Available online athttps://journal.rescollacomm.com/index.php/ijeer/index

International Journal of Ethno-Sciences and Education Research

e-ISSN 2776-7590

Vol. 3, No. 3, pp. 89-94, 2023

Comparative Analysis of Community Sentiment Against the Implementation of Booster Vaccination in Indonesia Using the K-Nearest Neighbor and Naïve Bayes Classifier Methods

Budiman[1], Wulandari[2*] Chairul Habibi[3]

[12,3] *Faculty of Technology and Informatics Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia*
**Corresponding author email: donghyunk80@gmail.com*

Abstract

Sentiment analysis is a person's opinion or view of a particular object that produces positive, negative, or neutral sentiments. The government's effort during the COVID-19 pandemic is to call for the implementation of a booster vaccination program to the public. Based on this, it produces several public sentiments, some of which are uploaded on the Twitter social media platform, which generate positive and negative sentiments. To find out the classification of public sentiment, the researchers carried out calculations using the K-Nearest Neighbor and Naïve Bayes Classifier methods. Based on the calculation results, it was found that the public sentiment was positive at 98% and negative at 2%. This means that the community is enthusiastic and supports the booster vaccination program. Then the comparison based on the calculation results, namely the K-Nearest Neighbor method with a K value of 3 resulting in an accuracy calculation of 97.33% and using the Naïve Bayes Classifier method, an accuracy calculation of 97.35% can be generated. So it can be seen that using the Naïve Bayes Classifier method has a higher accuracy than the K-Nearest Neighbor method.

Keywords: Comparative analysis, booster vaccination, K-Nearest neighbor.

1. Introduction

The era of globalization and the development of technology today are very rapidly developing. At the moment technology is constantly being developed and producing new discoveries that are beneficial to humans. Developers and researchers certainly periodically issue innovations and new ideas in developing more modern and useful technologies. New innovations are widespread and reach all parts of the world. People around the world use technology from government agencies, companies, education, workers, students, and the wider community. Everyone uses technology to help their daily lives. In helping with work in the office, at home, at school, in hospitals for health as well as especially the treatment of patients, researchers continue to research and make new discoveries in dealing with diseases so that they can be cured (Tomlinson, 2011).

The latest technology or inventions / innovations certainly require an analysis to find out the effectiveness of the technology or innovation that has been made. Analyzing technology or innovations that have been made has a lot of data that can be analyzed, one of which is the data processing analysis process in machine learning. Data processing can be done with various techniques and methods such as text mining. In text mining, researchers can analyze data in the form of text and produce outputs that can be represented by sentiment analysis. Sentiment analysis can determine sentiment in the form of negative, positive, and neutral from text data or sentences of opinions of people who have used the new technology or innovation. The text data obtained is of course very much, for this reason, in analyzing sentiment, techniques or learning methods are needed regarding grouping or classification of data, so learning methods such as the naïve bayes and k-nearest neighbor methods are needed. This method classifies data obtained from data sources which are then processed and analyzed to produce an accuracy of the analysis that has been carried out (Thanaki, 2018).

During the current Covid-19 pandemic, the government has issued new efforts to help the community in dealing with this pandemic by urging people to vaccinate. The implementation of vaccination is expected to minimize and keep the public from being exposed to covid-19. Currently, the government is no longer tightening people's access to travel or activities outside the home, but the government is appealing for people to vaccinate boosters. The Indonesian people have made many appeals from the government, so many people have given their opinions about the

vaccination. Their opinions are mostly expressed in a social media such as facebook, twitter, whatsapp, instagram and other social media. Twitter users in Indonesia are quite a lot, they issue their opinions on social media. People who already know the appeal for booster vaccination some have positive opinions but some also have negative opinions. Based on this, it makes researchers to classify people's opinions and what percentage of positive and negative opinions are in society. Researchers are interested in analyzing every opinion expressed by the public on twitter through sentiment analysis to find out the classification of negative, neutral, or positive sentiments expressed by the community against the appeal for the implementation of booster vaccinations. Based on this problem, researchers are interested in taking the title of the study, namely "Comparative Analysis of Public Sentiment towards the implementation of Booster Vaccination in Indonesia Using the K-Nearest Neighbor and Naïve Bayes Classifier Methods".

