simplest form of artificial neural networks is a feed-forward neural network where the data go through the different input nodes until they reach the output node using only one direction, which is obtained by using a categorizing activation function.

In [37], the authors proposed a market investment recommendation system to predict intraday stock returns. The authors tested many prediction methods to find the best resulting algorithm. The dataset includes 72 S&P 500 companies for evaluation. Using both historical market data with financial news, the authors implemented the modeling technique many times to select the best model. For the first time, they have applied a feedforward neural network algorithm. For the second time, they used a stepwise logistic regression (SLR). For the third time, they implemented the decision trees with a genetic algorithm (GA) proposed by [38]. The best result was obtained by using the neural network prediction technique, which indicates that the NN algorithm is profitable for any initial investment. The result also confirms that combining market data with financial news can predict the market movement with better accuracy.

In [39] the producers predicted the stock market movements based on sentiment analysis of comments and tweets extracted from Twitter and StockTwits famous social media sites. User comments are classified into four different categories, which are up, down, happy, and rejected. The market data of the popular companies like Apple, Microsoft, Oracle, Google, and Facebook was collected from the period of January 1, 2015, to February 22, 2016. Both market data and polarity data were fed to an artificial neural network to predict the movements of the stock. The best prediction result was obtained for Apple Company with MSE equal to 0.14.

In [40], the proposal adopted a two-layer RNN-GRU technique to forecast the Chinese stock market movements. The model exploited sentiment analysis of Sina Weibo (a very popular Chinese social network) news and posts. The authors constructed their sentiment dictionary using user posts on the website. The authors also collected stock prices of the Shanghai Shenzhen 300 Stock Index (HS300) to use as an input to the recurrent neural network (RNN) model with gated recurrent units (GRU). The experiments revealed that the news and posts on Sina Weibo can predict the market movements with MAE equal 0.625 and with MAPE equal to 9.38.

In [13] the authors proposed a multi-source multiple instance (M-MI) model to predict the stock market index movements. In the proposed frameworks, the authors collected data from multiple resources, namely, quantitative data of Shanghai Composite Index historical prices for each trading day, financial news data to extract events, and social media data taken from Xueqiu (a famous trader social network in China to explore user sentiments user posts). Then, the analyzed sentiments, events, and the stock historical data are given as input to the M-MI model to make the prediction. For event extraction, the authors used HanLP (the popular method used for text parsing to grab the syntax of a sentence). Event extracted is used to feed the Restricted Boltzmann Machines (RBMs), which is a creative theoretical artificial neural network. In the model, the authors also examined the importance of specific sources to the index movements by giving them specific weights. The proposed framework prediction accuracy was about 60%, which reveals many findings. Firstly, the integration of features from multiple resources can make a more effective prediction. Secondly, both news events and



market historical data have a more important effect on stock movements than social media sentiments. Thirdly, both news events and quantitative data have larger impacts on stock fluctuations than using sentiments alone.

Recently [41] applied a technique to forecast the stock directions. The authors used sentiment analysis of news headlines in addition to historical market data of Apple stock to predict the market trend. Hive ecosystem was used to preprocess the data, and the naive Bayes classifier was utilized to calculate the sentiment scores. With two inputs from news headlines sentiment score and historical numeric market data, the multilevel perception artificial neural network (ANN) is applied to forecast the stock movements. In the training procedure, the authors used back-propagation, and in the output layer, they used the identity function. Moreover, the model tested two different periods for training the data; in the first method, they trained a 3-year data period, and the second method trained a year data period. The result represents an accuracy of 91% in the first methods, while 98% accuracy was achieved in the second method, which indicates that stock price forecasting is more efficient for a shorter time.

More recently [42] predicted future market trends by using both market historical prices and financial news article sentiments as input to the neural network. The authors collected historical prices of the 20 biggest companies listed in the NASDAQ100 index to predict the fluctuations of the stock for the portfolio that consists of 20 firms historical stock prices, with a periodicity of 15 min, obtained from Google Finance API. For new article analysis, two approaches of feature selection have adopted the dictionary of Loughran and McDonald (2011) (L&Mc) and affective space [43]. The Loughran and McDonald dictionary is commonly used for market prediction and consists of many critical words for the classification task that represents negative, positive, and uncertain sentiments that can be found commonly in financial news, whereas affective space (AS) dictionary is a vector space dictionary that depends on the similarity and relationships between words as natural language processing methods. For dimensionality reduction, the affective space mapped each term to a 100-dimensional vector that allows concepts to be grouped based on their semantics and relations.

The proposed model with Loughran and McDonald's dictionary confirms to be more effective, resulting in an annualized return of 85.2%, while the use of affective space feature dictionary as an input to the neural network model proved to be more effective in obtaining high accuracy results. **Table 2** summarizes the studies that used NN extensively for market prediction techniques.

#### *4.2.2 Recurrent neural network*

Recurrent neural network is an important variant of artificial neural network that starts as normal with front direction but preserves the relevant data that may need to be utilized later. In other words, every node will act as a memory cell that remembers some information it had in the earlier step.

A well-known variant of RNN model is long short-term memory (LSTM), which was proposed by Hochreiter and Schmidhuber in 1997 [44]; it is a standard recurring neural network that solves the exploding gradient problem. LSTM can depict the long dependencies in a sequence by adopting a memory unit and a gate mechanism to determine how information stored in the memory cell can be used and updated [45]. Each LSTM is a set of cells or system modules that catch and store streams of data. The cells represent a transport line that carries data from the past and collects them for the present module from one module to another. Through the use of certain gates in each cell, data can be disposed of, filtered, or added for the next cells [46].

Reference	Data type	Methods	Feature selection of textual data	Feature representation	Measure used	Results
Martin [36]	CAC40 index data, Twitter	NN	Tokens	Average sentiment score	Direction accuracy	80%.
Geva and Zahavi [37]	72 companies in the S&P 500 index data, financial news	NN, SLR, and DT with GA	BOW	Calibrated sentiment scores and binary indicator	Return over the initial investment 200 k	NN: 8.57% SLR: -0.20% GA: 0.16%
Khatri and Srivastava [39]	Twitter and StockTwits, index data of Apple (APPL), Microsoft (MSFT), Oracle (ORCL), Google (GOOG), and Facebook (FB)	ANN	Predefined words	Sentiment score between 0 and 1	MSE	AAPL: 0.14 MSFT: 0.18 ORCL: 0.22 FB: 0.28 GOOG: 0.27
Zhang et al. [13]	Shanghai Composite Index, financial news data, and social media from Xueqiu	(RBMs) and ANN	For event extraction: HanLP, Sentence2vec For sentiments: latent Dirichlet allocation	Two polarities: positive or negative	Prediction accuracy	60%
Zhang et al. [13]	HS300 Index and Sina Weibo news and posts	Own sentiment dictionary and two-layer RNN-GRU	Positive and negative keywords	Probability value for fall or rise	MAE, MAPE, and RMSE	0.625, 9.381, and 0.803
Picasso et al. [42]	20 companies in NASDAQ-100 index and financial news articles	NN with L&MC and NN with affective space	L&Mc dictionary and AS dictionary	Counts of negative, positive, uncertainty, superfluous, and other words of the dictionary found in news and number of news in the slot	Accuracy	NN AS 68% NN L&MC 60%
Shastri et al. [41]	Apple stock and news headlines	Hive ecosystem, NB, and multilevel perception artificial neural network (ANN)	Unique positive and negative words	Sentiment score	MAPE, trend prediction accuracy of 1-year period	8.21 98%

**Table 2.**  
*The main study that used NN extensively for market prediction based on text mining.*

In the paper of [47], the proposal adopted a method to predict the stock market movements based on the bidirectional gated recurrent unit (BGRU), which is considered a variant of LSTM. The model used financial news that comes from Reuters and Bloomberg websites and historical stock prices to predict the market fluctuation with a better result. The S&P stock prices and news data were collected in the period of 2006–2013. Also, the model examined the method performance on the individual stock that comes from different sectors, namely, Google Inc., Walmart, and Boeing. In the proposed method, the authors used the word embedding model introduced by [48] to select the most efficient features from the collected financial news. In word embedding model, the words were encoded as vectors in a high-dimensional space, and then the analogy between words in meaning is interpreted to closeness in the vector space. The proposed model achieved accuracy equal to 59.98% in the S&P 500, whereas individual stock prediction accuracy was more than 65%. The authors also examined the performance of many LSTM variants like standard LSTM, GRU, and BGRU. The finding shows that BGRU obtained the best results compared to other LSTM variants.

However, conventional LSTM is unable to detect what is the most crucial part of the sentence for the sentiment categorization task. Therefore, [49] proposed a design mechanism capable of detecting the crucial part of the sentence related to a specific aspect and explained the architecture of attention-based LSTM in detail.

To predict the stock market directional movements, [50] proposed an Attention-based LSTM model (AT-LSTM) to predict the movements of Standard & Poor's 500 index and individual companies' stock price using financial news titles. The attention techniques were divided into two classes. The first class of attention assigns there weight to the news that contains positive sentiments to the stock market such as "raise," "growth," etc. While the second class of attention assigns there weight to the news that mentions the major companies in the S&P 500 such as "Microsoft" and "Google." Therefore, the attention model is trained continuously to assign more attention to the relevant news based on its content. The proposed method achieved more than 66% accuracy, and the company WALMART obtained a max accuracy of 72.06%. The results prove that attention mechanisms can achieve good results for market prediction in specific cases.

In [51] proposal support decision system based on deep neural networks and transfer learning was applied. To enhance the prediction accuracy, the authors pretrain the networks on a different corpus. The main aim of the study was to recommend the best deep learning techniques in terms of market prediction. The system provides its corpus with a length of 139.1 million words. The authors trained the deep neural networks by using the Adaptive Moment Estimation Algorithm (Adam), which can effectively solve sparse gradient problems. Then the use of transfer learning aims to initialize the weights of parameters with values that might be close to the optimized ones. In order to account for unbalanced classes in their dataset, they have used classification balanced accuracy that can be defined as the arithmetic mean of sensitivity and specificity. They also predicted the direction of nominal returns. The result proves that LSTM models surpass all traditional machine learning models based on the bag-of-words technique, specifically when they used transfer learning to pretrain word embeddings.

Recently [12] examined the effect of financial news articles on stock trend fluctuation either rise or fall. The financial new articles related to the Taiwan 50 Index were collected from Google. For textual data analysis and NLP tasks, the authors used their lexicon and then exploited the LSTM to make the final prediction. The use of LSTM features was joint with historical data and adjusted in each step. The results prove that individual stock prediction using the study polarity lexicon was better than the benchmark model. Moreover, the proposed model reaches an

accuracy of 76.32, 80.00, and 77.42% for each of the following stocks TSMC, Hon Hai, and Formosa Petrochemical, respectively, which reveals the effectiveness of the LSTM model in market prediction based on text analysis.

Another study proposed in [52] examined the effectiveness of using the LSTM technique to predict market movements, using market data and textual resources as input to the model. The authors analyzed user sentiments from forum texts about the CSI300 index using the naive Bayes algorithm and then using LSTM, which contains a merged layer, a ReLU layer, and a softmax layer to combine the investor sentiment taken from forum posts with the historical market. The fall or rise trend prediction accuracy achieved was 87.86%, outperforming other commonly used machine learning methods such as SVM algorithm by at least 6%, which highly indicates that LSTM can achieve a better result in prediction when using larger datasets. **Table 3** summarizes the recent studies that used RNN networks for stock market prediction based on text analysis.

4.2.3 Convolutional neural network (CNN)

Convolutional neural network used for natural language processing was first explained by Collobert and Weston in [53]. A typical convolutional neural network is composed of multiple convolutional layers at the bottom of a classifier. Conventional inputs for text processing are characters, phrases, paragraphs, or documents that are converted into a matrix representation. Each row of the matrix represents a token, which is typically a word or character [54].

In [16], framework proposal for stock market prediction based on long-term events and short-term events extracted from financial news articles about the S&P 500 index was applied. The collected financial news articles come from October

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Li et al. [52]	CSI300 index data, investors forum posts	LSTM, NB	Manually labeled sentiments by experts	Pos, neg, and neutral	Direction accuracy	87.86%
Huynh et al. [47]	S&P 500 index, financial news	BGRU	Word embedding	Real valued vectors	Prediction accuracy	59.98%
Kraus and Feuerriegel [51]	German ad hoc announcements	Transfer learning with RNN and LSTM	Word embedding (em)	Polarity score	Direction accuracy of nominal return	RNN 0.552 LSTM 0.576 LSTM-em 0.578
Liu [50]	S&P 500 index, financial news titles	AT-LSTM	Own word embedding trained with Skip-Gram	Word embedding and character-composition vector	Direction accuracy	More than 66% for each stock
Chen et al. [12]	Financial news articles, Taiwan 50 Index	LSTM and polarity lexicon	Word2vec and Skip-Gram	TF-IDF and polarity score	Accuracy	Up to 80% for Hon Hai stock

**Table 3.**  
*Recent studies that concentrate on RNN variants for market prediction based on text analysis.*



2006 to November 2013, which was released initially by Ding et al. [55]. The long-term events represent events over the past month, while the short-term events represent events on the last day of the stock price fluctuate. The proposed frameworks train the extracted events using a neural tensor network and then a convolutional neural network to predict both the short-term and the long-term impact of extracted events on stock price fluctuations. The proposed framework examined two different ways for representing the input to CNN. The first method (WB-CNN) used word embedding as input and convolutional neural networks for prediction. The second method (EB-CNN) used event embedding as input and convolutional neural networks for prediction. The experiments achieve accuracy of 61.73% for WB-CNN, while the EB-CNN method achieved an accuracy equal to 65.08%, which illustrates that the proposed model is more effective in stock market prediction than other models that predicted the S&P 500 index based only on stock historical data analysis. The model also proves that CNN can extract the longer-term influence of financial news events than traditional feedforward neural networks.

