Advances in Economics, Business and Management Research, volume 215
Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)

How Accurate are Predictions Made Using Big Data?

Haohao Su

Kings College Private High School, 1200 Leland Rd, Oshawa, ON L1K 2H5, Canada
Email: 3180887843@qq.com

ABSTRACT
Since the home computer phenomenally occurred in the 1980s, the methodology of information collecting has had a revolutionary improvement. There was a rapid increase in the number of sources of electronic information that are available to be collected. Big data is one of the consequences of these developments. Currently, it is widespreading and takes increasing roles, such as analysis and predicting. However, it is doubtful whether big data analysis is accurate or not. Therefore, this paper will concentrate on the accuracy of big data analysis and prediction with the following principles: the definition of big data, the standard of accurate big data, limitation of big data analysis, and possible future development.

Keywords: *Big data, data analysis, big data applications*

1. INTRODUCTION

First of all, it is critical to know the general meaning of big data and how people use it. D. Goldston gave the first definition of big data, it is a set of data with an innumerable number of data that is currently not available to use [1]. Though this explanation of big data might be thought of as out of date now, it is pretty interesting that as technology develops, it is achievable for humans to use big data. Also, it led to a change in the definition of big data, hence, scholars invented a complex and realistic system to maximize the benefit of big data, including collecting, storage, processing and analyzing and visualization [2]. The current big data systems could be seen in Fig. 1. In comparison, the old use of big data is shown in Fig. 2 [2]. It is obliviously that this breakthrough of technology has brought people many advantages in recent years. These benefits of big data have positive influences on multiple areas, such as the everyday life area and business area [3]. Furthermore, more landscapes attempt to expand with big data analysis, for example, the innovation areas and agriculture and artificial intelligence areas [4]. Big data prediction is one of the fields that benefited from the recent advancements in technology and is also the main topic that is focused on in this paper. Nevertheless, it is doubtful that whether the current big data analysis is available or not, and some scholars have given their answers on it, it is not always the best way [5]. Many areas have already suffered from the drawbacks of the current big data analysis system, for instance, the healthcare area meets difficulties that the current big data analysis system lacks details [6]. In big data prediction, the significant consequence of these drawbacks will lead to an inaccurate result. The course of the wrong big data prediction always account for three areas, big data system, customer and big data itself, and many of them were mentioned by Danah Boyd [7].

Fig. 1. The current running mode of big data analysis.

Fig. 2. The previous use of big data analysis.

After knowing the limitation of big data, scholars did a lot of research to improve, from the system, theory, or even the material of computer used [8]. These studies made a lot of achievements and appealed to people about the sustainable development of big data. With those study, it can be seen that big data have a bright future, and big data will contribute more as it grows. Nonetheless, there are also noticeable issues that may restrict the development of big data analysis, for example, the privacy problem [9]. There is a universal concern on the future of big data with these arguments, and scholars have

already given their answers like Deepa, N, and I will also provide my answer at the end of this paper [3].

In this paper, the current accuracy of big data analysis will be mentioned in section 2, and in the section 3, the possible factors lead big data analysis to less accurate will be discussed. In the section 4, the possible improving way will be suggested.

Table 1. Comparison of the accuracy of big data in different

Works	Accuracy	Area	Difficulty	Accurate or not
Ji et al. [10]	not given	Shop Floor Scheduling	lack of capacity high cost highly complex health problem lace of sample collection,techniques,infrastructure	inaccurate
Sahoo et al. [11]	0.98	Healthcare		accurate
Hekler et al. [12]	not given	Precise Medicine		inaccurate
Ghasemaghaei et al. [5]	not given	Business		inaccurate
Bhat et al. [4]	not given	Digital Agriculture		accurate

2.THE CURRENT SITUATION OF BIG DATA ANALYSIS

In 2015, IDC had predicted that big data would reach an annual spend of 48.6 billion dollars in the following four years [13]. Similarly, there is a prediction that the big data market will grow to 229.4 billion dollars in 2025 [3]. Although it is not guaranteed that will these forecasts come true or not, it can be seen that people have strong confidence in big data. The basement of this expectation is majorly from the unique performance made by big data analysis. The contribution of big data on weather forecasting area could reveal the edge of its influence. Big data analysis is widely used as a tool to overcome the challenge brought by climate change [14]. As Fathi mentioned, this single contribution to the weather forecasting area will provide people chain benefits in relevant industries like tourism and transportation [15].