2. Literature Review

2.1. Sentiment analysis

According to Fikri Aldi, sentiment can have meaning as an opinion or view based on excessive feelings about something. Sentiment is usually found in statements and sentences that have opinions. Sentiment is also useful to know the feelings that a person gives towards a particular topic or object. Sentiment analysis is one of the research areas contained in natural language processing, linguistic computing, and text mining. Sentiment analysis or can also be called opinion mining is a computational study of the opinions or opinions of others as well as the emotions contained in the entities, events, and attributes possessed. The task performed by sentiment analysis is to group the polarity contained in a text, whether it is contained in documents, sentences, or aspect-level features whether the opinions expressed are positive, negative, or neutral (Islam, 2020).

2.2. Text Mining

According to Fikri Aldi, text mining is one of the techniques in data mining that uses text data. Text mining can be defined as a process carried out to extract implicit knowledge hidden in textual data. The implicit knowledge gained from the extraction of text mining needs to be managed in depth because it has a different output compared to management in other data types so it needs separate analysis. Text mining is also part of information retrieval (information retrieval), which is a science or technique that studies procedures and methods to find and rediscover relevant information or sources of information that are related to what is needed. In general, there are 3 (three) main parts of text mining, namely text pre-processing, feature selection, and text analytics (Islam, 2020).

In general, the algorithms used in text mining have almost the same performance as the algorithms in data mining. The difference that is the main factor in data management using data mining and text mining is in the data type that is the object of work. Data mining works on structured data while text mining works on the other hand. In text mining, there are several processes that distinguish it from other data mining processing, namely text categorization, extraction of concepts or entities, text clustering, sentiment analysis, granular taxonomy production, document inference and entity relationship modeling (Hirji, 2001). A text can be defined as an unstructured piece of data consisting of a collection of strings called words. Even if the string set of a deep text is in a wide scope, of course it requires the meaning of the individual strings and the combination of those strings is set with rules called grammar to create the text. Text data used in the text mining process can be obtained from various sources including articles, news to social media. Texts obtained from these various sources, especially from social media, usually have a different structure and use non-standard language. This is where text mining acts as a technique that provides text processing with various differences in grammatical rules so that it can be obtained and concluded what information is contained in a text (Islam, 2020).

3. Materials and Method

3.1 Materials

According to Yudhanto Twitter was founded by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams in late October 2006. The service gained popularity since March 2007 (Skeels & Grudin, 2009). According to databoks twitter is one of the social media networks that is popularly used by people in Indonesia. According to a Statista report, there are 18.45 million users of the application founded by Jack Dorsey in the country as of January 2022. This achievement puts Indonesia as the 5th most Twitter user country in the world. According to the Reuters Institute Digital News Report 2021 survey report, Twitter is the social media that users use the most to search for news compared to Facebook, Youtube, Instagram, Snapchat, and TikTok. Currently it is still inferior to the growth of facebook, but twitter has its own advantages (Skeels & Grudin, 2009).

The administration of the first and second doses of vaccine is not enough to break the chain of transmission of covid-19. For this reason, a third dose or booster vaccine is needed. The purpose of the third dose of vaccine is to increase immunity to ward off infection with the covid-19 virus. As is known, over time the effectiveness of the first

and second vaccines began to weaken so that a booster vaccine was needed to reshape the antibody and extend protection against the body from the covid-19 virus. To be able to get the booster vaccine, you can check it in the PeduliLindungi application. Provided that the distance of administration with the second dose is at least three months. The injection of booster vaccines is different from the types of vaccines that have been given to the first and second vaccines. Some of the benefits obtained when receiving the third dose of vaccine are to ward off the entry of the covid-19 virus, increase body immunity, increase immunity and extend the protection period from the covid-19 virus. Giving booster vaccinations also results in side effects for the recipients that are felt by the vaccine recipients, usually namely pain at the injection site, heat, muscle pain, dizziness, and fatigue. All of these are body reactions showing that the body's immunity is resisting the vaccine received (Million, 2020).

