The Necessity and Public Concerns about Privacy in Big Data and Data Mining in Education

Chandra Wijaya  
School of Computer Science  
BINUS UniversityJakarta, Indonesia  
chandra.wijaya006@binus.ac.id

Mulia Darma  
School of Computer Science  
BINUS UniversityJakarta, Indonesia  
mulia.darma@binus.ac.id

Nicholas Adidharma Sukardi  
School of Computer Science  
BINUS UniversityJakarta, Indonesia  
nicholas.sukardi@binus.ac.id

*Abstract*—Many people are using big data and data mining to access information to make the best decision for their businesses and many more. However, some people are concerned that the use of big data and data mining may cause privacy issues and they think it’s dangerous to use that technology. This research will find out whether big data and data mining are dangerous or not using survey quantitative methods. The result of the survey is that most of the people think that big data is dangerous and some people think it’s not dangerous with their own opinion.

Keywords—Big data, Data mining, Privacy, Education, Public Opinion

# Introduction

Along with the advances in technology, Big data and Data mining has gained and received more attention due to the implementation and capability of the technology in many fields. Big data and Data mining provides a lot of methods and techniques that can be implemented and used to process data. With the use of Big data and Data mining in the field of education, analysis of data in a large scale can be produced in order to help improve said field. More specific topic that has been chosen for this research paper is how the method, techniques and implementations have a potential to produce concerns about privacy issues and the public opinion of said concerns that can surface.

From the application of Big data and Data mining itself, there has been research that uses Data mining to predict the percentage of dropout cases along with the factors that may contribute to the dropout case [1]. For the matter of privacy of Data mining and Big data, in a paper by Reidenberg, J. R., & Schaub, F.[2], policies to help the concerns and issues of privacy in the usage of Big data and Data mining is explored. [2] The use of Big data also causes privacy concerns like the InBloom case that have been discussed in paper from Reidenberg, J. R., & Schaub, F. (2018).

This research paper aims to explore the public opinion regarding the issues of privacy that occurs in the application/usage of Big data and Data mining. In this paper we will use the quantitative method of survey(s) to gain the opinions regarding the privacy concerns of using Big data and Data mining in the field of education. The plan of the execution of this research is by first creating and constructing the questions in the questionnaires for the survey. Then the said questionnaires will be given to people (doesn’t matter the background of the respondent) and asked to answers the questions in the questionnaires. After we have gathered enough respondent for us to conduct the analysis of the answers (around 50 respondents) we will begin to carefully analyse the contents of the answers and make out the opinion of the respondents and make a conclusion of the general opinion about the topic for this research.

# Literature Review

The use of Big data and Data mining have been common in a lot of fields, where Big data and Data mining can be used to process and extract hidden information from the data on a large scale. With many uses of Data mining and Big data where there is a lot of technique, methods, and algorithms, which leads to the concern of security and privacy. Bagga, S., & Sharma, A. (2018) [7] said that the definition of big data is a collection of large, distributed, multidimensional data. In the paper by Singh, D.S., & Singh.G [8]. (2017), there are three basic characteristics or 3V of data such as volume, variety, and velocity. However, that 3V have already expanded by many organizations and professionals to 4V by adding a new value characteristic and got expanded again to 5V by adding a veracity concept to the characteristic and there will be five characteristics such as volume that refers to the size of the data, velocity that refers to the speed at which data is generated, variety that refers to the different formats in which data is generated, value that refers to the ability of companies to analyze data and to provide a better understanding of the various key areas, and veracity that refers to the accuracy or truth of the data.

Big data not only used for processing data, but also can be used to predict instructor performance using data mining as paper from Ahmed, A.M., Rizaner, A., & Ulusoy, A.H. (2016) [3] said. Data mining also can be used to obtain hidden information like the cause of a result as the paper from Abu-Oda, G. S., & El-Halees, A. M. (2015) [1] said. Where data mining can be applied to get a reason why a student gets drop out and to predict dropout by using Decision Tree (DT) and Naive Bayes (NB). A paper by Kaur, P., Singh, M., & Josan, G.S. (2015) said that data mining can be used to identify the slow learners among students and display it by a predictive data mining model using classification based algorithms. [4].

[2] The use of Big data also causes privacy concerns like the InBloom case that have been discussed in paper from Reidenberg, J. R., & Schaub, F. (2018). One of many causes of this privacy issues is because the lack of big data privacy protection as stated by Zhang, D. (2018). A paper by Mishra, A.D., & Singh, Y.B. (2016) [6] said that there is many significant concerns about this privacy issues such as anonymization could become impossible, big data analytics are not complete accurate, legal protection exist for the involved individuals, security intelligence and compliance audit, unethical actions based on interpretations, privacy breaches and fraud incidents, discrimination, and data masking could be defeated to reveal personal information.