Furthermore, the growth of big data analysis technology is also one reason for the high expectation of the big data market. In 2013, the paper of Mohanty, S. had shown the simple usefulness of big data at that time [16]. Comparing a recent article about big data analysis, it is evident that big data has a board applicant in other areas [2]. This broad use should be one reason why many enterprises are fascinated by it. From these cases, it is allowed to see an edge of the benefits of big data prediction.

Macomb provided a theory called 3Vs, and it can be the basement for understanding the advantages of big data and the source of the high expectation of big data [13]. Big data has characteristics of large scale, fastupdating and diversity, also known as volume, velocity and variety in Macomb's approach. In short, the features of big data in scale provide enough sample for the prediction and the fastupdating could avoid the possible error account of the change of external environment, and the variety of big data could allow a further expansion of the use of big data. These may explain why people have high expectations of big data.

Additionally, a detailed introduction of the advantages of using big data prediction is mentioned by R Northcott [17]. Still, in short, it is undoubted that an accurate big data prediction could significantly improve the quality of people's life [17]. Based on that, it could be an excellent standard for us to define the accuracy of big data prediction, an ideal big data prediction should fit the requirement of the public's use and make people's life more convenient.

Nevertheless, it is doubtful that is big data prediction acurate or not. Based on this standard, it is possible to make a brief definition of the accuracy of big data prediction. As the PK Sahoo proved, big data prediction has an incredible accuracy of 98 percent in health care [11]. This could seem evident that the current big data prediction is accurate. It is suggested to use big data analysis prudently [18]. Conversely, some articles show that big data prediction is not strong enough to fit all the demands [10]. The accuracy of big data analysis in different areas can be seen in the Table 1. Macomb also briefly introduced the challenge that big data faced and identified the possible future difficulties [13]. These problems of inaccurate prediction are inevitable to meet the expectation of the market, therefore, it is essential to overcome the limitation of big data prediction. And before, that it is unavoidable to talk about what makes it not so accurate.

3.THE POSSIBLE REASONS FOR INACCURATE

Concerns abound, one of which is big data overestimated by people. Amazingly, big data have such rapid growth, and some scholars doubt the possibility of hype [19]. This concern is not a press without proof, evidence shows that the big data is not as ideal as people imagine [5]. The worried may come from three areas mentioned before, big data, systems and customers.

Firstly, the prediction made by big data may be wrong sometimes account of the limitation of big data. It can be observed that the usefulness of big data is related to the quality of data. M. Ghasemaghaei proved that the big data would be less beneficial when it fails to meet the requirement of the 3Vs, the lack of every characteristic will lead to a significant fall in the accuracy of the end of the prediction[5]. The possible influence of lack of any

3Vs could be observed in the Fig.3 and Fig.4 from Maryam's research [5]. The forecast result is also less reliable, therefore, Ghasemaghaei concluded that big data is usually good, but not all the time.

An early work hinted that there are differences between users, and it is not guaranteed whether the big data analysis will be less functional. In contrast, the users participated less in the application [20]. However, it can be sure that the amount of data collected will have a significant fall, assuming that most users are all 'listener' as Crawford described that. Also, evidence shows that some 'bots accounts' occur in social media [21]. This may provide an unreliable source of data. As a consequence, it will lead to a wrong prediction.

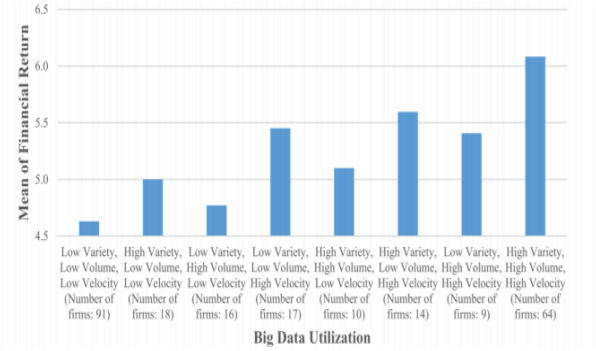

Fig. 3. The influence on usefulness in fiance areas when lack of 3Vs

Another argument that relates to big data could be the vague boundaries of morals. It has been a pushing demand for the legal definition of the privacy policy as the Internet was introduced. Some users may not want the enterprise to collect their information for business use, however, running a big data system is inevitable. It nowadays turns to an ethical issue about the boundaries of enterprise's authority and government [22]. Due to this, the users may try multiple ways to avoid the enterprise to draw their profile and attempts they made, including creating more than one account and switching off the 'cookies'. Though these methods have an achievement in protecting the users' rights, it is no doubt that these solutions are harmful to the accuracy of big data.

