Sentiment Analysis of Vaccine Booster during Covid-19: Indonesian Netizen Perspective Based on Twitter Dataset

Khoiril Hikmah[1], Abd. Charis Fauzan[2], Harliana[3]

[1,2,3]Program Studi Ilmu Komputer, Universitas Nahdlatul Ulama Blitar
[1,2,3]Jl. Masjid No. 22, Blitar, Jawa Timur, 66117, Indonesia
E-Mail: avp.khoir@gmail.com[1], abdcharis@unublitar.ac.id[2], harliana@unublitar.ac.id[3]

Abstract

Corona virus has become a global threat at the end of 2019. The spread of this corona virus is very fast to all countries in the world. The World Health Organization (WHO) has determined the status of the corona virus as a global pandemic called the Corona Virus Desease 2019 (Covid-19). Indonesia was detected first case of Covid-19 on March 2, 2020. After that, the number of Covid-19 cases in Indonesia increased every day and had a real impact on various sectors sectors, such as the economic and education. The Indonesian government has handled this health disaster, one of way that has been done is by holding a COVID-19 vaccine. Includes one dose vaccine, second dose vaccine and vaccine booster. The existence of this vaccine booster has received various opinions from Indonesian netizens who were conveyed through social media of Twitter. Therefore, this research aims to analyze the sentiments of Indonesian netizens about booster vaccination. In this study, data was collected from the Twitter dataset by crawling using the Rapidminer. Then, the data preprocessing stage is carried out consisting of: case folding, tokenizing, filtering and stemming. Sentiment classification is divided into positive sentiment, negative sentiment and neutral sentiment. Sentiment classification resolved using the Naive Bayes algorithm. This research resulted sentiment of vaccine booster during Covid-19 based on Indonesian netizen, include tweets of 23% positive sentiment, tweets of 15% neutral sentiment and tweets of 76% negative sentiment with an accuracy rate of 89%.
.

Keywords: Twitter, Vaccine Booster, Naive Bayes

I. INTRODUCTION

At the end of 2019, there was a COVID-19 outbreak in China, precisely in Wuhan, Hubei province. The spread of this corona virus is very fast and easy to all countries, so the World Health Organization (WHO) has declared the status of the corona virus as a global pandemic on 11 February 2020 [1]. Indonesia was first detected with the covid-19 virus with residents who were infected on March 2, 2020 which caused negative impacts in various sectors, such as the economic and education sectors [2].

The Indonesian government is taking measures so that the spread of the COVID-19 virus can be suppressed and the negative impacts can be minimized by implementing vaccinations of dose one, dose two and booster [3]. The existence of this booster vaccine has received various opinions from Indonesian netizens, which have been expressed through Twitter social media. The opinions of Indonesian netizens on Twitter social media can be analyzed using various analytical methods [4]. Sentiment analysis method is a method for analyzing data obtained from the internet [5]. With sentiment analysis, opinion polarities can be

collected, in order to predict the positive or negative feelings of Indonesian netizens [6].

Many researchers take the topic of sentiment analysis because this topic is very interesting to study. First, the research conducted by Sri Lestari [7] on sentiment analysis of the Sinovac vaccine using the Naive Bayes Algorithm found tweets with positive sentiment as much as 86% and tweets with negative sentiment as much as 14% with an accuracy rate of 92.96%. The two studies conducted by F. Septianingrum [8] on the issue of the covid-19 vaccine resulted in 78% accuracy and 80% recall value. There are also researchers who took the topic of this analysis for data classification in the form of text by I Wayan Desta Gafatia [9] with the topic of the pros and cons of the covid-19 vaccine with results that responded positively as much as 8%. Zulfikar Firmansyah [10] in his research, entitled Analysis of public sentiment on covid-19 vaccination based on opinions on Twitter using the Naive Bayes algorithm, resulted in an accuracy rate of 78% by testing using the confusion matrix and k-fold methods. Furthermore, research conducted by Rizal Al Habsi [11] with data of 488

tweets resulted in 51.4% positive sentiment and 48.6% negative sentiment with an accuracy rate of 82.65%. Based on previous research, Naive Bayes has advantages in accuracy results. On the other hand, this algorithm has the advantage that with small training data it is able to predict the parameters needed in the classification. It has been proven with a large amount of data that Naive Bayes is able to quickly publish databases and get a high level of accuracy.