In [17], writers proposed a model to predict the intraday stock market directional movements of the S&P index using financial news title and financial time series market data as input. The paper compared two commonly used deep learning methods, which are RNN and CNN algorithms using many text representation methods. The RNN method used in the paper was the LSTM model. The proposed model examined many types of text representation as an input to the CNN prediction model. The (W-CNN) represents a word embedding as input and a CNN as a forecast model. The (S-CNN) represents sentence embedding input and CNN forecast model. The (W-RCNN) word embedding input and RCNN forecast model. The (S-RCNN) represents sentence embedding input and RCNN forecast model. The (WI-RCNN) shows word embedding and historical time series input and RCNN prediction model. The (SI-RCNN) illustrates sentence embedding and historical time series data input and RCNN prediction model. Experiments on each of the previous models revealed that CNN is more effective than RNN on capturing

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Ding et al. [16]	Financial news articles	The train uses a neural tensor network, WB-CNN, EB-CNN	Word embedding WB, event embedding EB	Binary	Accuracy	WB-CNN 61.73 EB-CNN 65.08%
Vargas et al. [17]	Financial News articles, S&P index market data	Many text representation	Word embedding, sentence embedding	Binary	Accuracy	W-CNN 57.22% S-CNN 60.96% W-RCNN 60.22% S-RCNN 61.49% WI-RCNN 61.29% SI-RCNN 62.03%

**Table 4.**  
*CNN use for stock market prediction based on text mining results.*

semantic from new financial, and RNN is more efficient in capturing the context information for the stock market prediction. Moreover, the results prove that the sentence embedding for text representation is more effective than the word embedding. **Table 4** summarizes the studies that used CNN for stock market prediction based on sentiment analysis and NLP.

## 5. Other machine learning methods

Many other machine learning methods were used successfully and less frequently for market predation applications based on text mining. Summaries of these studies are illustrated in **Table 5**. In the study of [18], a method was proposed to predict the stock trend movements of three NASDAQ companies, namely, Yahoo Inc., Microsoft Company, and Facebook Inc. (FB Inc). The model used financial news sentiment analysis with historical stock data to predict the market with higher accuracy. The task is accomplished with two steps: Firstly, they used naive Bayes classifier to classify news sentiment into two classes, positive or negative. Secondly, to forecast the stock trend fall or raise, they used k-Nearest Neighbor algorithm (K-NN) (a clear algorithm that saves all possible instances of data and categorizes the new data based on a scale of closeness and is often used to classify a new data based on the current classification of its neighbors). The results show that the accuracies of sentiment analysis of news only can go up to 63%, while combining news sentiments with historical stock prices can achieve trend prediction accuracy up to 89.80%, which proves that adding historical stock prices to the classification model will be able to improve the prediction performance.

In the work of [56], the authors suggest a method to predict the daily up and down price fluctuation of four tech companies of NASDAQ stock, which are Apple (AAPL), Google (GOOG), Microsoft (MSFT), and Amazon (AMZN). The model analyze Twitter user messages in addition to three previous days of the stock price movement. The model constructs a named-entity recognition (NER) approach to identify and remove the noise of Twitter data. A decision tree approach was used to build the classification model. The proposed model achieved the highest accuracy of 82.93% in predicting the daily up and down changes of Apple Company, which indicates that using named-entity recognition method for noise removal of Twitter data can improve the accuracy results.

The research in [8] proposed a method to predict the stock market movements based on two feature extraction methods, using a novel aspect-based sentiment model to improve the prediction performance. The first methods tempt to excerpt hidden topics and sentiments together and use them for the prediction, while the aspect-based sentiment methods treat every message as a list of topics and correlative sentiment values. To build the prediction model, the authors used SVM with the linear kernel and collected data of 18 stocks for a period of 1 year from July 2012 to July 2013. Exploiting the aspect-based sentiment feature method obtained the best result with 54.41% average accuracy. The proposed model also proves to be 3.03% more effective than using the human sentiment method for stock movement prediction.

In [61] proposal a method to forecast the Indonesian stock movements based on Twitter sentiment analysis was introduced. Naive Bayes and random forest algorithm was used to find the user sentiments of the 13 most popular companies in Indonesia. The linear regression technique was used to build the prediction model. The highest accuracy was achieved by the categorization model using the random forest algorithm with 60.39% accuracy, whereas naive Bayes classifier was able to classify tweet data with 56.50% accuracy. For the price movement's prediction, the

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Vu et al. [56]	Twitter user messages, AAPL, GOOG, MSFT, and AMZN indices data	Decision tree approach, NER	Predefined bullish-bearish anchor words	The real number for daily Neg_Pos and Bullish_Bearish	Daily prediction accuracy	AAPL 82.93% GOOG 80.49% MSFT 75.61% AMZN 75.00%
Moniz and de Jong [57]	News stories for 598 global companies	Ensemble tree, LDA	LDA	Binary	F1- measure	0.508
Bing et al. [58]	30 NASDAQ and New York stock indices and Twitter	Association rule	Sentiment word list	TF-IDF, vector space model, which is an arithmetic model to represent text as vectors	Average accuracy	76.12%
Li et al. [59]	HSI 23 stocks indices and financial news	Multiple kernel learning	Word list extreme positive, positive, neutral, negative, and extreme negative	TF-IDF, vector space model	RMSE	0.139 for 30 m
Shynkevich et al. [60]	Five stock from the S&P 500 index, SS, and SIS news items	Multiple kernel learning	BOW	TF-IDF	Highest accuracy	81.63% for WLP stock with six kernels
Nguyen et al. [8]	Social media message board and 18 stocks index data	SVM	POS tagging Stanford CoreNLP for aspect-based sentiment	Average sentiment score or values	Prediction Accuracy	Aspect-based model 54.41%
Cakra and Trisedya [61]	Twitter data and many companies in Indonesia indices data	NB and RF and linear regression	Sentiment lexicon and sentiment shifters	Positive, negative, and neutral	Prediction accuracy	NB 67.37% RF 66.34%

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Ghanavati et al. [62]	Hong Kong market index and financial news articles and summaries	Loughran and McDonald dictionary, ML, and metric learning-based methods	Tokenization using OpenNLP tools	Sentiment value vectors	The average error rate	The average error rate of ML for large cape stock 0.15 The average error rate of ML for small cape stock 0.20
Khedr and Yaseen [18]	Financial news, index data of 3 NASDAQ companies	K-NN and NB	TF-IDF and N-gram	Values for pos, neg, and equal	Trend prediction accuracy	89.80%
Gálvez and Gravano [63]	Twelve stocks of the Merval Index and online message boards	Combining LSA with ridge regression	Latent semantic analysis (LSA)	Numbers for each special token	Maximum accuracy when using technical indicators and topics from the online message board	Up to 0.750
Liu and Wang [15]	China Security Index 300 (CSI300) and the Standard & Poor's 500 (S&P500). News reports and numerical data	LSTM and many textual representations	News embedding	Numerical Vectors	Accuracy	NBAD raises the accuracy of 2.32% and 1.35% higher than the best baseline models of the dataset
Maqsood et al. [64]	Many USA, Hong Kong, Turkey, and Pakistan company indices. Twitter	Event sentiment, linear regression (LG), support vector regression (SVR), and deep learning	A comprehensive dictionary with their own generated word list	Sentiment value that is calculated for each day separately	Average root mean square error (RMSE)	For each country using LG, SVR, and DL, respectively. US 4.35, 1.33, and 1.65 Hong Kong 0.90, 0.31, and 0.35 Turkey 0.27, 0.11, and 0.11 Pakistan 0.70, 0.34, and 0.33

**Table 5.**  
*Summaries of machine learning methods that were used successfully and less frequently for market prediction based on text mining.*



proposed models can predict the upcoming price fluctuation of either rise or fall with the accuracy of 67.37% achieved by the naive Bayes algorithm and 66.34% obtained by using Random Forest classifier.

Other research [62] introduced a stock market prediction service framework that allows users to choose different data sources and machine learning techniques. The authors gathered all news summaries and historical prices of all the stocks for a 1-year period. Using the Hong Kong market stock dataset for evaluation, they found that metric learning-based methods can improve the prediction results. The study also shows that adding news to the historical prices for stock market prediction will be more useful on large and popular stocks.

Recently [14] applied a numerical-based attention (NBA) method for multiple sources of stock market prediction. News headlines and numerical data combined to predict the stock prices. For evaluation, the authors collected news headlines and numerical data from two sources: the China Security Index 300 (CSI300) and the Standard & Poor's 500 (S&P500). They used NBAa-NBA<sub>d</sub> to denote different variations of the models with different textual representations. In these three datasets, the proposed structure accomplishes the best outcomes. Especially, NBA<sub>d</sub> raises the accuracy of 2.32 and 1.35% higher than the best baseline models on S&P500 and CSI300.

More recently, [64] investigated the effect of the most important event from 2012 to 2016 into the stock exchange prediction of four selected countries, which are the USA, Hong Kong, Turkey, and Pakistan. The events are then categorized into local and global events for each country according to their economic effects on the country stocks. Twitter data were gathered to find the sentiment for each one of these events. The model used a total of eight events for all countries. For classification, the authors investigated linear regression, support vector regression, and deep learning model for market prediction. The results revealed that linear regression achieves the worst prediction results compared to the other two methods used in their analysis, while the support vector regression achieves the best results. Event sentiment illustrates noted development in the forecasting results. For example, the US election 2012 event achieves the best prediction results in all methods, which indicates that a local event that appears in the USA has a very great effect on stock market future forecasting.

In [63], the authors predicted the Argentinian stock market by using online message boards with topic discovery methods in addition to daily historical stock prices. The authors exploited Latent Semantic Analysis (LSA) approach that finds the latent topics in the text. The experiments are trained with multiple combinations of features selected from online texts. The results show that the most predictive features are derived from the texts that contain the most relevant semantic content. Moreover, the experiments illustrate that combining LSA with ridge regression was able to identify the structure of the texts that later improves the prediction performance of the model.

In [57], the authors proposed a model that aims to find the influence of negative terms represented by the financial media on investor behavior. The proposed model relies on the counting of negative words from the dictionary and word counting methods to extract contextual information. The model also used a Latent Dirichlet allocation model to derive the financial media statements of negative influence. The model combines the two inputs in an ensemble tree to categorize the effect of financial media news on stock market fluctuation. The results indicate that there is a strong relationship between negative effect derived from financial media news and a company stock market fluctuation.

In the same year, authors in [58] suggested algorithm predicts 30 NASDAQ and New York stock exchange companies' movements. The algorithm used NLP

methods to categorize Twitter messages. Then the authors applied association rules to find interesting rules and associations between the stock movements and the Twitter messages. The collected tweets were about 15 million Twitter messages. The big data then stored it in MongoDB, which is an open-source database used to save and process the huge data. The suggested method has explained the relationships hidden in social media as a graph with several layers, with the top layer, intermediate layer, and the bottom layer attributes to show the relations. The proposed method has increased the dimensionality of whole variables that would measure the hidden and embedded data among the Twitter messages. The results indicate the outstanding performance of using tweet message sentiment to predict the stock market movements 3 days later.

In [60], the researchers exploited the multiple kernel learning method to integrate data from the stock special (SS) and subindustry special (SIS) news items effectively to predict future market movements. Multiple kernel learning (MKL) applies many different kernels to learn from various sections of data. Pairs of Gaussian, linear, and polynomial kernels were used to compare each model performance. For evaluation, the authors used five stocks from the S&P 500 index that belongs to managed healthcare subindustry. The results indicate that using Gaussian, linear, and polynomial kernels jointly in MKL achieves higher prediction results. The results also indicate that exploiting two types of news increases prediction accuracy in comparison with models that used only a single news source.

The study in [59] combined information on historical stock prices with financial market news to enhance the market forecasting accuracy of intraday trading status. For evaluation the model used the Hong Kong Stock Exchange (HKEx) tick prices; more specifically the authors used 23 stocks in Hang Seng Index10 (HSI) intraday prices in the year 2001. Multi-kernel support vector regression (MKSVR) was used with two subkernels: one for the news items and the other kernel for the stock historical prices. The results indicate that MKSVR outperforms other benchmark models that exploited only one source of information.

The evaluation measurements vary in all of the reviewed works; some of the researches calculate accuracy, F-measure, or recall and precision with accuracy being the most commonly used. However, other researchers calculated the error in prediction using mean absolute percent error (MAPE), mean squared error (MSE), or root mean square error (RMSE). The variances in using different evaluation measurements and exploratory data make an accurate comparison between different models difficult to achieve.

## **6. The reviewed work text source and period and number of collected items**

The textual data input comes from different several sources, and the period and the numbers of collected data are varied, and all are illustrated in **Table 6**.

The majority of writers have analyzed primary news websites like the Reuters and Bloomberg [16, 17, 37, 47, 50], Dow Jones [57], and Yahoo Finance [8, 18]. Most authors use financial news because it is associated with less noise compared to the general news. They either select the news text or the news headline as input to their machine learning model. Recently news titles and headlines are specifically extracted and are regarded to be more clear, concise, and associated with less noise [14, 16, 17, 50]. Other authors have examined less formal sources of news information such as Google News [12, 15]. Other researchers collect their textual information merely from social media websites especially Twitter to analyze the public user sentiments to predict the market more effectively [39, 56, 61, 64].

