Sentiment Analysis of Covid-19 Vaccine Tweets Utilizing Naïve Bayes

Abdurrahim Abdurrahim[1)], Lailis Syafa'ah[1)], Merinda Lestandy[2, a)]

[1]*Department of Electrical Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia*
[2]*Department of D3 Electrical Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia*

Corresponding author: [a)]merindalestandy@umm.ac.id

Abstract. COVID-19 is acknowledged as a transmitted from one person to another through contact, coughing, and sneezing. Twitter has served as one of the media outlets to raise awareness regarding COVID-19 problems. One of the government's objectives, based on the rising distribution, is pursued to preserve immunizations in stock. Hence, the vaccine information has become adequately available. However, immunization has sparked a range of reactions, including support and objection for vaccination. Attempts require a mechanism to distinguish tweets addressing immunization-related information. One notable method includes sentiment analysis, expressing a statement's negative, neutral, and positive feelings. A total of 5200 datasets were employed, with 4000 datasets classified as neutral, 300 datasets as negative, and 900 datasets as positive. The Naïve Bayes method and the TF-IDF (Term Frequency – Inverse Document Frequency) word weighting strategy are proposed to model the COVID-19 vaccine dataset, by comparing the three models of: Gaussian, Multinomial, and TF-IDF (Term Frequency – Inverse Document Frequency). According to study employing Naïve Bayes, the best model employing Bernoulli Naive Bayes is 80% with a data splitting of 30%.

INTRODUCTION

COVID-19 (Coronavirus Disease 2019) has been currently acknowledged as a disease due to a new type of coronavirus, which is Sars-CoV-2, transmitted from person to another by droplet contact (coughing and sneezing). According to data from the COVID-19 Task Force, as of December 5, 2020, 65,257,767, the confirmed positive cases of COVID-19 were 1,513,179 deaths, gathered from 220 countries, one of which was Indonesia, with a positive number of 569,707 patients and 17,589 deaths (www.covid19.go.id). Confirmed through the rapid spread of COVID-19, one of the government's measures is the provision of vaccinations with the aim of preventing the spread of COVID-19[1]. Information regarding the COVID-19 vaccination has proliferated through one of the popular media platforms such as twitter, that allows users to send characters up to 140 characters or commonly called twit or chirp[2]. Opinion presented in tweets regarding the vaccinations raise the broad range of thoughts in the community including both the supports and the objections towards vaccination [3]. One approach to navigate the public views regarding the COVID-19 vaccination includes sentiment analysis [4] to determine the content of a dataset in the form of text (documents, sentences, paragraphs, etc.) classifying: negative, positive or neutral presentation [5].

Sentiment analysis research employing machine learning has been conducted in recent years [6]–[9] and has been applied to various fields [10]–[17]. The Naïve Bayes method produces better results than those in the Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Conditional Random Field (CRF) for the sentiment analysis category [18][19]. The TF-IDF method includes both the number of occurrences of words such as BOW (Bag Of Word) and the examination of important and unimportant words from the document [20].

There are several studies which have been previously conducted to analyze public sentiment towards the COVID-19 tweet pandemic, one of which was research informing in this study that Indonesian people tend to demonstrate positive response to vaccination discourse by 30% and negative response by 26% towards the vaccination employing the Latent Dirichlet Allocation (LDA) method[1]. In accordance with Deden Ade et al.'s research on vaccine sentiment analysis (using Sinovac and Pfizer vaccine data), it is reported that views about vaccines classify the three presentations of: positive, neutral and positive; with Sinovac's results of 77% indicating positive vaccine tweets, 4% indicating neutral, 19% indicating negative. Meanwhile, Pfizer produces an accuracy of 81% positive views, 17% negative views and 3% neutral views [21]. Research on vaccines has been conducted by employing Naïve Bayes and SVM to navigate the public's views on vaccination actions. More specifically in this study, the analyzed vaccines include red, white vaccines and Sinovac vaccines with an average accuracy of 85.59% for Naïve Bayes and 84.41% for SVM[3]. Research [19] utilizes the Naïve Bayes method by adding bag of words as word weighting, generating

1st International Conference on Technology, Informatics, and Engineering
AIP Conf. Proc. 2453, 020052-1–020052-6; https://doi.org/10.1063/5.0094607
Published by AIP Publishing. 978-0-7354-4356-3/$30.00

an accuracy of 94% compared to using an ensemble feature with an accuracy of 88%. In line with previous research using the Naïve Bayes method with an accuracy of 80.90% compared to the KNN method of 73.34% and SVM of 63.99% [22]. On research [23] using the Naïve Bayes method and combining it with GA (Genetic Algorithm) produces a good accuracy of 87.50%. Method [24] produces a significant accuracy compared to KNN equal to 84.16%. This study utilizes the TF-IDF word weighting method demonstrating better results than that in BOW [20][25].