This research has a research gap with previous research. Fundamental differences are shown in specific datasets used than previous studies. Previous research used the covid-19 dataset before the booster vaccine, but this study offers a new dataset in the form of a booster vaccine during covid-19 using the naive Bayes algorithm. Based on the background described, this study aims to analyze the sentiment of the booster vaccine during covid-19 through perspectives of Indonesian netizens based on the Twitter dataset. This sentiment analysis determined whether the booster vaccine is widely accepted or not by Indonesian netizens.

II. THEORETICAL BASIS

A. Sentiment Analysis

Sentiment analysis is a process to determine sentiment along with grouping the polarity of the sentence text in order to determine the category of positive, negative or neutral sentiment [12]. This sentiment analysis is widely used for research in computer science along with the explosion of information on social media. The social media that is usually used for this sentiment analysis research is Twitter social media [13]. The impact of sentiment analysis is that the data will produce either positive or negative

B. Naive Bayes

Naive Bayes is a simple probabilistic classifier that calculates a set of probabilities by adding up the frequencies and combinations of values from a given data set [14]. The algorithm uses Bayes theorem and assumes all attributes are independent or not interdependent given by the value of the class variable. Naive Bayes predicts future opportunities based on past experiences [15]. The advantage of Naive Bayes lies in the small amount of training data used so that calculations can be carried out more quickly and efficiently [16]. The weakness of Naive Bayes is that if an error occurs in the selection of features, the level of accuracy will decrease and the calculation time will increase.

III. RESEARCH METHODS

In this study, to get the best results there are stages that are passed. The stages that are passed consist of data collection and data labeling, then the data preprocessing stage. Data that has preprocessed the data will be grouped using the Naive Bayes algorithm. The method used for the testing process is the confusion matrix. Figure 1 shows the stages carried out.
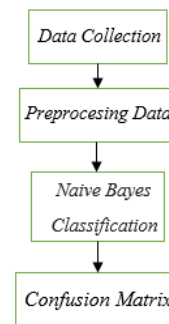

Figure 1 Research Stage

C. Data Collection

Collecting data in sentiment analysis in the form of Indonesian netizens' opinions about booster vaccines. At the data collection stage, researchers conducted by crawling twitter data using the Rapidminer application on April 4, 2022 with the keyword booster vaccine. The tweets obtained are in the form of Indonesian netizens' opinions about the booster vaccine. The data obtained is 658 data, but the data that can be processed is only 100 data which is divided into 2 parts, namely 80 training data and 20 test data.

D. Preprocesing Data

At the preprocessing stage, the data will be uniform in shape and format so that it becomes data that can be processed at a later stage. Preprocessing consists of case folding, tokenizing, filtering and stemming. The image of the preprocessing stage can be seen in Figure 2.
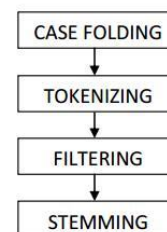

Figure 2 Data Preprocessing Stage

- *Case folding*
  This stage is a process of uniformity of letters that previously had uppercase letters changed to all lowercase letters.
- *Tokenizing*
  This stage is the stage of cutting the sentence into parts or words.
- *Filtering*
  The stage after tokenizing is the stage of eliminating words that appear in large numbers but are considered meaningless.
- *Stemming*
  The last stage in preprocessing is changing words into correct basic words

E. Classification

Naive Bayes is a simple probabilistic classifier that calculates a set of probabilities by adding up the frequencies and combinations of values from a given data set. The algorithm uses Bayes theorem and assumes all attributes are independent or not interdependent given by the value of the class variable. Naive Bayes predicts future opportunities based on past experiences.

Naive Bayes in this study is used to classify text documents. In the Naive Bayes algorithm, each document is presented with the input attribute "a1, a2, a3,,,, an" where a1 is the first word and the next until an (nth word), while V is a category label [8]. The next step is to find the highest value from the tested text category (VMAP) (McCallum, 1998). Calculate the probability value using the formula:

$P(W_k|V_j) = (n_k+1) / ($ Number of Frequency + Word Count)

$$(1)$$

Information :
P(Wk|Vj) : Probability of word weight according to category
nk : The value of occurrence of word frequency

as an example of article data for categories:
Positive: The Boster Vaccine tastes normal, doesn't it?
Neutral: Majauleng Sub-district Head Vaccines Boster in Tosora
Negative: After the booster vaccine, the side effects are like a paralyzed person

Then the number of words obtained is 18 and has the number of each frequency from each Positive category = 5, the number of frequencies in the neutral category = 5, the number of frequencies in the negative category = 8

As an example of the negative category of results after preprocessing, namely: "The booster vaccine has positive effects like covid". Calculated the probability value:

P(vaccine|negative)   = (1+1) / (8+18) = 0.076
P(boster| negative)   = (1+1) / (8+18) = 0.076
P(effects| negative)   = (1+1) / (8+18) = 0.076
P(positive| negative)   = (1+1) / (8+18) = 0.076
P(covid| negative)   = (1+1) / (8+18) = 0.076

Then it is entered into a probability table such as table 1.