Reference	Text type and source	Period	Number of collected items
Vu et al. [56]	Twitter user messages	April 1, 2011 to May 31, 2011	5,001,460 daily tweets
Porshnev et al. [31]	Twitter	April 132,013, to September 29, 2013	About 755 million tweets
Martin [36]	Twitter data.	February 27, 2013, to June 16, 2013	About 25,930 tweets
Li et al. [28]	23 stocks in Hang Seng Index10 (HSI) intraday prices and financial news from the website Caihua, <a href="http://www.finet.hk/">http://www.finet.hk/</a>	Intraday prices of the year 2001	28,885 pieces of news
Xu and Keelj [32]	StockTwits	March 13, 2012, to May 25, 2012	100,000 tweets
Bing et al. [58]	Twitter messages	October 2011 to March 2012	15 million Twitter messages
Li et al. [28]	Financial news articles from FINET (a main financial news seller in Hong Kong)	January 2003 to March 2008	Not mentioned
Geva and Zahavi [37]	Financial news from Reuters 3000 Extra Service	September 15, 2006, to August 31, 2007	51,263 news items
Moniz and de Jong [57]	News source is a corpus extracted from Dow Jones Newswires (DJNW). News articles are collected from financial blogs, online newspapers, financial magazines, and many online websites	January 1, 2009, to December 31, 2013	The corpus consists of 35,678 daily news stories
Ding et al. [16]	Financial news titles from Reuters and Bloomberg	October 2006 to November 2013	442,933 for training 110,733 for development 110,733 for testing
Cakra and Trisedya [61]	Twitter data	April 14, 2015 to April 30, 2015	Not mentioned
Nguyen et al. [8]	Texts in a message board from Yahoo Finance Message Board	July 2012 to July 2013	The different numbers of messages for each stock that follows between 89 and 11,220 in maximum
Shynkevich et al. [60]	News of 5 stock from the S&P 500 index that belongs to managed healthcare sub from LexisNexis database	September 1, 2009, to September 1, 2014	More than 400 news articles
Ghanavati et al. [62]	News summaries (source not mentioned)	June 1, 2014, and June 1, 2015	Not mentioned

Reference	Text type and source	Period	Number of collected items
Khatri and Srivastava [39]	Twitter and StockTwits	January 1, 2015, to February 22, 2016	Not mentioned
Gálvez and Gravano [63]	Message board texts from the webpage <a href="http://foro.ravaonline.com">http://foro.ravaonline.com</a> .	June 1, 2010, and July 31, 2015	More than 20,000 posts
Weng et al. [15]	Wikipedia company pages and Google news	May 1, 2012, to June 1, 2015	Not mentioned
Chen et al. [40]	Sina Weibo news and posts	January 1, 2015, to March 8, 2017	Not mentioned
Li et al. [52]	Forum posts from <a href="http://guba.eastmoney.com">guba.eastmoney.com</a>	January 1, 2009, to October 31, 2014	More than 18 million posts
Kraus and Feuerriegel [51]	German ad hoc announcements from <a href="http://www.dgap.de">www.dgap.de</a>	2010–2013	10,895 observations
Huynh et al. [47]	Financial news from Reuters and Bloomberg websites	2006–2013	5816 news for training 2904 news for testing
Khedr and Yaseen [18]	News data from different resources, Google finance, Reuters, wall street journal, <a href="http://marketwatch.com">marketwatch.com</a> , <a href="http://zacks.com">zacks.com</a> , Yahoo Finance, and <a href="http://economics.com">economics.com</a> , <a href="http://nasdaq.com">nasdaq.com</a>	Not mentioned	Not mentioned
Vargas et al. [17]	Financial news title from Reuters and Bloomberg	October 2006 to November 2013	13,149 for training 1976 for development 2046 for testing
Liu [50]	Financial news titles collected from Reuters and Bloomberg	2006–2013	445,262 for training 55,658 for development 55,658 for testing
Zhang et al. [13]	Financial news articles from financial news websites in China and Xueqiu social media posts	2015–2016	38,727 news in 2015 and 39,465 news in 2016 6,163,056 posts for 2015 and 2016
Xie and Jiang [27]	Financial news of Wallstreetcn, Stockstar, China news, and many other resources	2008 and 2015	2,302,692 news items
Long et al. [30]	Financial news from <a href="http://ifeng.com">ifeng.com</a> financial channel in China	September 2012 to March 2017	18 news per day at maximum
Shastri et al. [41]	News headlines from <a href="http://www.nasdaq.com/">http://www.nasdaq.com/</a>	2013–2016	Not mentioned
Picasso et al. [42]	News articles from <a href="http://intrinio.com">intrinio.com</a> API	July 3, 2017 to June 14, 2018	Not mentioned



Reference	Text type and source	Period	Number of collected items
Chen et al. [12]	News articles from Google	January 4, 2016, to December 29, 2017	130,000 articles
Liu and Wang [14]	News headlines from five famous financial news websites in china	January 1, 2016, to December 31, 2016	780,920 financial news headlines
Maqsood et al. [64]	Twitter data	2000–2018	11.42 million tweets

**Table 6.**  
*Summaries of the reviewed work text source, period, and number of collected items.*

Also, as **Table 6** illustrated, the data were collected in a variety of periods; some few papers collected data in several months, while others extracted data within a maximum of 7-year period, which resulted in more sufficient data and better results in prediction.

However, it can be noted that the insufficiency of highly structured datasets containing text data of markets prevents researchers from accumulating their analysis and assessment efforts with others. Another problem is the imbalanced dataset that has been used by many researchers, which is discriminating the accuracy of prediction. In future, potential researchers are encouraged to locate new datasets for market forecasting based on text mining analysis.

Market predictive text mining could become much more advanced by concentrating on a particular source of text, such as a specific social media website or the new news source from specialized financial news websites. As mentioned in Section 3 of this research, there is a strong relationship between the behavioral economics and the market fluctuations; due to this fact focusing on behavioral economics studies and its impact on market movements will be of great research opportunity in the future.

7. The reviewed work findings, limitations, and future work

Developments in sentiment analysis approaches and deep learning have enabled the development of stock market prediction systems to turn future web content, tweets and financial, and news contents into investment decision systems. Online text mining processes are evolving and have been intensively investigated using machine learning advancements, and this trend will continue to achieve progression especially for market prediction.

Many researchers believe that analyzing only the historical prices of the stock market will be able to predict the stock market movement [3–5]. However, other researchers combine both textual information with historical prices of stock to predict the stock market movements [8, 13, 15, 47, 62]. The previous studies’ major limitation is that they depend heavily on either structured data (historical stock prices) or unstructured data (news articles or social media). However, for the researchers that used both structured and unstructured data, the major limitation for most of them is that they combined either news articles or social media with past stock prices to predict the stock movements and they neglect the critical impact of combining social media and financial news information’s with time series market data to improve the forecasting results.

Reference	Finding	Limitation and future work	Year
Vu et al. [56]	Using the named-entity recognition method for noise removal of Twitter data improves the accuracy results.	Increase the collected tweet data and the collection period and expand the number of companies.	2012
Porshnev et al. [31]	It is possible to increase the prediction accuracy using human sentiment analysis and a lexicon-based approach.	They need to expand the training period to achieve better outcomes. Use more effective sentiment analysis method to increase the prediction accuracy.	2013
Martin [36]	Twitter sentiment analysis using the neural network can be used to predict the stock market movements.	Adding a different source of information such as financial news articles will be able to improve the prediction performance more.	2013
Li et al. [28]	The sentiment analysis model performs better than a bag-of-words model inaccuracy measures. There was a small difference between using the two models, LMD and HVD.	We need to automatically expand the HVD and LMD dictionaries without affecting the accuracy of the dictionary.	2014
Xu and Keelj [32]	The result shows the outstanding performance of SVM for the sentiment classification task.	Expand the data analysis period. Use a more effective expanded lexicon. Exploit the user profile features.	2014
Geva and Zahavi [37]	NN algorithm is profitable for any initial investment. Combining market data with financial news can predict the market movement with better accuracy.	Study the effect of using other prediction models, and investigate the impact of using different textual data processing.	2014
Moniz and de Jong [57]	There is a strong relationship between negative affect derived from financial media news and a company stock market fluctuation.	Adding social media data to the dataset to improve the prediction performance.	2014
Bing et al. [58]	The study algorithm has an outstanding performance in using tweet message sentiment to predict the stock market movements 3 days later.	Needs to add other textual sources for social media data such as Facebook. Adding news items to the dataset.	2014
Li et al. [59]	The results indicate that MKSVR outperforms other benchmark models that used only one source of information.	Adding more sources of textual data. Apply more subkernel using the same textual data. Positive and negative news could be classified by using the use of sentiment analysis to categorize positive and negative news. The use of multiple subkernels for each news in different sentiment classes.	2014

Reference	Finding	Limitation and future work	Year
Ding et al. [16]	CNN can extract the longer-term influence of financial news events than traditional feedforward neural networks.	Adding different textual data sources and improvement in classification algorithm will yield a better result.	2015
Nguyen et al. [8]	The proposed model proves to be more effective than using the human sentiment method for stock movement prediction.	They have to define the number of topics and sentiment beforehand. The model can predict the stock movements either up or down only and can be improved to predict the degree of the movements. Adding different text data sources like financial news.	2015
Cakra and Trisedya [61]	The highest classification accuracy was achieved by using the random forest classification model.	Have to expand the data collection period. Needs to improve the sentiment classification model by adding different features.	2015
Shynkevich et al. [60]	Using of Gaussian, linear, and polynomial kernels jointly in MKL achieves higher prediction results. Exploiting two types of news increases the prediction accuracy in comparison with models that used only a single news item.	Add historical stock prices to the dataset with the news articles to enhance the prediction results.	2015
Khatri and Srivastava [39]	It is better to invest in a company whose sentimental score is high and positive rather than choosing a close price as an indicator of stock movements.	The datasets should be taken for a longer time to achieve better results.	2016
Ghanavati et al. [62]	The metric learning methods can improve the results. Adding news to the historical prices for stock market prediction will be more useful on large and popular stocks.	Needs to add the different sources of textual information like social media.	2016
Weng et al. [15]	Incorporating data from multiple sources will improve the efficiency of market prediction.	The use of different rank values selected from a different data source. Expand the work to include the certainty level of the prediction, which can be achieved by using Bayesian Belief Networks (BBN) or ensemble methods. Try to forecast the actual price instead of the movement. Adding other data sources also will increase the prediction performance.	2017
Chen et al. [12]	News and posts on Sina Weibo can predict the market movements.	The use of more improved machine learning techniques for sentiment analysis such as interdependent Latent Dirichlet allocation (ILDA) will improve the prediction performance.	2017

Reference	Finding	Limitation and future work	Year
Li et al. [52]	There is a strong relationship between investor sentiments and CSI300 prices.	Utilized only naive Bayes algorithm for classification and did not test other classification methods that may achieve better results.	2017
Huynh et al. [47]	BGRU obtained the best results in predicting the market compared to other LSTM variant.	Adding another textual source of information such as social media may enhance the model performance.	2017
Kraus and Feuerriegel [51]	LSTM models surpass all traditional machine learning models based on the bag-of-words technique, specifically when using transfer learning to pretrain word embeddings.	Increasing the number of collected news for a longer time and applying the deep learning model will improve the predictive performance.	2017
Vargas et al. [17]	CNN is more effective than RNN on capturing semantic from financial news. RNN is more effective in capturing the context information for the stock market prediction. Sentence embedding for text representation is more effective than the word embedding.	Exploiting the reinforcement learning models to train the proposed methods on trading simulation may yield better results.	2017
Khedr and Yaseen [18]	Adding historical stock prices to the classification model will be able to improve the prediction performance.	Adding technical analysis and social media sentiment analysis will improve the prediction results.	2017
Gálvez and Gravano [63]	The results indicate that the most predictive features derived from the texts that contain the most relevant semantic content. Moreover, the results prove that combining LSA with ridge regression was able to identify the structure of the texts, which improves the prediction performance of the model.	Adding even sentiment and more text resources such as social media data will improve the results.	2017
Checkley et al. [21]	There is a causal link between Twitter sentiments to stock market returns, volatility, and volume. Among all five stocks, market volatility and volume seem to be more predictable than market direction or return.	The consideration of event sentiment may affect the market return more and improve the forecasting result.	2017
Bujari et al. [24]	Some of the proposed ad hoc forecasting models well predict the next day direction of the stock movements for some particular companies with 82% of success, and there is no unified method to be used with all cases. The more volume of a tweet will yield better prediction results. There is a strong correlation between tweet posts and the trend movements for some companies.	Investigate another source of textual information such as online financial news.	2017



Reference	Finding	Limitation and future work	Year
Zhang et al. [13]	Both news events and market historical data have a more important effect on stock movements than social media sentiments. Both news events and quantitative data have larger impacts to drive stock fluctuations than sentiments.	Increasing the dataset collection period may improve prediction performance.	2018
Liu [50]	Adding news articles was able to predict the individual stock prices with better accuracy compared to predicting the market using time series prices alone.	Predicting price changes at a different time horizon in the future to achieve better performance. The study used the full corpus as input for the prediction model, which may add noise to the data and affect prediction accuracy.	2018
Xie and Jiang [27]	Both audience numbers and news quality have a crucial impact on the stock market.	Have to develop a better sentiment evaluation system.	2019
Long et al. [30]	Content structure hidden in daily financial news can successfully predict the stock market movements. Financial news influence on stock movements lasts for 2–3 days.	Adding the structural information to the prediction model will be able to improve the prediction performance. The use of different models to process news texts may also improve the results.	2019
Shastri et al. [41]	Stock price forecasting is more efficient for a shorter time.	Upgrade the sentiment analysis task by increasing the words that may affect the stock movements more.	2019
Picasso et al. [42]	The model with the LMD dictionary is more effective in annualized return measure, while the use of AS dictionary proved to be more effective in obtaining high accuracy results.	The model could not achieve overwhelming results compared to using news set alone. The use of advanced feature fusion methods will improve the results. Collect more news data for a longer period.	2019
Chen et al. [12]	Individual stock prediction using the study polarity lexicon was better than the benchmark model.	The research did not analyze detailed data; it only has the data that can be achieved by any public users.	2019
Liu and Wang [14]	NBAd structure accomplishes the best outcomes. Market predictions of the stock price at the minute time frame obtain better outcomes than those at day level.	Apply the NBA model in an index or industry-level data.	2019
Mudinas et al. [19]	In some cases, sentiment emotions contribute to Granger-cause stock price fluctuates, but the finding was not inclusive and must be examined for each case. For some stocks, adding sentiment emotions to the machine learning market prediction model will increase the prediction accuracy. SVM achieved better and more balanced results.	Enhancing the sentiment classification model and increasing the number of collected items will yield a better result.	2019

Reference	Finding	Limitation and future work	Year
Souza and Aste [22]	Multiplex network approach incorporating information from both social media and financial data can be used to forecast the causal relationship framework with high accuracy.	Investigate the impact of financial crises by expanding the historical data period. Use different techniques of the financial correlation establishment and apply it to portfolio management mechanisms.	2019
Wu et al. [23]	Adding news articles was able to reduce the RMSE that proves that the economic news has crucial impacts on market returns. The negative news has more influence on the stock market returns than positive news articles.	The research only tested the news texts published in the Knowledge Management Winner newspaper. Future study may include other online news datasets. Apply the proposed model to examine the stocks of smaller companies.	2019
Maqsood et al. [64]	Not all the main events have a crucial impact on stock market movements. More crucial local events affect the performance of the prediction model. Support vector regression gives the best prediction performance	Needs to exploit more than one social media website to produce sentiment analysis for a specific event. The use of financial news may improve the prediction result.	2020

**Table 7.**  
*Summaries of the reviewed work findings, limitations, and future work.*

Moreover, as **Tables 2–5** indicate, the main trends in recent studies are to utilize deep learning methods instead of conventional machine learning to analyze the stock market textual information in the news or social media due to the advantages of DL that offer overconventional machine learning. DL promises enough amount of data and training time that conventional machine learning methods are unable to handle effectively.

Many recent studies only exploit sentiment analysis of textual data, and they neglect the important influence of historical stock prices, which affect their prediction accuracy results; this suggests that the incorporation of data from multiple sources will improve market prediction effectiveness. The more data fed into the prediction model, the better accuracy can be achieved.

Machine learning models described previously have been discussed to show how SVM and LSTM are highly preferred by investigators because of their high accuracy result in text classification and market prediction, whereas many other machine learning methods like K-nearest neighbors (k-NN), random forest (RF), linear regression, decision tree, artificial neural networks (ANN), etc. illustrate promising results for text mining and sentiment analysis task for market analysis but are least frequently used and need to be further investigated.