3.2 Method

This research uses quantitative research methods. According to Sugiyono, quantitative data is a research method based on positivistic (concrete data), research data in the form of numbers that will be measured using statistics as a calculating test tool, related to the problem under study to produce a conclusion (Sugiyono, 2010). The Nearest neighbor method according to Kubat is one of the classification techniques in which an object is identified into a certain group or class based on the tendency of its neighboring objects. One of the techniques in the nearest neighbor is k-nearest neighbor. The value of k in this case is the number of objects or entities that will be taken into account in the identification vote. K-nearest neighbor has several advantages, namely being tough against noise and effective data training when the training data is large (Bentley, 1984). The naïve bayes method is a method used to qualify comment data to obtain analytical sentiment. To perform sentiment classification will use the data from the pre-processing process. After the data is successfully trained, it will then be tested using test data to test the results of the accuracy of the classification carried out (Rahardi, 2022)

4. Results and Discussion

4.1 Result

The results of the analysis of public sentiment towards the implementation of the booster vaccination, it can be seen the positive and negative sentiments from the data that has been collected and processed. It is known that the positive sentiment of the community towards the booster vaccination was 97% and the negative sentiment was 3%. In the calculation of the K-Nearest Neighbor method with a K value of 3, the resulting calculation accuracy is 97.33% with a positive value of 98% and a negative of 2%. Using the Naïve Bayes Classifier method, an accuracy calculation of 97.35% can be obtained with a positive value of 98% and a negative of 2%. Next, testing the model using the confusion matrix resulted in an accuracy calculation of 92% with a positive value of 92% and a negative of 8%. Based on the calculation results of the method that has been obtained, it can be seen that using the Naïve Bayes Classifier method has higher accuracy compared to the K-Nearest Neighbor method.

4.2 Discussion

This study uses a data set in the form of tweets from the Indonesian people uploaded on Twitter social media. Retrieval of data on Twitter using the keyword "vaccination booster". Tweets uploading public sentiment regarding booster vaccinations, taken by scrapping data using an instant data scraper which is already available on the Google Chrome extension. The results of the data that has been retrieved using the instant data scraper are stored in files of type xlsx. The data taken was 10,173 data, from January 1 2022 to June 30 2022. It can be seen that the tweet data uploaded by the Indonesian people is very diverse and there are different sentiments from each upload. The data that has been stored is then labeled to determine the sentiment that exists in the data that has been retrieved.

Labeling of the data set that has been taken is done manually for labeling the data into positive and negative category classes. It can be seen that the positive sentiment from the existing data has a greater frequency compared to the negative sentiment. At this stage, if there is the same data, it will be deleted. This process changes the previous 10,173 data into 6,496 data. The chart shows that positive sentiment is greater than negative sentiment. There were 9,495 data, 6,322 positive comments and 173 negative comments.

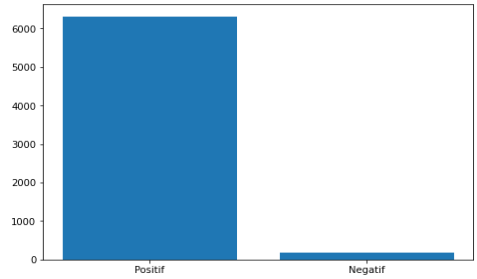

Figure 1: Data Labeling Chart

Cleaning, namely cleaning data in which there are symbols, punctuation marks, numbers, hashtags and so on. This process includes case folding by changing all uppercase or capital letters to lower case in the document. This stage will be assisted with the help of the RegEx library or regular expressions. In addition to changing uppercase to lowercase, it also removes punctuation, hashtags, symbols, and so on.