Based on the aforementioned research, the researchers in this study utilize the Naïve Bayes method with 3 models of: Gaussian, Multinomial, Bernoulli; the results would be further compared to navigate the best model and the weighting of TF-IDF words along with the twitter regarding COVID-19 vaccine data.

METHODS

This study aims to navigate the accuracy with the best model of Naïve Bayes. The dataset was obtained from the Kaggle website with a total data of 5000 data. The data was manually labeled by the annotator with 3 classifications, demonstrating negative, neutral and positive views. From figure 1, in the data pre-processing flow, a data cleaning process was conducted to eliminate unnecessary words or characters, then the tokenizing stage was employed to convert sentences into words. After the sentences were separated into words, the data would then be separated into 2, including training data and test data. The process of calculating word weights was performed by implementing the TF-IDF method. The next flow is progressed with the classification test phase implementing the Naïve Bayes method. In this model, training data is required as learning and testing data for the data testing process. The final flow was concluded by testing the performance of the 3 models (gaussian, multinomial, and Bernoulli).

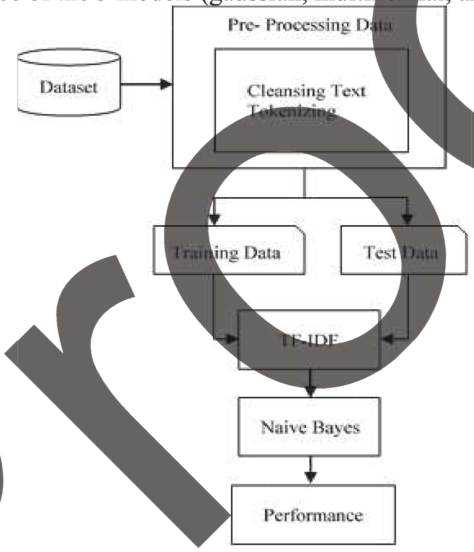

FIGURE 1. Research Stages

Dataset

The employed data is twitter vaccine data with a total of 5000 data, manually labeled with the results of the sentiments of each twitter which described in table 1.

TABLE 1. Twitter Datasets

Tweets	Sentiment
Let's succeed in the vaccination program for a better Indonesia #Vaksin #indonesiasehat https://t.co/tp5rxZae6g	Positive
Vaccination of Health Workers in Surabaya to be Completed soon #news #vaccin #surabaya #Indonesia https://t.co/sXeObfZXII	Neutral
Let's support the national vaccination program, for Indonesia to return to health, safety and prosperity. #Covid19 #VirusCorona https://t.co/us0B6ZKADe	Positive
After getting the #vaccine #Sinovac for 2 days from my chest to my throat experienced a burning sensation. https://t.co/dwr6BsCSaB	Negative

TF-IDF

TF-IDF Word Weighting is defined as a technique to assign a weight (value) to a word. There are 3 aspects in determining the weighting, which include: term frequency (TF), inverse document frequency (IDF), and normalization. In this study, TF-IDF was employed as a word weighting method. TF-IDF is determined by 2 factors, as follows:

1. Calculation of the number of occurrences of term j or term frequency with the symbol$tf_{t,d}$
2. Calculation of the number of occurrences of all documents from TF or commonly acknowledged as document frequencyidf_t

The calculation of TF*IDF is formulated in the following Equation 1.

$$W_{t,d} = Wtf_{t,d} \times idf_t \tag{1}$$

Where $Wtf_{t,d}$ is word weight in each document, $tf_{t,d}$ is number of occurrences of the term in the document, *N* is number of documents in the document set, *df* is number of documents containing term, idf_t is inverse weight of df value, and $W_{t,d}$ is TF-IDF weighting

Naive Bayes

Naïve Bayes algorithm is described as an algorithm implemented in text categorization. The basic idea of Naïve Bayes lies in the combination of word probabilities and the word category to predict possible document categories [26]. This method combines the previous data with new data resulting in a simpler method but having high accuracy [27]. The calculation of the Naïve Bayes method is illustrated in Equation 2

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2}$$

where P(H|X) is probability of hypothesis H based on condition X, X is data *training* with a recognized class (label), H is data with class (label) C, P(H) is probability of hypothesis X, P(X) is probability of X being observed, P(H|X) is probability of X, based on the conditions on the hypothesis H.