However, the reviewed work has some limitations; one of the main limitations is the insufficiency of highly structured datasets containing text data on markets for certain periods that researchers can use to integrate their analysis and assessment efforts; another problem is the imbalanced dataset that has been used by many researchers, which make discriminating result in prediction.

Future work should focus on predicting the movement of the stock market using structured data (past stock prices) along with textual data from different resources like financial news and social media. Moreover, to achieve better results

in predicting the stock market, the text mining procedure should improve feature selection, feature representation, and dimensionality reduction methods.

In general, many techniques will be able to improve the prediction methods such as adding the structural information to the prediction model, expanding the training period, using more effective expanded lexicons, adding different sources of information such as financial news articles, increasing the number of collected news for longer period, applying the deep learning models, upgrading the sentiment analysis task by increasing the words that may affect the stock movements more, using of more improved machine learning techniques for sentiment analysis such as Interdependent Latent Dirichlet allocation (ILDA), adding historical stock prices to the dataset with the news and social media information, and considering of event sentiments analysis as illustrated in **Table 7**.

## 8. Conclusion

Knowledge of stock movements by a fraction of a second can lead to high profits investors can make which makes stock market studies a major motivation for a researcher. The great advances and success of natural language process and sentiment analysis of online news based on machine learning and deep learning have gained huge popularity recently in the financial domain especially in market prediction models. This survey has discussed the recent current studies on market prediction systems based on text mining techniques with comprehensive clarifying of the model's main limitations and future improvement methods. The survey was undertaken on many major portions such as text preprocessing, machine learning algorithms, evaluation mechanisms, findings, and limitations associated with detailed discussion and explanation of the most successful used techniques. Moreover, this review provides a serious attempt to address the problem of market prediction based on the most recent text mining methods and provide a clear view of the future research direction. Recently, more extensive observations into the financial markets are required in the current dynamic world, since the absence of it can have a detrimental effect on the investments around the globe. It is therefore essential to undertake prediction models based on text mining research as a practical solution that can lead to a much greater degree of confidence in the understanding of market movements and make valuable investments. With the considerable amount of textual data available online, the need to build specialized text mining systems gradually evolves for each field of market analysis.

This study is intended to support other researchers to place the different theories in this research area more easily into practice and become able to make key decisions in the development of future models. The researches mentioned in this paper proved the effectiveness of text mining and sentiment analysis methods in predicting market movements. By comparing many ML methods such as SVM or decision tree and deep learning models like LSTM or CNN, we discussed some of these model's limitations and future work and debated the best result obtained by each one of these models. After all, the proposed survey displayed the need of improving the prediction methods such as adding the structural information, considering of event sentiments analysis, using more effective expanded lexicons, increasing the number of collected news, expanding the training period, applying the deep learning models, adding different sources of information, upgrading the sentiment analysis task by increasing the words that may affect the stock movements more, and using unified benchmark dataset and evaluation measures.

IntechOpen

IntechOpen

### **Author details**

Faten Subhi Alzazah\* and Xiaochun Cheng  
Department of Computer Science, Middlesex University, London, UK

\*Address all correspondence to: [fatensubhi@gmail.com](mailto:fatensubhi@gmail.com)

### **IntechOpen**

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 



## References

- [1] Gupta A, Dhingra B. Stock market prediction using hidden Markov models. In: 2012 Students Conference on Engineering and Systems. IEEE; 2012. pp. 1-4
- [2] Asadi S, Hadavandi E, Mehmanpazir F, Nakhostin MM. Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction. *Knowledge-Based Systems*. 2012;**35**:245-258
- [3] Saravanan S, Mala S. Stock market prediction system: A wavelet based approach. *Applied Mathematics and Information Sciences*. 2018;**12**:579-585. DOI: 10.18576/amis/120312
- [4] Chung H, Shin KS. Genetic algorithm-optimized long short-term memory network for stock market prediction. *Sustainability*. 2018;**10**(10):3765
- [5] Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*. 2019;**164**:163-173
- [6] Agarwal B, Mittal N, Bansal P, Garg S. Sentiment analysis using common-sense and context information. *Computational Intelligence and Neuroscience*. 2015;**2015**
- [7] Rajput V, Bobde S. Stock market forecasting techniques: Literature survey. *International Journal of Computer Science and Mobile Computing*. 2016;**5**(6):500-506
- [8] Nguyen TH, Shirai K, Velcin J. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*. 2015;**42**(24):9603-9611
- [9] Sun A, Lachanski M, Fabozzi FJ. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*. 2016;**48**:272-281
- [10] Schumaker RP, Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*. 2009;**27**(2):1-9
- [11] Ming F, Wong F, Liu Z, Chiang M. Stock market prediction from WSJ: Text mining via sparse matrix factorization. In: 2014 IEEE International Conference on Data Mining. IEEE; 2014. pp. 430-439
- [12] Chen MY, Liao CH, Hsieh RP. Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. *Computers in Human Behavior*. 2019;**101**:402-408
- [13] Zhang X, Qu S, Huang J, Fang B, Yu P. Stock market prediction via multi-source multiple instance learning. *IEEE Access*. 2018;**6**:50720-50728
- [14] Liu G, Wang X. A numerical-based attention method for stock market prediction with dual information. *IEEE Access*. 2018;**7**:7357-7367
- [15] Weng B, Ahmed MA, Megahed FM. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*. 2017;**79**:153-163
- [16] Ding X, Zhang Y, Liu T, Duan J. Deep learning for event-driven stock prediction. In: Twenty-Fourth International Joint Conference on Artificial Intelligence; 2015
- [17] Vargas MR, De Lima BS, Evsukoff AG. Deep learning for stock market prediction from financial news

- articles. In: 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). IEEE; 2017. pp. 60-65
- [18] Khedr AE, Yaseen N. Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*. 2017;**9**(7):22
- [19] Mudinas A, Zhang D, Levene M. Market trend prediction using sentiment analysis: Lessons learned and paths forward. 2019. arXiv preprint arXiv:1903.05440
- [20] Granger CW. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*. 1969;**1**:424-438
- [21] Checkley MS, Higón DA, Alles H. The hasty wisdom of the mob: How market sentiment predicts stock market behavior. *Expert Systems with Applications*. 2017;**77**:256-263
- [22] Souza TT, Aste T. Predicting future stock market structure by combining social and financial network information. *Physica A: Statistical Mechanics and its Applications*. 2019;**535**:122343
- [23] Wu GG, Hou TC, Lin JL. Can economic news predict Taiwan stock market returns? *Asia Pacific Management Review*. 2019;**24**(1):54-59
- [24] Bujari A, Furini M, Laina N. On using cashtags to predict companies stock trends. In: 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE; 2017. pp. 25-28
- [25] Dumais S, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: *Proceedings of the Seventh International Conference on Information and Knowledge Management*; 1998. pp. 148-155
- [26] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: *European conference on machine learning*. Berlin/Heidelberg: Springer; 1998. pp. 137-142
- [27] Xie Y, Jiang H. Stock market forecasting based on text mining technology: A support vector machine method. 2019. arXiv preprint arXiv:1909.12789
- [28] Li X, Xie H, Chen L, Wang J, Deng X. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*. 2014;**69**:14-23
- [29] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*. 2011;**66**(1):35-65
- [30] Long W, Song L, Tian Y. A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. *Expert Systems with Applications*. 2019;**118**:411-424
- [31] Porshnev A, Redkin I, Shevchenko A. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In: 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE; 2013. pp. 440-444
- [32] Xu F, Keelj V. Collective sentiment mining of microblogs in 24-hour stock price movement prediction. In: 2014 IEEE 16th Conference on Business Informatics, Vol. 2. IEEE; 2014. pp. 60-67
- [33] Uysal AK, Murphey YL. Sentiment classification: Feature selection based

approaches versus deep learning. In: 2017 IEEE International Conference on Computer and Information Technology (CIT). IEEE; 2017. pp. 23-30

[34] Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*. 2018;5(1):3

[35] Singhal P, Bhattacharyya P. *Sentiment Analysis and Deep Learning: A Survey*. Bombay: Center for Indian Language Technology, Indian Institute of Technology; 2016

[36] Martin V. Predicting the French stock market using social media analysis. In: 2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization. IEEE; 2013. pp. 3-7

[37] Geva T, Zahavi J. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision Support Systems*. 2014;57:212-223

[38] Goldberg DE. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley; 1989

[39] Khatri SK, Srivastava A. Using sentimental analysis in prediction of stock market investment. In: 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE; 2016. pp. 566-569

[40] Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS. Stock market prediction using neural network through news on online social networks. In: 2017 International Smart Cities Conference (ISC2). IEEE; 2017. pp. 1-6

[41] Shastri M, Roy S, Mittal M. Stock price prediction using artificial neural model: An application of big data. *EAI*

*Endorsed Transactions on Scalable Information Systems*. 2019;6(20)

[42] Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*. 2019;135:60-70

[43] Cambria E, Fu J, Bisio F, Poria S. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In: Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015

[44] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997;9(8):1735-1780

[45] Rao G, Huang W, Feng Z, Cong Q. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*. 2018;308:49-57

[46] Siami-Namini S, Namin AS. Forecasting economics and financial time series: ARIMA vs. LSTM. 2018. arXiv preprint arXiv:1803.06386

[47] Huynh HD, Dang LM, Duong D. A new model for stock price movements prediction using deep neural network. In: *Proceedings of the Eighth International Symposium on Information and Communication Technology*; 2017. pp. 57-62

[48] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*; 2013. pp. 3111-3119

[49] Wang Y, Huang M, Zhu X, Zhao L. Attention-based LSTM for aspect-level sentiment classification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; 2016. pp. 606-615



- [50] Liu H. Leveraging financial news for stock trend prediction with attention-based recurrent neural network. 2018. arXiv preprint arXiv:1811.06173
- [51] Kraus M, Feuerriegel S. Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*. 2017;**104**:38-48
- [52] Li J, Bu H, Wu J. Sentiment-aware stock market prediction: A deep learning method. In: 2017 International Conference on Service Systems and Service Management. IEEE; 2017. pp. 1-6
- [53] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th International Conference on Machine Learning*; 2008. pp. 160-167
- [54] Ho CC, Baharim KN, Fatan AA, Alias MS. Deep neural networks for text: A review. In: *The 6th International Conference on Computer Science and Computational Mathematics*. Langkawi, Malaysia; 2017
- [55] Ding X, Zhang Y, Liu T, Duan J. Using structured events to predict stock price movement: An empirical investigation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014. pp. 1415-1425
- [56] Vu TT, Chang S, Ha QT, Collier N. An experiment in integrating sentiment features for tech stock prediction in twitter. In: *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*; 2012. pp. 23-38
- [57] Moniz A, de Jong F. Classifying the influence of negative affect expressed by the financial media on investor behavior. In: *Proceedings of the 5th Information Interaction in Context Symposium*; 2014. pp. 275-278
- [58] Bing L, Chan KC, Ou C. Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In: 2014 IEEE 11th International Conference on e-Business Engineering. IEEE; 2014. pp. 232-239
- [59] Li X, Huang X, Deng X, Zhu S. Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. *Neurocomputing*. 2014;**142**:228-238
- [60] Shynkevich Y, McGinnity TM, Coleman S, Belatreche A. Stock price prediction based on stock-specific and sub-industry-specific news articles. In: 2015 International Joint Conference on Neural Networks (IJCNN). IEEE; 2015. pp. 1-8
- [61] Cakra YE, Trisedya BD. Stock price prediction using linear regression based on sentiment analysis. In: 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE; 2015. pp. 147-154
- [62] Ghanavati M, Wong RK, Chen F, Wang Y, Fong S. A generic service framework for stock market prediction. In: 2016 IEEE International Conference on Services Computing (SCC). IEEE; 2016. pp. 283-290
- [63] Gálvez RH, Gravano A. Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Scienc*. 2017;**19**:43-56
- [64] Maqsood H, Mehmood I, Maqsood M, Yasir M, Afzal S, Aadil F, et al. A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management*. 2020;**50**:432-451



# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Modern Business Intelligence: Big Data Analytics and Artificial Intelligence for Creating the Data-Driven Value

*Ahmed A.A. Gad-Elrab*

## Abstract

Currently, business intelligence (BI) systems are used extensively in many business areas that are based on making decisions to create a value. BI is the process on available data to extract, analyze and predict business-critical insights. Traditional BI focuses on collecting, extracting, and organizing data for enabling efficient and professional query processing to get insights from historical data. Due to the existing of big data, Internet of Things (IoT), artificial intelligence (AI), and cloud computing (CC), BI became more critical and important process and received more great interest in both industry and academia fields. The main problem is how to use these new technologies for creating data-driven value for modern BI. In this chapter, to meet this problem, the importance of big data analytics, data mining, AI for building and enhancing modern BI will be introduced and discussed. In addition, challenges and opportunities for creating value of data by establishing modern BI processes.

**Keywords:** Business Intelligence, Big Data Analytics, Artificial Intelligence, IoT, Data mining, Data governance

## 1. Introduction

Recently, in the fourth industry revaluation, there is a very huge amount of created and generated data by computer machine such as GPS, sensors, website or application systems or by people through social media (twitter, Facebook, Instagram, or LinkedIn) [1]. Every moment, the data servers store huge amount of data which are produced by organizations. This is a huge amount of data comes from website, social media, tracking, IoT applications, sensors, and online news articles. Also, the advancement in computing and communication technologies have facilitated collecting a large volume of heterogeneous data from multiple sources. This data consists of structured and unstructured, complex and simple information.

Currently, business gets a revenue from the analysis of such data with unstructured form up to 80% [2]. So, the organization can improve the business productive process due to this analysis of unstructured data that contains valuable information. In addition, it is significant for education, security, healthcare, and manufacturing.

This can be achieved through big data analytics, artificial intelligence, and data management in order to achieve the business intelligence.

BI is the technologies, tools, systems, and applications for the compilation, analysis, combination, and exhibition of the business report with active business decision executing way. This way will give unlimited help to gain, learn and control their data to further decision making for developing business processes and procedures [3]. Also, BI can be described as the ability of a firm to make meaningful data using which is collected every day from business processes and operations [4].

Business intelligence (BI) plays an importance role for helping the decision marker to get the insights for improving productive or better and fast decision. In addition, BI can enhance and assist the effectiveness of operational rules and its impression on corporate-level decision-making, superintendence system, administration, budgeting, and financial recording which gives better strategic alternatives in dynamic business environments [5]. Also, BI can improve the organizational performance by identifying new opportunities, revealing new business insights, highlighting potential threats, and enhancing decision making processes among many other benefits [6, 7].