From the paper we have reviewed, we found the use of Big data and Data mining bring us an advantage in education, not only bring an advantage but it will cause a problem and privacy concern.

# METHODOLOGY

The Methodology that we used and would like to introduce to the readers is the methodology of using the quantitative method of using survey(s). As the question of this paper is aimed at determining the general public's concerns about the usage of Big data and the public’s opinion regarding the necessity of the usage of Big data in education, we needed qualitative data and the method that best suits in collecting said data is by using survey(s).

The data is collected via surveys that our team has made filled with the questions to our paper’s subject, the questions our team designed are aimed at gathering answers in form of opinions, the form that these questions take vary between multiple-choice questions and short answers questions. The survey takes form on an online questionnaire. The sample size is currently at the time of writing at 60 responses.

* The questions in the questionnaire are as follows, the first three questions are the introductory questions aimed at getting the personal information of the respondent.
* The next question “Apakah anda berasal dari jurusan Computer Science” (“Do you major in Computer Science?”) is aimed at getting the information whether the respondent has background education in Computer Science and can be used to analyze the other answers.
* Next is the question of “Apakah anda mengetahui mengenai Big Data” (“Do you know about Big data?”) to factor in respondents that do not major in Computer Science but do know about Big data and also those who major in Computer Science but don't/haven’t learned about Big data.
* Next the question “Apakah menurut anda penggunaan Big Data berbahaya?” (“Do you think the usage of Big Data is dangerous?”). This is aimed at getting the opinion about the usage of Big data.
* Next the question “Berdasarkan jawaban anda sebelumnya kenapa?” (“Based on your previous answer, why?”). This question is to gather the reason for the previous questions' answer that made them think the usage of Big data is dangerous and vice versa.
* Next the question “Menurut anda keamanan data adalah hal penting?” (“Do you think that data security is important?”). This question seeks answers from the respondents in the form of their opinion about data security.
* Next the question “Jika tempat anda (universitas, sekolah, tempat kerja) menggunakan big data. Apakah akan menjadi suatu hal yang baik?” (“If your university, school or workplace implements big data, do you think it is good?”). This question asks about the opinion of the respondent on how they would think if big data were to be implemented in their spaces.
* Next the question “Berdasarkan jawaban anda sebelumnya kenapa?” (“Based on your previous answer, why?”).

Our method for analyzing the data collected from the respondents answers in our questionnaire is by first transforming the data from the Google Form into .csv and then into an SQL database with the help of the online file converter into SQL, SQLizer. Our analysis consists of Content Analysis where we try to grasp and uncover the meaning of words in the answers. This Content Analysis is with the consideration of the question “Do you major in Computer Science?”.

Other Tools that we used to help us in our analysis is by using Google Colab to use the pyplot and toPandas to help us visualize and analyse the data that we got into more visible and easier form to understand.

# RESULT

Here are the results and the content analysis of the data that we have gathered via the methodology that we have chosen. For the Questions that receive “Ya” (Yes) or “Tidak '' (No) as an answer we can visualize using pie charts, while opinions we would have to analyze its content. In the results we can also link two answers together and visualize them for better understanding.

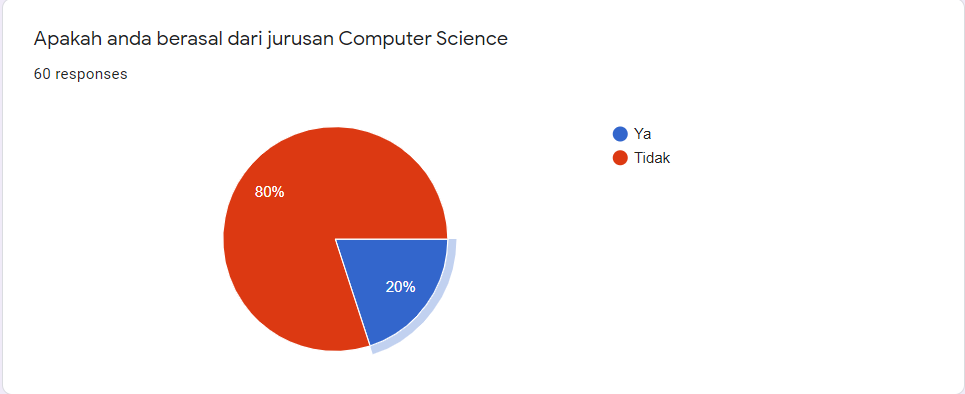


Figure 1. Pie chart (1)

We knew that the people we ask are 20% from Computer Science and 80% are not from Computer Science. From what we know, people in Computer Science also know about big data. But from what we get only 1 person that didn’t know about big data. And the others know about big data. This will be taken into consideration for the analysis for the answers after this one.