What's more, big data itself also limits the accuracy of big data prediction. Big data became widespread as the Internet became universal, whereas, Internet restricted big data in the source of finance. The information collected from the Internet will be easy to meet the need of volume, however, the data will have low variability. Although now it's hard to observe the influence brought from this limitation, it will no doubt restrict the upper limit of big data prediction.

The nature limitation of data also restrict the development of big data, an article had mentioned that companies would only collect part of the data instead of the 'whole data' [7]. The consumers of the enterprise limited the data it collects, therefore, the data might be partial and lead to a wrong result. It is almost impossible for enterprises to exchange each others' data to overcome this difficulty, therefore, it becomes a tricky question for companies, and it is hard to develop a solution about it.

Nonetheless, big data has these limitations, it still meets enterprises' needs and contributes to many industries. However, it is inevitable to overcome all these difficulties so that the big data can reach people's expectations of the 229.4 billion dollars market [3]. It is believable that big data will have a bright future account of the steadily breakthrough in the big data area.

4.POSSIBLE IMPROVING

Recently, scholars gave their ideas for improving big data analysis, which are creative and groundbreaking. For instance, Hekler introduced a 'small data' concept to help the big data analysis [12]. This article hints that people could combine big data analysis and the traditional qualitative method, which has a smaller scale but is highly precise, to improve the accuracy of big data prediction. It is a loss to focus on big data only and miss the small data.

Besides the systematical improvement, some advances are also made in other areas, for example, the computer material can also make big data prediction more accurate. Jablonka revealed that the evolution of the material could provide a larger scale of data collected, therefore, the volume of big data will increase, and the result will be more accurate as of the sample grows [8].

Also, for the problems mentioned before, scholars provide many solutions. For the issues of the multi-account users, Hu Z served a new method to deter the multi-account and reduce the possible error it brought [23]. The data source will be more accurate, therefore, the result will be better. Furthermore, people are also trying their best to improve data security to avoid the leak of users' information, like the theft of data in Facebook in 2018 [24]. After that case, people are trying their best to find a new method to avoid similar accidents happening again, IoT is an excellent example [9]. The study that improves the firework will give the crowds more confidence, therefore, they will trust the enterprise and provide more data for prediction.

Fig.4. The influence on usefulness in customers perspective when lack of 3Vs

Some scholars attempt to analyse the trend of the appliance of big data and try to predict the future of big data [6]. According to Kuo, Y. H's research, it could be observed that the big data had explored three stages from operating to analysis and then the system. And Kuo predicted that the future of big data would possibly be in the personalised production research area, however, other research proved that the use of big data might be out of our imagination. A recent study of the application of big data in map areas is noticed, and it hints to people that the depth is not the only developing direction and also the width [2].

Finally, there is still a heated argument about the privacy issue. The broad line of privacy in online information is not straightforward, especially when it is relative to more than one country. However, scholars are doing their best to fix this question, and it is believable that the big data will be better with their contribution [25]. Furthermore, people are redefining this concept's big data, which is essential for future development [26]. It can be guaranteed that big data will have a bright future with all these works.

5.CONCLUSION

In conclusion, this paper first shows the usefulness of big data in some cases that present people's expectations and the wide use of big data analysis. Secondly, provide some examples to evaluate the accuracy of big data analysis. The requirement of big data analysis accuracy is mentioned, and the possible consequence of inaccurate big data analysis is listed. In the third part, the possible reasons are shown according to references. It is essential to deal with the problems about the public, big data and the analyzing method. Finally, new advances and possible future development are suggested. The future developing direction of methodology and privacy policy is recommended.

REFERENCES

[1] David Goldston, "Big data: Data wrangling," Nature News, vol. 455, no. 7209, pp. 15–15, 2008.

[2] Jin Wang, Yaqiong Yang, Tian Wang, R Simon Sherratt, and Jingyu Zhang, "Big data service architecture: a survey," Journal of Internet Technology, vol. 21, no. 2, pp. 393–405, 2020.

[3] Natarajan Deepa, QuocViet Pham, Dinh C Nguyen, and et al., "A survey on blockchain for big data: Approaches, opportunities, and future directions," arXiv:2009.00858, 2020.