The first issue in business is big data management with various data formats which is the serious management problem due to that the current tools are not adequate for managing such massive big data volumes [8]. The new challenges in terms of data integration complexity, storage capacity, lack of governance, and analytical tools gives an importance for solving the big data management problem related to pre-processing, processing, security and storage. The big data management in abundant data, generated by heterogeneous sources for using in BI and decision-making, is a complex process. Therefore, some form of big data may be managed by 75% of organizations. The goal of managing big data is ensuring the effectiveness of security, storage, and analytics applications of big data [9].

Unfortunately, the practical implications using big data analytics in enhancing business intelligence remains comparatively immature and under-researched because the existing research models are mainly focused on the benefits and challenges of business intelligence and big data. So, the most important issues are studying the implications of big data analytics on business intelligence in for data collected from various sources and exploring the future directions to find further developments in use of big data analytics for business intelligence.

The second issue in BI is the determination of the most appropriate data mining technique, which is one of the most critical responsibilities. Based on business nature and difficulty suffered or object kind in the business there is a need for determining the data mining optimal technique [10]. In the data mining process, most core techniques identify the character of the reclamation option of data and its mining process. Based on the results, the data mining technique will be highly productive [11]. There are many data mining techniques as association rule, clustering, classification, decisions tree, and neural networks are profoundly successful and practical.

Data Mining is regarding to interpret the huge volume data and extract knowledge from its various objects. For some businesses, the purposes of data mining are recognized to identify different trims, develop marketing abilities, and predict the prospect based on earlier observations and modern inclinations. There is a requirement for examining the data for sustaining divestments and additional purposes of an entrepreneur. Furthermore, data mining could continue practiced for recognizing unusual performance and a strange behavior of representatives practicing on some technologies could be identified [12].

The third issue in BI is artificial intelligence (AI). AI is the main step in the technology evolution that has been actively pursued since British mathematician

and code breaker Alan Turing envisioned a clear way forward in his groundbreaking 1950 paper, "Computing Machinery and Intelligence." At the time, computer technology could not keep up with Turing ideas. But, due to the advancement in computing, AI was established. At Oxford University, the Future of Humanity Institute introduced a 2018 report for surveying a panel of AI researchers on timelines for Strong AI. This report found that in 45 years, 50% chance of AI will outperform humans in all tasks and in 120 years it will automat all human jobs. As well as, AI will bring many opportunities for creating new jobs. Also, removing the need to do tedious and repetitive tasks is one of the great values of AI, as many experts said. Instead, users can focus on their main skills and values. For reducing human error, shrinking labor costs, and subsequently increasing profit, the application of technology in many industries and business has been aimed. This was true for the advancements made during the fourth Industrial Revolution (FIR) on through to the birth of the computer, and still true for the era of AI.

In this chapter, the importance of big data analytics, data mining, AI for building modern BI and enhancing will be introduced and discussed. In addition, challenges and opportunities for creating value of data by establishing modern BI processes.

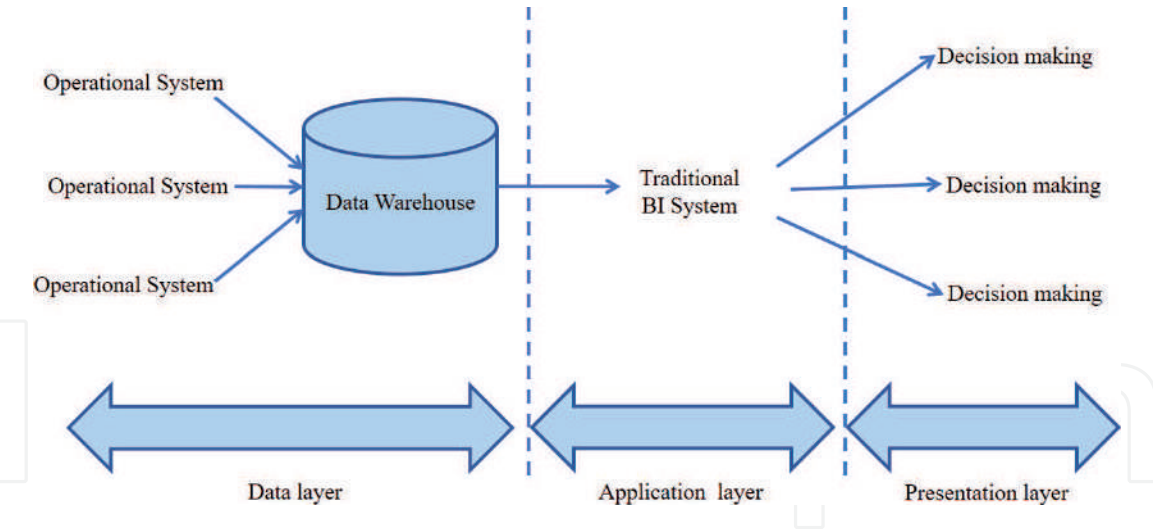
## **2. Business intelligence (BI)**

Business Intelligence (BI) can be described as an automated process for deriving models and insights form raw data that are collected from heterogeneous data sources and are organized in a systematic way for improving business operations and processes. In enterprise BI architectures, the best practice is splitting the data collection and data organization processes that are associated with back-end architecture from data analysis and display to a user through the frontend. In BI, the processed transactions generate data, which are stored in Operational Data Sources called Online Transaction Processing servers (OLTP). With OLTP, the data is stored in a structured data repository called data warehouse after extraction and transformation processes. With data warehouse, there are different query optimization techniques can be applied for speeding-up of data analysis and running the analytics query. To achieve this speed-up, data warehouse creates subsets of the data warehouse called data marts. Also, reporting mechanisms for accessing transaction data stored in data warehouse are used in traditional BI systems. Therefore, analyzing these transaction data can help us for detecting patterns and predicting business trends.

Recently, the data sources of BI are not only traditional data sources as transaction data, but they include modern data sources as mobile devices and sensor data, and web messages which were sent by company intranets and profiles of employees and customers. Most of modern data sources are unstructured, for example, posted messages in online social networks (OSN) and data from various sensors. Therefore, the main challenge is how to maintain these modern data sources as traditional relational database and achieve query efficiency. From the data analysis perspective, additional data means additional opportunities for discovering more insights. However, the big data challenges remain the big problem from the analytic perspective.

Due to the increase in data, there are expanded opportunities within the scope of BI, which is not only a mechanism to analyze historical data trends, but it can combine data from sensors and other real time personal information for inferring insights that are not commonly available that is called situational BI [13]. For business operations, BI is called operational BI, which provides insights in real time to





**Figure 1.**  
*A traditional BI system.*

these operations as getting instant feedback for a call center operation as benefits from their work. In addition, the analytics rules may be composed depend on meta-information of the exposed data to his/her which can be considered as a self-service BI. Therefore, these new BI approaches must be managed carefully such that the compliance models and governance of enterprise are not violated.

The three-tier architecture of traditional BI system is shown in **Figure 1**. This architecture consists of three layers: 1) Presentation layer, 2) Application layer, and 3) Database layer. The main challenge with this three-tier architecture, is how to fulfill service level objectives such as minimal throughput rates and maximal response time. This is because, the data storage management at the low-level layers is hidden from the application layer which makes some difficulties to predict execution times.

However, traditional BI systems are efficient in extracting and analyzing data, but they are rigid, slow, time-consuming, and requiring knowledge experts for maintenance. Therefore, many research works have been done for adding modern features to improve the three-tier architecture, which will establish the next generation BI.

### 3. Modern business intelligence (MBI)

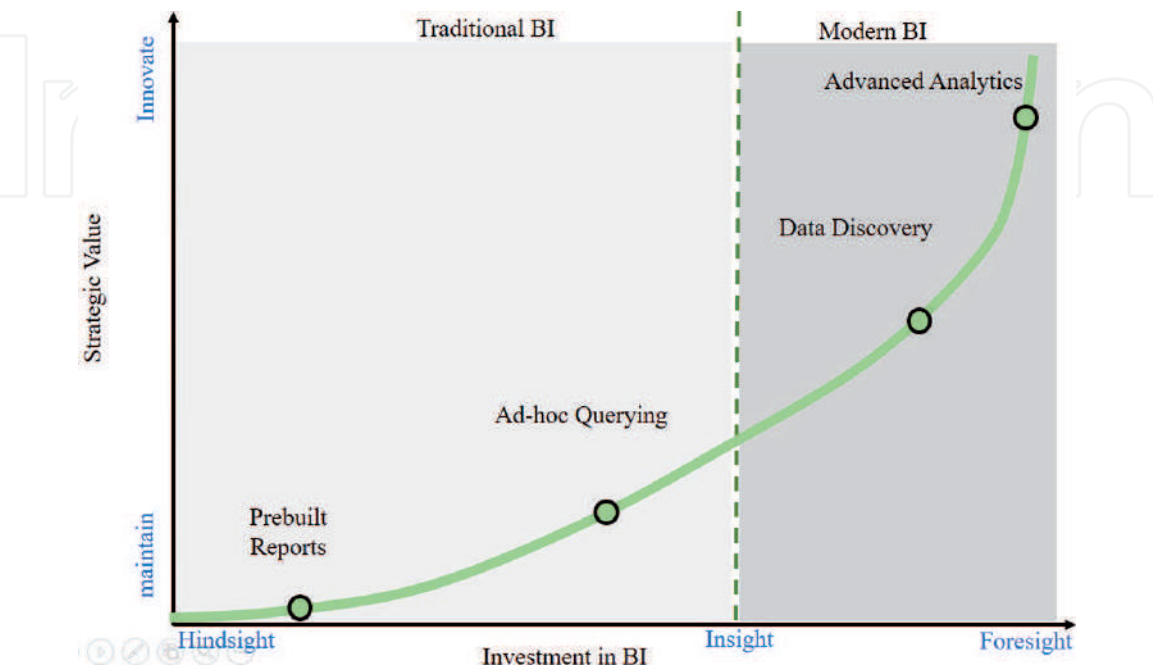
In the traditional BI platforms, the main goal is giving answer “What happened?” Question by providing the efficient analyses. While, the BI modern platforms are giving the answer for “What is happening, what will happen, and why?” which offers the ability to monitor and obtain a continuous development of organization within fast analytics, while for accomplishing objectives of mission using predictive analytics.

Traditional business intelligence platforms over the past two decades have mainly succeeded to provide users with historical comprehensive reports and easy-to-use custom analysis tools. Due to the underlying data architecture, which consists of a central data storage solution such as an enterprise data warehouse (EDW), the availability of BI functionality is largely. EDWs form the backbone of traditional data management platforms and usually connect vast network systems of data source into a central data warehouse. The data is then consolidated, refined, and pulled into different reports and dashboards after converting data in EDW to display old business information, such as weekly revenue metrics or quarterly sales.

Although, this traditionally BI provides a basis for these types of dashboards and interim reports.

While users have gained immense value from traditional platforms for historical reports capabilities, there are more users now require data analysis technologies that need direct access to data without depending on IT professionals. Federal agencies highlighted the following challenges associated with traditional BI solutions in analytics [13]:

1. *On-Demand Analysis Capabilities Lacking*: advanced users of BI today do not need to wait for answers to more business complex problems. Additional users need capabilities of self-service for linking and analyzing specific datasets depend on their own understanding, for any purpose, and at any time.
2. *Needed Predictive Analyses*: Historical reports capabilities provide just one puzzle piece: an insight about what happened in the past. Companies look to predictive analytics or insight about the future to forward thinking and truly be driven by data. With predictive models, the companies can use patterns and forecasts to get next actionable steps using their data.
3. *Mixed Data Types Analysis*: Traditional BI platforms have largely focused on structured data, but today users require the ability for viewing and analyzing semi-structured, unstructured data and third-party data. In recent years, the massive number of produced information has increased, partly due to new data mining technologies, the Internet of Things (IoT), the proliferation of data sensors and automated data collection tools. Now, advanced BI users and data scientists need access to unutilized data in different formats to mix data types and create their own algorithms, where on demand insights are available to make accurate and quick decisions. A lot of organizations that lack the processes, technology, and people needed to extend data-analysis capabilities to the next level become frustrating. These challenges need a strategy and platform for analytics that goes faraway the traditional BI platforms scope, as shown in **Figure 2**.



**Figure 2.**  
*Grows of BI platforms based on insights: Hindsight to insights to foresight [14].*

Integration of traditional and modern BI Platforms is essential to laying the groundwork for enterprise-wide data transformation and organizations are truly concerned for getting rid of IT infrastructure and starting over. Data warehouses play a major role in existing data platforms, which provide the data that fully cleaned, organized, and managed for most businesses and companies. The data warehouse gives business managers, executives and others ability to obtain insights from historical data with relative ease without deep technical knowledge. The obtained data from data warehouse is very accurate due to careful testing, IT cleaning, and accurate knowledge of data layers. However, traditional BI challenges create a demand to increase EDW with different form of optimized architecture for fast access to ever-changing data: Lake Hadoop Data.

Organizations look to upgrade their platforms of analytics are beginning to adopt the data lakes concept. Data lakes store information in its raw and unfiltered form, whether structured, semi-structured, or unstructured. Unlike the stand-alone EDW, the data lakes themselves perform little of the automated data cleaning and transfer operations, allowing data to be swallowed more efficiently, but they transfer the greatest responsibility for preparing and analyzing data to business users.

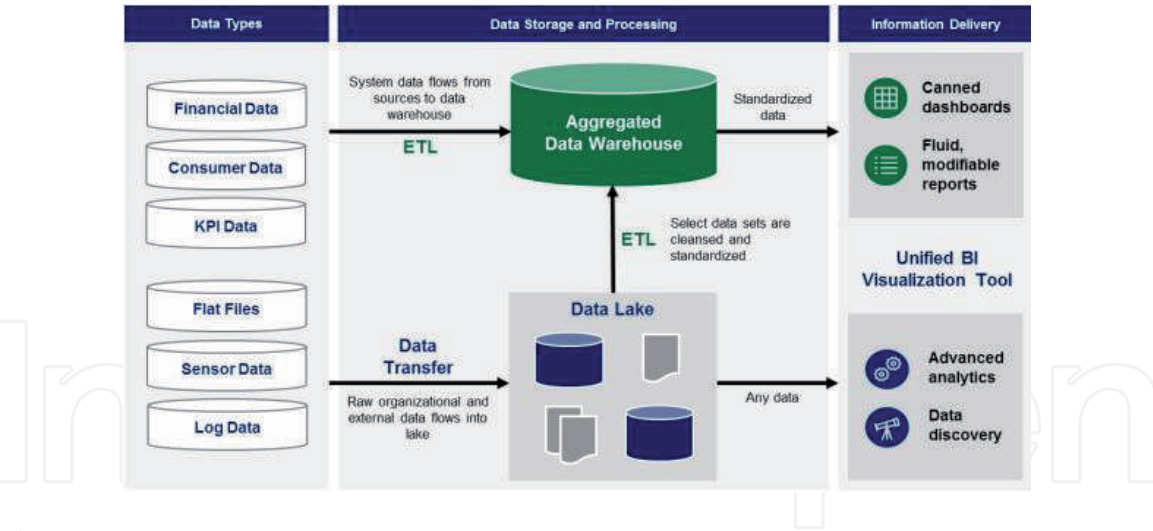
Data Lakes can offer a low-cost solution by using Hadoop's Distributed File System (HDFS) for efficiently storing various types of data and analyzing them in their original structure. As shown in **Figure 3**, a data lake coupled with the data warehouse to identify the next generation of BI and provide the optimal basis for data analysis.