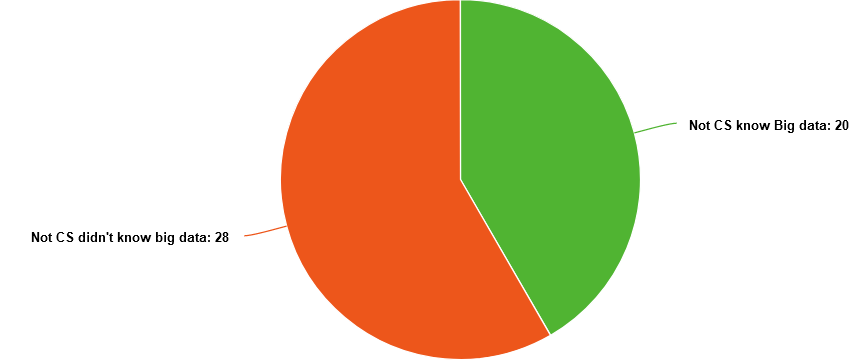


Figure 2. Pie chart (2)

From the data that we compiled, there are 48 people/respondents (80% of total of our respondents) are not from the Computer Science major. And from those 48 people, 20 of them know about Big data while the rest 28 respondents don’t know about Big data.

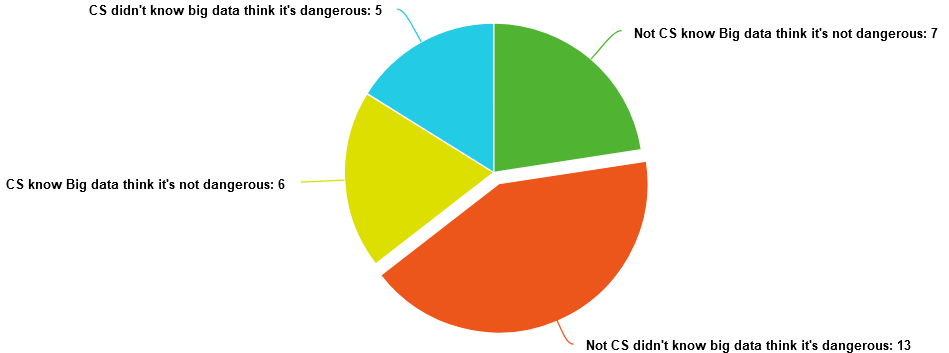


Figure 3. Pie chart (3)

From this chart we can see that more than half the people that know Big data think it’s dangerous. Some of them say they only feel that Big data is dangerous but some of them think that big data is dangerous because it has the potential to be abused by some people for their own profit. And the people that do not think it’s dangerous say that if the security that they use on big data is secure it will be a good way to get information from the data.

Chart, pie chart

Description automatically generated

Figure 4. Pie chart (4)

All of the people in Computer Science agree that big data is a good way to get information from data even though there are 5 people that think Big Data is dangerous but they still support the use of Big data. But there are also people that are not from Computer Science who think that big data is good. Also, there are a few people that don’t think big data is good. Part of them said they don’t know about big data as a whole but the other said it would be dangerous if the data got leaked.

With these answers we get that more than half of the people that knew about big data think it’s dangerous but there will always be a problem and with this current development we as a people who didn’t even know about big data will feel safe. With the use of one of many ways of encryption, this method can be used to make data useless if you don’t have the key. If other people got the data but didn’t get the key it would be useless for them because the data couldn’t be read. The other method is to build a firewall that can be used to prevent an attack.

##### References

1. Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, *5*(1), 15.
2. Reidenberg, J. R., & Schaub, F. (2018). Achieving big data privacy in education. *Theory and Research in Education*, *16*(3), 263-279.
3. Ahmed, A.M., Rizaner, A., & Ulusoy, A.H. (2016). Using data mining to predict instructor performance. Procedia Computer Science 102 (2016), 137 – 142.
4. Kaur, P., Singh, M., & Josan, G.S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science 57 ( 2015 ), 500 – 508.
5. Zhang, D. (2018). Big Data Security and Privacy Protection. Atlantis Press, 77, 275-278.
6. Mishra, A.D., & Singh, Y.B. (2016). Big Data Analytics for Security and Privacy Challenges. International Conference on Computing, Communication and Automation, 50 – 53.
7. Bagga, S., & Sharma, A. (2018). Big Data and its Challenges: A Review. 4th International Conference on Computing Sciences, 183 – 187.
8. Singh, D.S., & Singh.G. (2017). Big data – A Review. International Research Journal of Engineering and Technology (IRJET), 4(4), 822 – 824.