[4] Showkat Ahmad Bhat and NenFu Huang, "Big data and ai revolution in precision agriculture: Survey and challenges," IEEE Access, vol. 9, pp. 110209–110222, 2021.

[5] Maryam Ghasemaghaei and Goran Calic, "Assessing the impact of big data on firm innovation performance: Big data is not always better data," Journal of Business Research, vol. 108, pp. 147–162, 2020.

[6] S Smys, "Survey on accuracy of predictive big data analytics in healthcare," Journal of Information Technology, vol. 1, no. 02, pp. 77–86, 2019.

[7] Danah Boyd and Kate Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," Information, communication & society, vol. 15, no. 5, pp. 662–679, 2012.

[8] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and et al., "Bigdata science in porous materials: materials genomics and machine learning," Chemical reviews, vol. 120, no. 16, pp. 8066–8129, 2020.

[9] Lei Zhang, Yu Huo, Qiang Ge, and et al., "A privacy protection scheme for iot big data based on time and frequency limitation," Wireless Communications and Mobile Computing, vol. 2021, 2021.

[10] Wei Ji and Lihui Wang, "Big data analytics based fault prediction for shop floor scheduling," Journal of Manu- facturing Systems, vol. 43, pp. 187–194, 2017.

[11] Prasan Kumar Sahoo, Suvendu Kumar Mohapatra, and ShihLin Wu, "Analyzing healthcare big data with prediction for future health condition," IEEE Access, vol. 4, pp. 9786–9799, 2016.

[12] Eric B Hekler, Predrag Klasnja, Guillaume Chevance, and et al., "Why we need a small data paradigm," BMC medicine, vol. 17, no. 1, pp. 1–9, 2019.

[13] In Lee, "Big data: Dimensions, evolution, impacts, and challenges," Business horizons, vol. 60, no. 3, pp. 293– 303, 2017.

[14] Gunasekaran Manogaran and Daphne Lopez, "Spatial cumulative sum algorithm with big data analytics for climate change detection," Computers & Electrical Engineering, vol. 65, pp. 207–221, 2018.

[15] Marzieh Fathi, Mostafa Haghi Kashani, Seyed Mahdi Jameii, and et al., "Big data analytics in weather forecasting: A systematic review," Archives of Computational Methods in Engineering, pp. 1–29, 2021.

[16] Soumendra Mohanty, Madhu Jagadeesh, and Harsha Srivatsa, Big data imperatives: Enterprise 'Big Data' warehouse, 'BI'implementations and analytics, Apress, 2013.

[17] Robert Northcott, "Big data and prediction: Four case studies," Studies in History and Philosophy of Science Part A, vol. 81, pp. 96–104, 2020.

[18] Pablo Olivera, Silvio Danese, Nicolas Jay, and et al., "Big data in ibd: a look into the future," Nature Reviews Gastroenterology & Hepatology, vol. 16, no. 5, pp. 312– 321, 2019.

[19] Rob Kitchin, "Big data–hype or revolution," The SAGE handbook of social media research methods, p. 27, 2017.

[20] Kate Crawford, "Following you: Disciplines of listening in social media," Continuum, vol. 23, no. 4, pp. 525–535, 2009.

[21] Zafar Gilani, Reza Farahbakhsh, and Jon Crowcroft, "Do bots impact twitter activity?," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 781–782.

[22] Kate Galloway, "Big data: A case study of disruption and government power," Alternative Law Journal, vol. 42, no. 2, pp. 89–95, 2017.

[23] Zhongshun Hu, Fei Chen, Chenhao Xie, and et al., "Towards multiaccount users detection and connection

[24] based on telecom data," in 2016 3rd International Conference on Mechatronics and Information Technology. Atlantis Press, 2016, pp. 195–199.

[25] Valentinus Paramarta, Muhammad Jihad, Ardhian Dharma, and et al., "Impact of user awareness, trust, and privacy concerns on sharing personal information on social media: Facebook, twitter, and instagram," in 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2018, pp. 271–276.

[26] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi, and et al., "Privacy issues and data protection in big data: a case study analysis under gdpr" in 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018, pp. 5027–5033.

[27] Maddalena Favaretto, Eva De Clercq, Christophe Olivier Schneble, and et al., "What is your definition of big data? researchers' understanding of the phenomenon of the decade," PloS one, vol. 15, no. 2, pp. e0228987, 2020.