In the system shown in **Figure 3**, EDW receives system data from different sources through the ETL process (Extract, Transform, and Load). After the data is cleaned, transformed, and standardized, it will be ready for analysis by a diverse group of users using dashboards and reports.

In the interim, a data lake collects raw data from single or multiple source systems or all systems, and the data is absorbed and ready for discovering or analyzing processes. The result: a broader user base for exploring and creating relationships between vast amounts of various data for individual analyzes, upon request.

### 3.1 Features of modern BI

1. *Operational BI (real-time)*: Today, the competitive pressure of businesses has increased the requirement for almost real-time BI, which called operational BI. The operational BI goal is reducing latency between data analysis time and data acquisition time. Reducing response time enables the system to take suitable action when an event exists. With operational BI realization, companies can discover patterns or time trends across flow of operational data.
2. *Situational BI*: it enables situational awareness. In companies, BI positioning is important where a rapid turnaround in positions, commonly external business trends, has affected business [15]. However, this external data, which mostly comes from the intranet of company, external vendor, or the Internet, is unstructured. Moreover, this unstructured data must be combined with other structured data from the local data warehouse of the company for supporting real-time decision marking. For example, the company may want to know if its users and customers are posting negative or positive comments about their new products. Through the analysis of these comments, companies can provide immediate comments to the development team for making the product more competitive and qualified. Another example is important for a company



**Figure 3.**  
Data sources, data warehouse, and data analytics in modern BI [14].

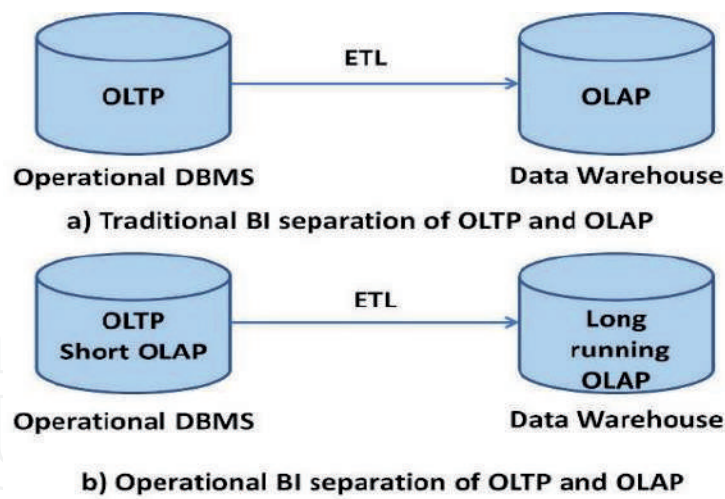
to know whether natural disasters have affected their contract suppliers. Recognizing natural disasters and enable businessmen to take appropriate measures to reduce losses [16].

3. *Self-Service BI (SSBI)*: it enables end users for generating analyzes and analytical queries without involving of the IT department. In SSBI, the user interface of applications must be easy to use and intuitive, therefore technical knowledge of the data repository is not needed. In addition, the user should be allowed for accessing or expanding data sources organized by IT, but also non-traditional sources.

### 3.2 Data architecture

1. *Background*: Traditional business application architecture has three layers: data, application, and presentation. In the three-tier architecture, execution time is very difficult for predicting, due to the relationship between processes of low-level data management and operations of high-level. Usually, workload management solutions are built on top of general-purpose DBMS, which need time delays for executing parallel requests. With modern business applications, this will create challenges for functions as operational information in real time. Therefore, technologies that enable simultaneous business transaction and analytical queries to be performed on the same data are important. Organizations today use the ETL to extract data, make transfers, and upload data that is converted into a data warehouse. This model is based on two types of business process critical processes: Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP). OLTP is used for managing business operations, such as processing of an order. OLAP is used for supporting strategic decision making as sales analytics.
2. *Challenges*: Traditionally, OLAP and OLTP workloads are performed on the same database system. However, workloads of OLAP mostly consist of bulk reads on data only that is updated by OLTP, constantly. Therefore, transaction-processing performance may be unexpected due to competition for a resource when both workloads are performed in a single database. Thus, it is necessary to separate workloads from OLAP and OLTP. **Figure 4-a** describes ETL-based background information where OLAP and OLTP are separated.





**Figure 4.**  
Classification of database systems.

In this architecture, each workload of OLAP must wait until the data in the data pool is completely refreshed and visible which will cause delays. Today, for reducing the delay, BI operating systems execute OLTP and short-term analytical queries together on the DBMS, as shown in **Figure 4-b**. These workloads are called short OLAP workloads. However, long-term OLAP workloads may be conflicted with many short OLTP transactions that make changes to the database. So, high synchronization is needed to deal with resource competition, which produces lower utilization of all resources.

Also, the commercial database management system (DBMS) uses special techniques as shadow copy [17], for handling mixed workloads with lower overheads. That is, on different logical versions of the data, different workloads will be separated and performed. Therefore, the additional space may be increased, which increases the infrastructure costs and requirements. Therefore, in current disk-based DBMSs a major challenge is managing these mixed workloads (OLAP and OLTP) [18].

### 3.3 Current BI systems

1. *Extended systems of traditional BI*: Current traditional BI technologies can perform OLAP queries and OLTP transactions on the same database without interfering with each other. Combining these mixed workloads with the same system needs extreme performance improvements due to the huge explosion in dynamic data size.
- “In-memory database (IMDB)”: Today, in most BI systems, OLAP and OLTP mixed workload on a one system can be handled using an In-Memory database (IMDB) (also, called Master-Memory). This technique needs that the system stores all data in the main memory, because it is faster than the optimized databases on disk and the internal optimization algorithms use less CPU instructions and are simpler. In case of querying data, this technique provides more predictable and faster disk performance by reducing the time of search. However, the IMDB systems can lack durability because of stored information losing when the device is reset or loses power. Many IMDB systems have proposed various mechanisms for supporting durability such as snapshots, non-volatile DIMM, non-volatile RAM, transaction logging, and high availability.

System	Type	Methods	Achievements
H-store [20]	IMDB	Distributed, row-store technique	High OLTP throughput rate
Radu Stoica [21]	Hybrid	Data reorganization	High performance, reduce paging I/O, and improve memory hit rate
Siberia [22]	Hybrid	Cold data access and migration mechanisms	Acceptable access rates with 7–14% throughput loss

**Table 1.**  
*Systems of modern BI that use various methods to hold most or all the data in the main memory.*

**Table 1** shows systems of modern BI that use various methods to hold most or all the data in the main memory for obtaining high OLTP productivity. For example, a distributed set of shared devices is used to run H-Store system, where the data is completely located in the main memory. The H-Store can execute transaction processing at high productivity rates, by removing traditional DBMS features as buffer management, lock and close. Recently, the H-Store prototype was marketed by a startup called VoltDB [19].

- “Hybrids with on-disk database”: The main-memory has become big enough for handling most OLTP databases, nevertheless this may not constantly be the best choice. For OLTP workloads by using access patterns, where some records are “cool” (rarely or not accessed at all), others, “hot” (accessed frequently). So, the coldest records are stored on fast secondary storage devices in the modern systems to ensure good performance. For example, Stoica and Ailamaki [19] suggested a way to migrate primary memory DB data to cheaper and larger secondary storage. In [20], for improving major memory heart rates and reducing I/O operating system migration, relational data structures are reorganized using access statistics for workloads of OLTP. Recently, Siberia was introduced as a cold data management framework in Microsoft Hekaton IMDB [21]. Like [19], it does not require storing an entire database in the main memory.

Hekaton focuses on how records are migrated to and from a cold store and how records are accessed and updated in a cold store in a consistent manner for transactions. So, only some tables can be declared and managed in the main memory by Hekaton. Experience evaluation shows that when cold storage is located on commodity ash, Siberia can lead to an appropriate productivity loss of 7–14%, given that cold data access rates are for an improved main memory DB.

2. *Modern features of BI Systems:* There are three modern information survey indicators: operational biological investigation, situational temporary survey, and self-service self-examination. Whereas, the H-Store system is only for OLTP transaction processing, a modern system called HyPer can handle it mixed workloads of both OLTP and OLAP are extremely high throughput rates using a low-overhead mechanism to create differential shots [22]. This system is used an unlocked approach which allows all OLTP transactions to be carried out in sequence or on special sections. In parallel with OLTP processing, HyPer system performs OLAP queries on the same shot and consistent.

Castellanos et al. [23] proposed a new platform called to notify business managers to situations that could affect their business. SIE-OBI integrates the functions

required for exploiting relevant rapid flow information from the web. They proposed new schemes for extracting and linking information that obtained from the web with the stored historical data in the data warehouse to reveal position patterns. The relevant information is extracted only from two or more different unstructured data sources, usually one stream of internal slow text and stream of external fast text. This time and effort minimization platform were built to build slow and fast data streams that integrate structured and disorganized flows, and to analyze them in almost real time.

### 3.4 Data governance

1. *Background:* in DAMA I [24] data governance is defined as “the exercise of authority and control over the management of data assets, planning, supervision and control over the management and use of data”. Data governance describes the responsibilities and roles of the organization in promoting desired behavior in the use of data [25]. Data management differs from data management, which involves setting data quality standards, making decisions and implementing them [26]. It is also different from BI Governance, which aims to provide a dedicated decision-making framework through the governance of all activities within the BI environment [27]. DAMA I [28] identifies 10 data management functions as shown in **Figure 5**. The data management function is high-level supervision, planning, and control of all other functions. There are four data management functions related to the next generation of biological information that requires fast access to data, external data utilization, and analyzing data by users, generally. Data architecture management includes setting of data standards, maintaining, and developing structures of enterprise data and linking application projects and architecture. The department of data quality focuses on planning, implementing, and controlling activities that apply techniques of quality management to measure, evaluate, improve and ensure the use of data. Data storage and business intelligence management focus on providing decision support data for reporting, query and analysis. Metadata management focuses on activities to enable easy access to high-quality metadata, such as architecture, integration, control, and delivery.

2. *Deploying Next Generation BI in Data Governance:* Data management has become vital for the organization as the data becomes inherent. The business derives its business value and decides based on the information derived from the data. Consequently, data control is required to ensure the quality of the data that directly affects the quality of the decisions made by the organization [29]. More effective data governance (DG) can lead to a higher scale of decision-making. To achieve effective data governance, maturity models of enterprise data governance help to understand DG and to determine what the next expected plan is [30]. Many data management maturity models [31] have been proposed for directing an organization to understand what data management level is. In [29], Oracle anticipated that the maturity model of data governance would help the organization in locating it in its data governance system evolution, identifying steps of short-term needed to reach the next level, and enhancing capabilities of data management. In the Oracle model, the highest level of maturity is the integration of data management with BI.

The next generation of BI supports almost insights of real-time with using of external information that generates a large data amount and its manipulations. So, this requires very mature DG for providing data quality, reliability, and integrity.

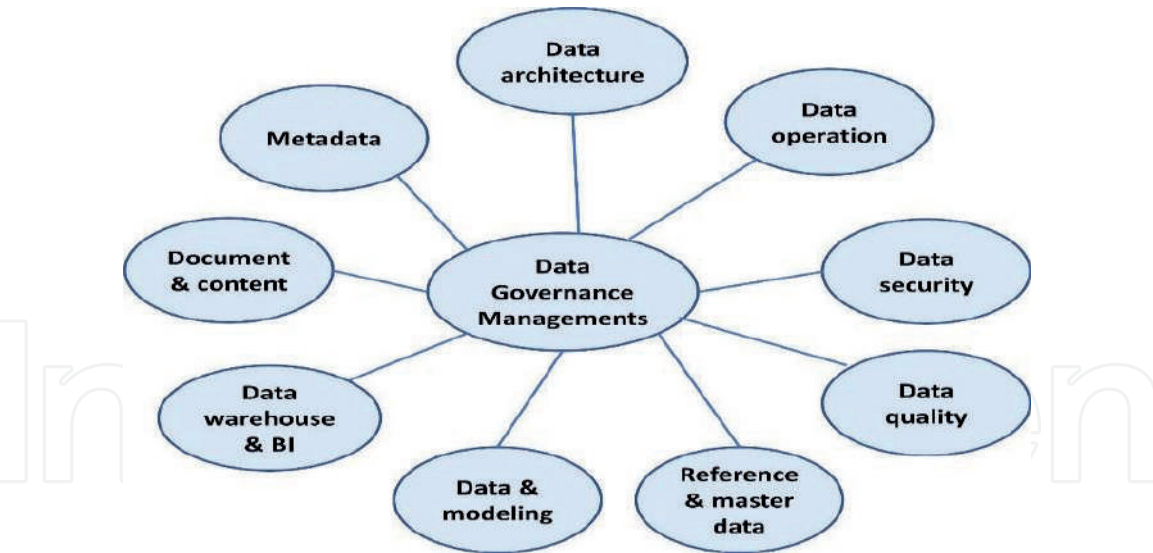


Figure 5.  
Framework of data governance as defined in Dama I [27].

The three characteristics are crucial to extract accurate insight through techniques of data mining. For example, in a “self-service” BI (for example Tableau and QlikTech), allows users to discover insight from many data sources without modeling the data environment and implementing complex ETL operations, which is one of the most time-consuming and difficult tasks in BI. So, these new features allow users to easily access data, get quick results and visual data visualization. To enable the evolution of the next generation of biological information, data management is critical to the reliability of data from the discovered vision. For example, in the case of BI self-service, the fact that end users can access and process their data reduces the reliability of BI results [32]. In data management, useful functions to ensure reliability can be considered such as tracking data ratios to source and creating records of how data is processed or transferred. However, integrating data governance into the next generation of biological information has faced some challenges due to the requirements of flexible and reliable responses while there is an enormous amount of external data and public user engagement.