Table 1: Cleaning

Before	After
BIN Gelar vaksinasi booster khusus di aceh utara	bin gelar vaksinasi booster khusus di aceh utara

The results of the cleaning process, the data set will be processed by breaking it down from sentences into word for word. At this stage it will be assisted by the NLTK library. This process is to facilitate the sentiment process based on word for word and the calculation process using the classification method.

Table 2: Filtering

Before	After
bin gelar vaksinasi booster khusus di aceh utara	['bin', 'gelar', 'vaksinasi', 'booster', 'khusus', 'di', 'aceh', 'utara']

The results of the tokenizing process will then select words that are not important and delete these words. This stage is assisted by the NLTK library, which uses stopwords. Examples of omitted words are I, and, or, in, to, ok and others.

Table 3: Stemming

Before	After
['bin', 'gelar', 'vaksinasi', 'booster', 'khusus', 'di', 'aceh', 'utara']	['gelar', 'vaksinasi', 'booster', 'khusus', 'aceh', 'utara']

The stemming process is carried out by changing words that have affixes to become basic words by removing affixes in front of or after the base word. This stage is assisted by the Sastrawi and Swifter libraries.

Table 4: Tokenizing

Before	After
['gelar', 'vaksinasi', 'booster', 'khusus', 'aceh', 'utara']	['gelar', 'vaksinasi', 'booster', 'khusus', 'aceh', 'utara']

Based on the results of the tokenization carried out, it can be seen that the frequency of the words that are most widely used, the following is the result of word visualization in the form of wordcloud.

Figure 2: Wordcloud Data

The data is divided into two, namely training data and testing data. The data shared is 70% training data and 30% testing data. Data sharing is done to be used when performing classification calculations.

The K-Nearest Neighbor classification is carried out using the help of a python library, namely Neighbor, KneighborsClassifier, confusion matrix, and f1_score. The first step is the same as the previous method, namely installing the required library. Furthermore, all libraries are declared and proceed to call the data set that will be used. After that, the calculation process for the K-Nearest Neighbor method is carried out, namely classification is carried out using the KNeighborsClassifier library. Then determine the K value in the K-Nearest Neighbor method with a K value of 3 which is obtained from the search results for the maximum value. Based on the highest score, K is 3.

From the results of data accuracy, it can be seen that the precision and recall values in each class can be said to be the level of the system's ability to find accuracy between the information requested by the user for the positive class by 98%, for the negative class by 55%. While the success rate of the system in retrieving information for the positive class is 100%, for the negative class is 13%. This means that the success of the system performance in retrieving negative information in documents is very low. It is known that sentiment analysis using the K-Nearest Neighbor method has a positive sentiment of 98% and 2% negative.

The classification of the naïve Bayes classifier method is carried out using the help of the python library, namely Naïve Bayes, MultinomialNB, confusion matrix, and f1_score. The first step is the same as the previous method, namely installing the required library. Furthermore, all libraries are declared and proceed to call the data set that will be used. After that, the calculation process for the naïve Bayes classifier method is carried out, namely the classification is carried out using the MultinomialNB library. Then the performance of the algorithm used in the test data is obtained.

From the results of data accuracy, it can be seen that the precision and recall values in each class can be said to be the level of the system's ability to find accuracy between the information requested by the user for the positive class of 98%, for the negative class of 0%. While the success rate of the system in retrieving information for the positive class is 100%, for the negative class is 0%. This means that the success of the system performance in retrieving negative information in documents is very low. It is known that sentiment analysis using the K-Nearest Neighbor method has a positive sentiment of 98% and 2% negative.

The accuracy of the two methods used is known to have values that are not much different, there is still a slight difference. The following is the accuracy value of the two methods used.