Table 1. Probability of each word

Word	Probabilitas		
	Positive	Neural	Negative
Vaccine	0.086	0.086	0.076
booster	0.086	0.086	0.076
Effects	0.043	0.043	0.038
positive	0.043	0.043	0.038
Covid	0.043	0.043	0.038

After getting the word probability value in each category, then calculate the category probability using the formula:
P(Vj) = number of Documents for each Category / Total Documents

$$(2)$$

Known: number of
positive documents = 1
neutral documents = 1
negative documents = 1

So, the probability of the document:
P(positive)   = 1/3 = 0.33
P(neutral)   = 1/3 = 0.33
P(negative)   = 1/3 = 0.33

IV. RESULTS AND DISCUSSION

In this sentiment analysis research, the language used to analyze using the PHP programming language. Labeling is done manually according to the category, namely positive, negative or neutral. Furthermore, the data preprocessing process with 4 stages, namely Case folding, Tokenizing, Filtering, Stemming. The total data set that will be used is 100 tweets which are divided into 80 training data and 20 test data.

The training data that has gone through the preprocessing stage will be used as material for the test data classification process with the Naive Bayes algorithm. Testing the Naive Bayes algorithm will use the Confusion Matrix.

Table 2. Example of preprocessing Stages

Preprocessing Stages	Result
Original Sentence	*Vaksinasi ketiga atau booster ini bertujuan untuk memutus mata rantai penularan COVID-19*
Case Folding	*vaksinasi ketiga atau booster ini bertujuan untuk memutus mata rantai penularan covid-19*
Tokenizing	*"vaksinasi", "ketiga", "atau", "booster", "ini", "bertujuan", "untuk", "memutus", "mata", "rantai", "penularan", "covid-19"*
Filtering	*vaksinasi ketiga booster bertujuan memutus mata rantai penularan covid-19*
Stemming	*vaksin tiga booster tujuan putus mata rantai tular covid-19*

At preprocessing stage, the data will be processed into data that is ready to be analyzed. The data obtained from Twitter is displayed as is, so that irrelevant data

is possible so that the removal of irrelevant parts of the data is a step that must be passed so that an efficient system can be realized. There are 4 stages of preprocessing, including: case folding, tokenizing, filtering and stemming. The preprocessing stages can be seen in Table 2.

Table 3. Classification Examples

Tweet	Sentiment
Vaksin booster rasanya biasa saja	Positive
Vaksin booster astra nampol banget gila efeknya	Negative
Desa di wilayah Kabupaten Demak sampai saat ini menggiatkan vaksin ketiga booster	Neutral

Discussion

This study collected 658 data from twitter with the keyword "booster vaccine" data that can be processed only 100 data which is divided into 2 parts: 80 training data and 20 test data.
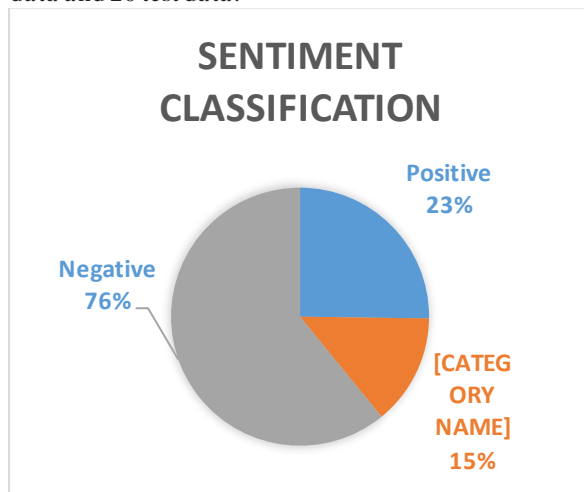

Figure 3 Pie chart Polarity of tweets

From the results of the analysis of the sentiment of the twitter user community with the topic of the booster vaccine, it leads to a negative sentiment with a percentage of 70% because this vaccine is a requirement for the 2022 munik and the effect of the booster vaccine itself.