3. *Data governance challenges:* There are two main advantages to the next generation of biological information that affects the data management model. Decision making in the next generation BI, should be more effective and faster between a huge amount of data that comes from many data formats and sources. However, data from many sources makes data management more difficult to manage and sophisticated to control properly. This can also lead to ineffective decisions being made. In case of data comes from different conflicting sources, more research and analysis of the data and the different sources of that data to determine what is true and accurate or its approximation must be done by the decision-maker, which will be costly operations. Therefore, management of data across heterogeneous sources in the next generation BI system is very important. In the next generation of personal information, especially “self-service” business users participate in procedures of decision-making.

In general, the central IT organization and many data supervisors have been involved in data management initiatives and have a metadata repository for the data management platform and a set of data management tools to deal with varied data. In advance, they standardize common data definitions of master data and reference data that are widely shared across many enterprise applications. When they receive



disparate data, they match it to define predefined shared data, determine its quality, determine which rules, convert, and merge them. However, in the next generation of BI, users also select, manipulate or merge their data names themselves using various “self-service” tools of BI. They may want to upload to the DB and share their vision with others. Participation of business user in the data process can lead to data in a mess where the same data can be converted and combined in various ways through data managers and a central organization using tools of data management and by business users who have tools of BI for “self-service”. Consequently, meta-data sharing criteria are crucial through this sharing to transfer shared data, shared data names, and shared integration rules [33].

4. *Data governance model for next generation BI*: The data management model design is designed to centralize versus. Decentralization and hierarchy versus cooperative. Central design assigns all decision-making authority in the central IT department while decentralized design assigns authority to individual business units [25].

The term big data is a group of huge and complex data sets from various sources where data the management and traditional application processing techniques face difficulties to process it. Big data is a collection of a large amount of structured or unstructured data that is processed and analyzed for informed decision-making or evaluation. These data can be taken from various sources including browsing history, geographic location, social media, medical records, and purchasing record. Big data is made up of complicated data that will smash the processing power of traditional simple database systems [34]. In [35], the authors mentioned that, there are three main characteristics associated with big data: (1) Volume is a feature used for describing the vast data amounts that big data uses. Usually, the range of data amounts starts from GB to YouTube. Big data should be able to handle any data amount even with its highly anticipated growth. (2) Variety is a feature used for describing various types of data sources that are used as portion of a large data analytics system. Currently, there are many data storage formats used by computers all over the world. One format is the structured data such as databases and. Csv, video, short message service (SMS) and excel papers. Unorganized data can be in the handwritten notes form. All data from these sources will be ideally used for Big Data Analytics. (3) Velocity is a feature used to describe the speed at which data is generated. It is also used to describe the speed at which generated data is processed. With the click of a button, an online retailer can quickly view big data about a specific customer. Speed is also important to ensure that data is updated and updated in real time, allowing the system to perform at its best. This speed is necessary as real-time data generation helps organizations accelerate operations. Which can save institutions a large amount of money.

Today, many companies are increasingly interested in using technologies of big data to support their BI, so that it becomes very important to understand the different practical issues from previous experiences in BI systems. Today’s BI systems sense the world and harness these data points for recommending the best possible options and forecast results, accurately. As BI systems continue to be built in real time, the demand for data collection, integration, processing, and visualization increases almost in real time. BI systems are characterized by high sensitivity opportunities as seen in sensors with the rich diversity of sensors ranging from mobile phones, personal computers and health tracking devices to technologies of Internet of Things (IoT) designed to give contextual and semantic sound to entities that could not previously contribute Intelligent in key decisions. So, many companies are analyzing big data today.

Big data analytics is needed and is machine learning techniques because of often distributed data sets, and its privacy and size considerations are evidence of distribution techniques, where data is on platforms with different computing capabilities and networks. The benefits of application diversity and big data analytics pose challenges. As an example, every hour the servers of Walmart handle more than million transactions for a customer, and this information is stored into databases that contain larger than 2.5 petabytes of data, which is 167 times the number of books in the Library of Congress. Herein, CERN's Collider Hadron Collider produces around 15 petabytes of data annually, and that is enough to fill over 1.7 million double layer DVD discs annually [36]. Big data analytics are used for education, health care, media, insurance, manufacturing and government. Big data analyzes of business intelligence and decision support systems that enable healthcare organizations to analyze data size, diversity and tremendous speed have been developed across a wide range of healthcare networks to support evidence-based decision-making and action [37]. Therefore, it is clear from the discussion that data management and big data analytics [38] are important in BI for 4 reasons:

1. *Better-decision-making (BDM)*: Big data analytics can analyze current and old data for making predictions about the future. So, companies can make not only better current decisions, but also preparing for the future.
2. *Cost reduction (CR)*: Big data technologies like cloud-based analytics and Hadoop offer great cost advantages when storing large data amount. In addition, he provided insights on the effect of various variables.
3. *New products and services (NPS)*: With the ability to measure needs and satisfaction of customers through analytics, the strength comes for giving customers what they want. So, more companies are creating new products and services to meet customer needs.
4. *Understand the market conditions (UMC)*: By analyzing big data, we can get a better understanding of current market conditions for retrieving important information. In addition, there are a few features and challenges that must be considered in the tools and techniques of big data analytics, and they include scalability and fault tolerance as well [39–41]. The following **Table 1** represents a few of the widely used tools with the advantages of Big Data Analytics.

The rapid development of business intelligence and analysis attracted the attention of researchers. The reason is that organizations no longer rely on traditional technologies as data grows exponentially. This huge amount of data requires advanced analytical techniques in order to convert it into valuable information that helps organizational growth. BI&A is the contemporary methodology for extracting value from this vast amount of data, driving strategic decision-making, and forecasting and benefiting from future opportunities.

BI&A is necessary in most organizations. BI&A has proven effective support in decision making. In addition to that data and IT infrastructure is clearly influenced by the good use of BI&A practices. Nowadays, business intelligence and analysis have played a vital role in most institutions and sectors due to their value and benefits. BI&A helps organizations gain a better view of their private data and thus improves fact-based decision-making. These methodologies and data analysis also help to maintain competitive advantage in addition to resolving technical and quality problems that will enhance the performance and productivity of enterprises [42, 43].

According to Abai et al. [44] BI&A helps to build an integrated framework that supports speeding up organizational performance. Many factors and technological developments have shaped the past and present trends of BI&A. With the rapid development of technology, it is not enough to use traditional analytical techniques. The future direction of business intelligence and analysis will expand to include areas of diversity. According to Chen et al. [45]. The success opportunities associated with data analysis technologies have generated future interest in business intelligence and analytics. Additionally, BI&A contains different practices and methodologies that can be applied to different sectors; Health care, security, market intelligence, e-government, and others. According to Mohammed and Westbury [46] BI&A is contributing to future development systems. By mapping all the facts, BI&A has become soon biotechnology in developing cities by supporting real-time information that will turn countries into smart cities.

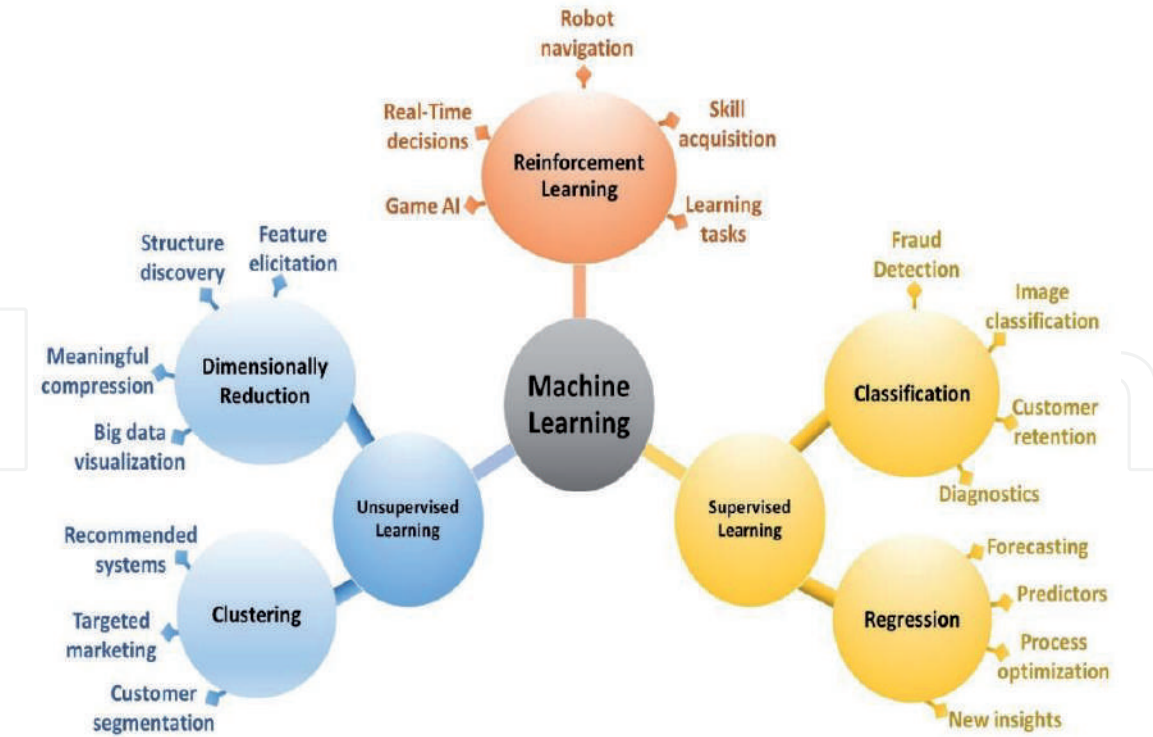
One of the most important responsibilities in the data mining process is choosing the appropriate data extraction technology. The nature of work and the type of object or difficulty experienced by the work provides appropriate guidance for identifying the best techniques [47]. Application of data mining techniques There are some generalized approaches that can indicate enhanced efficiency and cost-effectiveness. Many of the basic techniques that are performed in the data mining process, determine the nature of the mining process and the option of data recovery.

Artificial intelligence (AI) represents a step in the evolution of technology that has been actively pursued since the British mathematician and code-breaking Alan Turing was conceived as a clear way forward in his pioneering research of 1950, "Computing and Intelligence." At the time, computer technology could not keep up with Turing's ideas. But as computing advanced, Amnesty International advanced. Most of the artificial intelligence that we see today is narrow artificial intelligence (ANI), which means it can perform a well-defined task. A 2018 report by the Future of Humanity Institute at Oxford University has surveyed a group of AI researchers in the schedules of strong AI. She found "50% chance of artificial intelligence outperforming humans in all tasks in 45 years and automating all human functions in 120 years." However, AI will bring with it many opportunities to create new business opportunities as well. As many experts have pointed out, one of the great values of artificial intelligence is its ability to eliminate the need for strenuous and repetitive tasks. Alternatively, users can focus on their core values and skills. Technology was applied in many industries mostly aimed at reducing human error, reducing labor costs, and thus increasing profit. This was true of the progress made during the Industrial Revolution until the birth of the computer, and still true of the emergence of artificial intelligence.

Artificial intelligence has advanced significantly in the past few years due to a number of factors, starting with a massive increase in the computing power available. The once-trained AI model now takes days or even hours with machine learning (more on this soon). Another factor is wider data access. You may have heard that the data is "new oil" or something similar. However, the data must be processed using advanced tools such as analyzes and machine learning algorithms to reveal useful information. This processing is where the AI in BI becomes an invaluable tool.

Machine learning is the engine of artificial intelligence systems. It strengthens artificial intelligence models by analyzing complex data sets. Machine learning enhances models by analyzing complex data sets through a set of self-acquired rules and knowledge as shown in **Figure 6**. The machine learning model learns from big data and from frequent human interactions so that it can provide information and answers related to the user's interests or goals. Big data refer to very large data sets





**Figure 6.**  
*Machine learning model.*

that can be mathematically analyzed to reveal patterns, trends, and correlations, especially about human behavior and interactions. In the space of artificial intelligence, deep learning represents a major leap forward in technology. As we just touched on, programmers write a code that directs the device how to interpret a series of words, pictures, or commands to reach a decision and execute an order. The end user then introduces the entry (data), while internal engineers may define more specific rules for interpreting and analyzing that data. Finally, the system provides outputs (analysis) based on the specific inputs and defined rules. In [48], the authors proposed a demand-forecasting model by BI with machine learning.

**3.5 Why BI needs AI?**

Does it matter if the constant awareness of the original, or is the copy going to be alive anyway 19? For better or worse, the future comes faster than we realize. There will be no before or after artificial intelligence but a slow transition for a decade or more. As we have seen with Google Glass, it's currently impossible to guess what acceptable results would look like. But how much can we trust in our future assistants? Will they work with us or unknown entities? If we do not ask the right questions now, we'll get the default app. It will be free, but what will include small prints? Good morning John. Here's today's program. Any questions? Perhaps it does not matter after all: Using a good learning algorithm, the program will know what we need and what we need to do, better than we can ever guess. The power of statistics will win the war against the gods and we will lose our soul. It is known that job candidates can lose their chances in a decisive way when they think that no one is watching by bad behavior or rejection of reception staff and waiting staff. Once NLPs and other AIs are widespread, it will not be long before the same literature test is introduced. Looking at 2050, the future of humanity lies in the transition to a civilization of the first kind. We are type 0, extinct. We are about to become half-gods. Most likely, we will merge with our own processing technology and each of us will have our own virtual world to dominate it with absolute control in every aspect



of it, and the countless millions of planets of “life” that we may control or merge with as well. Just as video game programmers have absolute control over the worlds they create. Immortals, omniscient and omnipresent, are all capable of our universes. Of course, he can explore this universe as well, maybe contact directly with his creative being and know that we are characters in his game. Our last question will be morality and maturity. Will we only have one universe? Or does force drive us into madness and transform us into “invaders of the universe” and penetrate the universes of others, based on greed, against the desire for more force? Will we be good? Or evil? Or both? Will we be able to achieve wisdom, and secure peaceful and harmonious coexistence with all other demigods, or will we go to war? Or will we merge into one excessive force? Or are we tired one day from the divine and start the final game again, and transform ourselves into a universe that we will have to evolve for billions of years for us to be re-created one day? Maybe this is exactly what is happening.

### 3.6 Improving BI with AI

In this section, we explore how BI's AI raises and improves the way of an organization that are used for analysis and interpretation the lifeline of its business.