Fold	KNN	NBC
0	0.929231	0.935385
1	0.975385	0.978462
2	0.989231	0.990769
3	0.973846	0.981538
4	0.958462	0.961538
5	0.976888	0.976888
6	0.983051	0.990755
7	0.979969	0.981510
8	0.981510	0.981510
9	0.958398	0.956857

Figure 3: Accuracy Comparison of the Two Methods

The average value of the accuracy of the two methods, namely for the K-Nearest Neighbor method is 97% and the Naïve Bayes Classifier is 97.3%. Both of these methods have the same high accuracy value, but have different values. Even though the difference is very slight, the Naïve Bayes Classifier method is larger than the K-Nearest Neighbor method.

	KNN	NBC
average	0.970597	0.973521

Figure 4: Average Comparison Accuracy of the Two Methods

5. Conclusion

Based on the results of testing the classification method that has been carried out, the researcher can draw the following conclusions. The results of the analysis of public sentiment towards the implementation of booster vaccinations, positive and negative sentiments can be identified from the data that has been collected and processed. It is known that the positive sentiment of the community towards the booster vaccination was 97% and the negative

sentiment was 3%. In the calculation of the K-Nearest Neighbor method with a K value of 3, the resulting calculation accuracy is 97.33% with a positive value of 98% and a negative of 2%. Using the Naïve Bayes Classifier method, an accuracy calculation of 97.35% can be obtained with a positive value of 98% and a negative of 2%. Next, testing the model using the confusion matrix resulted in an accuracy calculation of 92% with a positive value of 92% and a negative of 8%. The data collected has many positive values and only a few negative data, meaning that it can be concluded that the community is very enthusiastic and supports the implementation of booster vaccinations in Indonesia. This means that the government's program to carry out booster vaccinations in Indonesia has been well implemented by the community. The results of the performance comparison of the methods used, namely the K-Nearest Neighbor and Naïve Bayes Classifier methods, produce a comparison using K-Fold Cross Validation, namely for K-Nearest Neighbor has an accuracy value of 97.33% and Naïve Bayes Classifier has an accuracy value of 97.35%. This means that the Naïve Bayes Classifier method has a higher level of accuracy compared to the K-Nearest Neighbor method. Both of these methods produce the same sentiment analysis, namely positive sentiment of 98% and negative sentiment of 2%. Both methods have a high accuracy value so that it is sufficient and good to be used in a system.

Acknowledgments

This research was supported by [Indonesian University of Education], [Ma'soem University] and [Indonesian Cooperative University]. We thank you for helping in the smooth running of this community service activity.

References

Bentley, J. (1984). Programming pearls: algorithm design techniques. *Communications of the ACM, 27*(9), 865-873.

Hirji, K. K. (2001). Exploring data mining implementation. *Communications of the ACM, 44*(7), 87-93.

Islam, S., Ab Ghani, N., & Ahmed, M. (2020). A review on recent advances in Deep learning for Sentiment Analysis: Performances, Challenges and Limitations. *Compusoft, 9*(7), 3775-3783.

Million, M., Jarrot, P. A., Camoin-Jau, L., Colson, P., Fenollar, F., ... & Raoult, D. (2020). Natural history of COVID-19 and therapeutic options. *Expert review of clinical immunology, 16*(12), 1159-1184.

Rahardi, M., Aminuddin, A., Abdulloh, F. F., & Nugroho, R. A. (2022). Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia. *International Journal of Advanced Computer Science and Applications, 13*(6).

Skeels, M. M., & Grudin, J. (2009, May). When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proceedings of the 2009 ACM International Conference on Supporting Group Work*, 95-104.

Sugiyono, S. (2010). *Educational Research Methods: Quantitative, Qualitative, and R & D Approaches.* Bandung: CV. Alfabeta.

Thanaki, J. (2018). *Machine Learning Solutions: Expert Techniques to Tackle Complex Machine Learning Problems Using Python.* Britania Raya: Packt Publishing.

Tomlinson, B. (Ed.). (2011). *Materials development in language teaching.* Cambridge University Press.