V. CONCLUSION

The results of the analysis of public sentiment using Twitter social media with the keyword "booster vaccine" obtained positive sentiment as much as 23%, neutral sentiment as much as 15%, and negative sentiment as much as 76% with an accuracy rate of 89%. This shows that the booster vaccine is not very good according to Indonesian netizens because the booster vaccine is a requirement for going home in 2022 and the effects obtained are causing body aches, hands cannot be lifted, etc.

REFERENCES

[1] J. S. Asri and S. Wahyu, "Analisis Sentimen Menerapkan Lexicon-Learning Based Untuk Melihat Opini Masyarakat Mengenai Protokol Kesehatan Dan Perkembangan Vaksin Covid-19 Di Indonesia Menggunakan Dataset Twitter," *Konf. Nas. Ilmu Komput.*, pp. 530–536, 2021.

[2] B. Laurensz and Eko Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 2, pp. 118–123, 2021, doi: 10.22146/jnteti.v10i2.1421.

[3] E. Nurhazizah, R. N. Ichsan, and S. Widiyanesti, "Analisis Sentimen Dan Jaringan Sosial Pada Penyebaran Informasi Vaksinasi Di Twitter," vol. 10, no. 1, pp. 24–35, 2022.

[4] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," vol. 8, no. 2, pp. 100–109, 2020.

[5] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012045, 2021, doi: 10.1088/1757-899x/1088/1/012045.

[6] K. K. Agustiningsih, E. Utami, O. Muhammad, and A. Alsyaibani, "Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings," vol. 1, pp. 39–46, 2022.

[7] S. Lestari and S. Saepudin, "ANALISIS SENTIMEN VAKSIN SINOVAC PADA TWITTER MENGGUNAKAN ALGORITMA NAIVE BAYES," 2021.

[8] F. Septianingrum, J. H. Jaman, and U. Enri, "Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia dengan Metode Naive Bayes Classifier," vol. 5, pp. 1431–1437, 2021, doi: 10.30865/mib.v5i4.3260.

[9] I. W. D. Gafatia and N. Hadinata, "Analisis Pro Kontra Vaksin Covid 19 Menggunakan Sentiment Analysis Sumber Media Sosial Twitter," *J. Pengemb. Sist. Inf. dan Inform.*, vol. 2, no. 1, pp. 34–42, 2021, doi: 10.47747/jpsii.v2i1.544.

[10] Z. Firmansyah and N. F. Puspitasari, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Berdasarkan Opini Pada Twitter Menggunakan Algoritma Naive Bayes," *J. Tek. Inform.*, vol. 14, no. 2, pp. 171–178, 2021, [Online]. Available: https://doi.org/10.15408/jti.v14i2.24024.

[11]   R. Al Habsi, R. A. D. Anggoro, M. A. Valio, Y. Widiastiwi, and ..., "Analisis Sentimen Terhadap Vaksin Covid-19 di Jejaring Sosial Twitter Menggunakan Algoritma Naïve Bayes," ..., no. September, pp. 239–248, 2021, [Online]. Available: https://conference.upnvj.ac.id/index.php/sena mika/article/view/1714%0Ahttps://conference .upnvj.ac.id/index.php/senamika/article/downl oad/1714/1403.

[12]   M. I. Aditama, R. I. Pratama, K. H. U. Wiwaha, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19," *J. Inf. Eng. Educ. Technol.*, vol. 4, no. 2, pp. 90–92, 2020.

[13]   C.- Di Twitter, F. A. Hasanuddin, M. Akbar, and M. Farid, "Analisis Kluster Terpaan Selektif Dalam Jaringan Informasi Vaksin," vol. 08, no. January, pp. 713–732, 2022.

[14]   E. Prabowo and A. C. Fauzan, "Implementasi Algoritma Naive Bayes Classifier Untuk Penentuan Kelayakan Penerima Beasiswa Peningkatan Prestasi Akademik (PPA) Bagi Mahasiswa," *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 3, no. 2, pp. 165–177, 2021, doi: 10.28926/ilkomnika.v3i2.296.

[15]   J. Jtik, J. Teknologi, S. H. Ramadhani, and M. I. Wahyudin, "Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," vol. 6, no. 4, 2022.

[16]   A. C. Fauzan and S. D. Purwanto, "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Learning Vector Quantization dalam Sistem Identifikasi Boraks pada Bakso Daging Sapi," vol. 7, no. 2, pp. 43–50, 2021, doi: 10.24014/coreit.v7i2.15085.