1. *Turning Business Users into Data Experts (TBUDE)*: Typically, business analysts (BA) and IT officials control the access to data and its interpretation. Although these occupations are crucial until now. With the AI tools in today's BI tools, including LOB, NLIs users no longer need to depend on data science experts for analyzing their data. AI allows users to obtain actionable answers easily and directly for helping “democratize” data. In other words, it gives users a two-way conversation ability with their data and feel empowered for acting with answers in a reliable way. Here is an example of how AI works in practice: a certain organization is deploying a BI solution that uses an advanced NLI and instead of waiting for system administrators or data scientists for analyzing data, the manager of business unit arrives at the BI solution directly. The manager makes the data available by calling or downloading and asks questions in simple language. Then, the user receives insight into these questions along with a dashboard and visuals ready for presentation to help communicate these answers. A pre-trained model of AI can target even specific tasks of BI such as visualization recommendations, scenarios of “what-if”, and prediction for helping managers to make important decisions for their business.
2. *Helping You Explore Your Data (HYEYD)*: There is something inherently satisfying to explore your data with the right tool of AI that supports artificial intelligence. In minutes, you can move from loading data sets to revealing hidden facts in the data and introducing these results in beautiful visualizations. At a starting moment the data is available, the AI in the system of BI heavy lifting by sorting automatically, marking columns and joining matching data across groups. Accessing the NLI is the first step in data exploration for the user. The AI tool will suggest questions that might be helpful if you get stuck. You can also start with the basics, like “How did the retail store department perform during the X period?” The AI will provide answers and suggest ways to explore data to get additional insights into performance. Exploration is exciting because you can continue to delve deeper into visions that only AI can achieve. What embodies the imagination of users is imagination. Visuals are an essential feature of all modern BI solutions, but with AI enabled AI solutions, users receive suggested, automated visualizations that best fit the answers to their questions.

3. *Learning from the End User (LFEU)*: The leading AI systems in BI systems are customized and improved all times through machine learning that indexes and learns traditional questions and behaviors of a user. The more user interacts with the tool of BI, the better the AI will know what this user wants in the presentation and analysis. If the user usually uses forecast data, the system will begin to prepare and present the data in the prediction model via dashboards.
4. *Automatically Cleansing and Prepping Data (ACPD)*: To successfully interpret it, your data must be organized in a unified and searchable manner. As any business knows well, multiple datasets cause multiple headaches. What if names are formatted as first name/last name in one spreadsheet, and last name/first name in another? What if there are duplicate records? What if there are records in one dataset and not the other? What if the data in one set is daily, and the other is monthly?  
AI in BI reduces data cleansing and contact preparation and provides massive aspirin for headaches. By setting up data automatically (one of the biggest artificial intelligence in saving time), you can move from making data available to working with it in minutes, instead of hours or days. The future AI function will allow users to enter structured and unstructured data without skipping any win; A big change since most of the data being created today - such as photos, videos and audio - is disorganized. Removing barriers to effective analysis is one of the ways in which the advanced AI in BI tool helps users who are not data scientists to access and interpret their data.
5. *Gaining Competitive Advantage (GCA)*: AI now makes a critical difference between the companies that enable it to succeed and those that will be left behind soon. Gartner predicts that by 2021, 75% of pre-prepared reports - such as those used to extract data - will be either replaced or strengthened using automated insights. The robust AI in BI tools also provides improved accuracy for critical operational use reporting. If they do not, the data and analytics leaders should plan to adopt Enhanced Analyzes (AI) immediately in their business as the capabilities of the platform mature. Rita Sallam, vice president of Gartner Research, warned at a recent conference that “data and analytics leaders should examine the potential impact of business” from increasing reliance on predictions using enhanced and automated insights “and adjusting business and business models accordingly, or risking losing the competitive advantage of Those who do. “ AI are already offered in BI solutions today, and those companies that adopt technology are poised to succeed more safely than those that do not. By uncovering trends and correlations in data and proposing ways to interpret results in natural language along with providing the best coordination for presenting these results, AI saves time and provides actionable insights to increase profitability and avoid potential problems before they arise.

## 4. Conclusion

In this chapter, the traditional and Modern BI were reviewed in detail which became a critical and important process and received a great interest in both industry and academia fields. So, the data management, data mining and machine learning techniques are needed for extracting the insights from big data. By using such techniques, business intelligence gets better decision making, cost reduction, new products and services and understand the market conditions. In addition, the importance of big data analytics, data mining, AI for building modern BI and

enhancing were introduced and discussed. Also, challenges and opportunities for creating value of data by establishing modern BI processes were described and how AI raises and improves the way an organization analyzes and interprets the lifeline of its business are explored. In the future work, we will study more AI tools to enhance the processes of BI and solving the cybersecurity problems in modern BI.

## **Acknowledgements**

This research was supported by the Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt. In addition, it was partially supported by King Abdul-Aziz University, Jeddah, Saudi Arabia. I thank both for providing guidance to finish this research. I also thank IntechOpen Limited for giving the opportunity for publishing this research work as a book chapter in E-Business.

## **Author details**

Ahmed A.A. Gad-Elrab<sup>1,2\*</sup>

<sup>1</sup> King Abdul-Aziz University, Jeddah, Saudi Arabia

<sup>2</sup> Faculty of Science, Al-Azhar University, Cairo, Egypt

\*Address all correspondence to: [asaadgad@azhar.edu.eg](mailto:asaadgad@azhar.edu.eg)

## **IntechOpen**

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Wael M.S. Yafooz Abidin, S. Z., & Omar, N. (2011, November). Challenges and issues on online news management. In *Control System, Computing and Engineering (ICCSCE)*, 2011 IEEE International Conference on (pp. 482-487). IEEE.
- [2] <https://technative.io/unstructured-data-the-hidden-threat-in-digital-business/>
- [3] Balachandran, B. M., & Prasad, S. (2017). Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence. *Procedia Computer Science*, 112, 1112-1122.
- [4] Kimble, C. and Milolidakis, G. Big Data and Business Intelligence: Debunking the Myths. *Global Business and Organizational Excellence*. 35, (2015), 23 – 34.
- [5] Richards, G., Yeoh, W., Chong, A. Y. L., & Popovic, A. (2017). Business intelligence effectiveness and corporate performance management an empirical analysis. *Journal of Computer Information Systems*, 1-9.
- [6] Xia B.S. and Gong P. Review of business intelligence through data analysis. *Benchmarking: An International Journal*. 21, (2014), 300-311.
- [7] Kowalczyk M. and Buxmann P. (2014). Big Data and Information Processing in Organizational Decision Processes: A Multiple Case Study. *Business & Information Systems Engineering*. 5, (2014), 267-278.
- [8] Wael M.S. Yafooz Abidin, S. Z., & Omar, N. (2011, November). Challenges and issues on online news management. In *Control System, Computing and Engineering (ICCSCE)*, 2011 IEEE International Conference on (pp. 482-487). IEEE.
- [9] Siddiqua, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, 71, 151-166.
- [10] Fahad, S. A., & Alam, M. M. (2016). A modified K-means algorithm for big data clustering. *International Journal of Computer Science Engineering and Technology*, 6(4), 129-132.
- [11] Fahad, S. A., & Yafooz, W. M. (2017). Design and Develop Semantic Textual Document Clustering Model. *Journal of Computer Science and Information Technology*, 5(2), 26-39. doi:10.15640/jcsit.v5n2a4 .
- [12] Thuraisingham, B. (2014). *Data mining technologies, techniques, tools, and trends*. CRC press.
- [13] A. L"oser, F. Hueske, and V. Markl, "Situational business intelligence," in *Business Intelligence for the Real-Time Enterprise*. Springer, 2009, pp. 1-11.
- [14] Deloitte report, "Modern Business Intelligence: The Path to Big Data Analytics", April 2018.
- [15] A. Lser, F. Hueske, and V. Markl, "Situational business intelligence," in *Business Intelligence for the Real-Time Enterprise*, ser. Lecture Notes in Business Information Processing, M. Castellanos, U. Dayal, and T. Sellis, Eds. Springer Berlin Heidelberg, 2009, vol. 27, pp. 1-11. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-03422-0\\_1](http://dx.doi.org/10.1007/978-3-642-03422-0_1)
- [16] M. Castellanos, C. Gupta, S. Wang, and U. Dayal, "Leveraging web streams for contractual situational awareness in operational bi," in *Proceedings of the*



- 2010 EDBT/ICDT Workshops, ser. EDBT '10. New York, NY, USA: ACM, 2010, pp. 7:1-7:8. [Online]. Available: <http://doi.acm.org/10.1145/1754239.1754248>
- [17] R. Elmasri and S. B. Navathe, *Fundamentals of database systems*. Pearson, 2014.
- [18] H. Kuno, U. Dayal, J. Wiener, K. Wilkinson, A. Ganapathi, and S. Krompass, "Managing dynamic mixed workloads for operational business intelligence," in *Databases in Networked Information Systems*, ser. Lecture Notes in Computer Science, S. Kikuchi, S. Sachdeva, and S. Bhalla, Eds. Springer Berlin Heidelberg, 2010, vol. 5999, pp. 11-26. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-12038-1\\_2](http://dx.doi.org/10.1007/978-3-642-12038-1_2)
- [19] "Voltdb," <https://voltdb.com/>.
- [20] R. Stoica and A. Ailamaki, "Enabling efficient paging for main memory OLTP databases," in *Proceedings of the Ninth International Workshop on Data Management on New Hardware*, ser. DaMoN '13. New York, NY, USA: ACM, 2013, pp. 7:1-7:7. [Online]. Available: <http://doi.acm.org/10.1145/2485278.2485285>
- [21] A. Eldawy, J. Levandoski, and P.-A. Larson, "Trekking through siberia: Managing cold data in a memory-optimized database," *Proc. VLDB Endow.*, vol. 7, no. 11, pp. 931-942, Jul. 2014. [Online]. Available: <http://dx.doi.org/10.14778/2732967.2732968>
- [22] A. Kemper and T. Neumann, "Hyper: A hybrid OLTP & OLAP main memory database system based on virtual memory snapshots," in *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ser. ICDE '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 195-206. [Online]. Available: <http://dx.doi.org/10.1109/ICDE.2011.5767867>
- [23] M. Castellanos, C. Gupta, S. Wang, U. Dayal, and M. Durazo, "A platform for situational awareness in operational {BI}," *Decision Support Systems*, vol. 52, no. 4, pp. 869 – 883, 2012, 1) *Decision Support Systems for Logistics and Supply Chain Management* 2) *Business Intelligence and the Web*. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S016792361100217X>
- [24] D. M. International, *The DAMA guide to the data management body of knowledge*. Technics Publications, Bradley Beach, 2009.
- [25] K. Weber, B. Otto, and H. "Osterle, "One size does not fit all— a contingency approach to data governance," *Journal of Data and Information Quality (JDIQ)*, vol. 1, no. 1, p. 4, 2009.
- [26] V. Khatri and C. V. Brown, "Designing data governance," *Communications of the ACM*, vol. 53, no. 1, pp. 148-152, 2010.
- [27] R. S. Seiner, "Real-world data governance bi governance and the governance of bi data," <http://www.slideshare.net/Dataiversity/realworlddata-governance-bi-governance-and-the-governance-of-bi-data14889552> Accessed:2015-11.
- [28] D. M. Association et al., "Dama dmbok functional framework (version 3.02)," DAMA International, 2008.
- [29] NASCIO, "Data governance - managing information as an enterprise asset part 1 - an introduction," *NASCIO Governance Series*, 2009.
- [30] —, "Data governance part iii: Frameworks - structure for organizing complexity," *NASCIO Governance Series*, 2009.
- [31] P. Aiken, M. D. Allen, B. Parker, and A. Mattia, "Measuring data

- management practice maturity: a community's self-assessment," *Computer*, vol. 40, no. 4, pp. 42-50, 2007.
- [32] B. Potter and R. Software, "Self-service bi vs. data governance," <https://tdwi.org/articles/2015/03/17/self-service-bi-vs-datagovernance.aspx>, Mar. 17. 2015.
- [33] M. Ferguson, "Is self-service bi going to drive a truck though enterprise data governance?" <http://intelligentbusiness.biz/wordpress/?p=489> Accessed:2015-10.
- [34] Hung, P. C. K. *Big data applications and use cases, the springer international series on applications and trends in computer science*. Switzerland: Springer International Publishing AG, 2016
- [35] Dave, P. What is big data - 3 vs of big data. Retrieved from SQL Authority, 2013 Blog: <http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-varietyday-2-of-21/>.
- [36] Sparks, B. H., & McCann, J. T. (2015). Factors influencing business intelligence system use in decision making and organisational performance. *International Journal of Sustainable Strategic Management*, 5(1), 31-54.
- [37] Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3-13
- [38] Wael M.S. Yafooz., Abidin, S. Z., Omar, N., & Idrus, Z. (2013, December). Managing unstructured data in relational databases. In *Systems, Process & Control (ICSPC)*, 2013 IEEE Conference on (pp. 198-203). IEEE.
- [39] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In *System sciences (HICSS)*, 2013 46th Hawaii international conference on (pp. 995-1004). IEEE.
- [40] Zhou, Z. H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. *IEEE Computational Intelligence Magazine*, 9(4), 62-74.
- [41] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286.
- [42] Lautenbach, P., Johnston, K. and Adeniran-Ogundipe, T.. Factors influencing bussiness intelligence and analytics usage extent in south african organaisations. *S.Afr.J.Bus.Manage*, 48(3): 23-33, 2017.
- [43] Wang, Te-Wei,et al. "Depicting Data Quality Issues in Business Intelligence Environment Through a Metadata Framework." *Applying Business Intelligence Initiatives in Healthcare and Organizational Settings*, edited by Shah J. Miah and William Yeoh, IGI Global, 2019, pp. 291-304. <http://doi:10.4018/978-1-5225-5718-0.ch016>.
- [44] Abai, N. H., Yahaya, J. and Deraman, A. "An integrated framework of business intelligence and analytic with performance management system. A conceptual framework." In *Proceedings of the 2015 Science and Information Conference* . London. pp. 452-56, 2015.
- [45] Chen, H., Chiang, R. H. and Storey, V. C. Business intelligence and analytics. *From Big Data To Big Impact*, 36(4): 1165 – 1188, 2012.

[46] Mohammed, J. and Westbury, O. Business intelligence and analytics evolution, applications, and emerging research areas. *International Journal of Engineering Science and Innovative Technology (IJESIT)*, 4(2): 193-200, 2015.

[47] M. A. Khan *et al.*, "Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning," in *IEEE Access*, vol. 8, pp. 116013-116023, 2020, doi: 10.1109/ACCESS.2020.3003790.

[48] Fahad, S. A., & Alam, M. M. A modified K-means algorithm for big data clustering. *International Journal of Computer Science Engineering and Technology*, 6(4), 129-132, 2016.