RESULTS AND DISCUSSION

Data Preprocessing

The COVID-19 vaccine tweets contain unstructured data depicting unimportant letters and symbols. This dataset requires a preprocessing stage in several stages such as cleansing which contains deletion of urls, mentions, hashtags, stop words, and tokenizing stages. The utilized libraries in this study include: hard, sklearn and pandas as well as google collab. Examples of data preprocessing are illustrated in Table 2 and Table 3.

TABLE 2. Preprocessing Results

Original Tweets	Pre-processing Results
Let's succeed in the vaccination program for a better Indonesia #Vaksin #indonesiasehat https://t.co/tp5rxZae6g	Let's succeed in the vaccination program for a better Indonesia
Vaccination of Health Workers in Surabaya to be Completed soon #news #vaccin #surabaya #Indonesia https://t.co/sXeObfZXII	Vaccination of Health Workers in Surabaya Will Soon Be Completed
Let's support the national vaccination program, for Indonesia to return to health, safety and prosperity. #Covid19 #VirusCorona https://t.co/us0B6ZKADe	Let's support the national vaccination program, for Indonesia to return to health, safety and prosperity

TABLE 3. Tokenizing Results

Preprocessing Results	Tokenizing Results
Let's succeed in the vaccination program for a better Indonesia	['let', 'us', 'success', 'kan', 'program', 'vaccination', 'for', 'indonesia', 'which', 'more', 'good']
Vaccination of Health Workers in Surabaya Will Soon Be Completed	['Vaccination', 'Energy', 'Health', 'at', 'Surabaya', 'Soon', 'Completed']
Let's support the national vaccination program, for Indonesia to return to health, safety and prosperity	['Come on', 'support', 'program', 'vaccination', 'national', 'for', 'Indonesia', 'back','healthy','safe','and', 'prosperous']

Word Weight

The results of word preprocessing would be later converted into values utilizing word weighting. This process aims to calculate the weight of each utilized value, in which more features generate more documents from the calculations at this stage acknowledged as TF (Term Frequency) and IDF (Inverse Document Frequency). TF presents the number of words that appear in each document. IDF depicts a calculation of the term or the number of documents for each word, signifying that rare appearance of word in a document generates greater IDF value [20]. Table 4 depict the result of word weighting with the TF-IDF method. The calculation of the TF and IDF values which produce more accurate values for each document.

TABLE 4. TF-IDF And TF*IDF Calculation Results

Term	TF-IDF					TF*IDF		
	D1	D2	D3	DF	IDF	D1	D2	D3
let	1	0	0	1	0.477121	0.477121	0	0
we	1	0	0	1	0.477121	0.477121	0	0
success	1	0	0	1	0.477121	0.477121	0	0
right	1	0	0	1	0.477121	0.477121	0	0
program	1	0	1	2	0.176091	0.176091	0	0.176091
vaccination	1	1	1	3	0	0	0	0
for	1	0	1	2	0.176091	0.176091	0	0.176091
Indonesia	1	0	1	2	0.176091	0.176091	0	0.176091
that	1	0	0	1	0.477121	0.477121	0	0
more	1	0	0	1	0.477121	0.477121	0	0
good	1	0	0	1	0.477121	0.477121	0	0
power	0	1	0	1	0.477121	0	0.477121	0
health	0	1	0	1	0.477121	0	0.477121	0
…	…	…	…	…	…	…	…	…
prosperous	0	0	1	1	0.477121	0	0	0.477121

Data Labeling

The total vaccine data for COVID-19 tweets is 5,200 with a category of neutral sentiments, positive sentiments and negative sentiments that showed in figure 2. The number of each category if 4000 for neutral sentiments, 900 for positive sentiments and 300 for negative sentiments

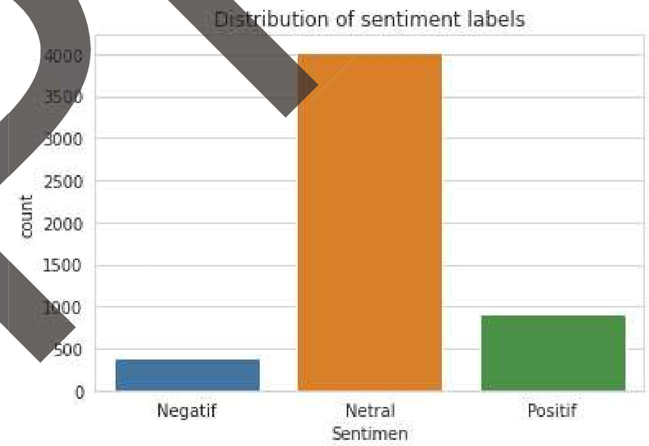

FIGURE 2. Vaccine Sentiment Label Sharing Graph

Sentiment Classification Model

In this classification model, the data is modeled as presented in the following by changing the sentiment to 0 being neutral, 1 being positive, and -1 being negative. For example, Table 5 showed one of the data modeling .

TABLE 5. Vaccine Sentiment Classification

Tweet	Sentiment
Let's succeed in the vaccination program for a better Indonesia #Vaksin #indonesiasehat https://t.co/tp5rxZae6g	1
Vaccination of Health Workers in Surabaya to be Completed soon #news #vaccin #surabaya #Indonesia https://t.co/sXeObfZXII	0
After getting the #vaccine #Sinovac for 2 days from my chest to my throat experienced burning sensation. https://t.co/dwr6BsCSaB	-1

After the data process is inserted into the model, the next step is progressed with the process of separating the data into 2 parts of data train and data testing. Training data is utilized as a model in the data learning process and then data testing is utilized to test the data. Testing the test data is split by 30% with a random size of 42 after the data separation process, the dataset would be tested obtaining the following results in table 6.

TABLE 6. Naïve Bayes Accuracy Comparison

Test Size	*Accuracy*		
	Gaussian Naïve Bayes	Naïve Bayes Multinomial	Bernoulli Naïve Bayes
0.2	0.75	0.79	0.79
0.3	0.75	0.79	0.80
0.4	0.74	0.78	0.79

CONCLUSIONS

This study tests the Naïve Bayes method with the Naïve Bayes weighting model using COVID-19 vaccine tweet data. The Naïve Bayes method produces significant accuracy values with several tests such as data testing with 3 different formats of 20%, 30% and 40%, with the best results obtained through utilizing 30% data testing with 80% accuracy results due to faster model training process using Bernoulli, compared to the preprocessing Bernoulli process shifting the value to 0 and 1 when condition 0 reflects a condition that does not have document features and vice versa. Gaussian method is hence more suitable for continuous data and Multinomial is deemed more suitable for discrete data.

ACKNOWLEDGMENTS

The author would like to express the gratitude towards the Department of Electrical Engineering, Faculty of Engineering, University of Muhammadiyah Malang, which has provided full support for the success of this research.

REFERENCES

1. F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Health Information Management Journal ISSN*, **8**, 2655–9129 (2020).
2. S. Suryono, E. Utami, and E. T. Luthfi, "Analisis Sentiment Pada Twitter Dengan Menggunakan Metode Naïve Bayes Classifier," *Seminar Nasional Geotik 2018*, 9–15 (2018).
3. B. Laurensz and Eko Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, **10**, 118–123 (2021),
4. V. K. S. Que, A. Iriani, and H. D. Purnomo, "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, **9**, 162–170 (2020),
5. M. Syarifuddin, "Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn," *Inti Nusa Mandiri*, **15**, 23–28 (2020).
6. C. R. Aydın, T. Güngör, and A. Erkan, "Generating Word and Document Embeddings for Sentiment Analysis," in *Proceedings for CICLing 2019*, no page (2020).
7. A. F. Anees, A. Shaikh, A. Shaikh, and S. Shaikh, "Survey Paper on Sentiment Analysis : Techniques and Challenges," *EasyChair*, 2516–2314 (2020).
8. R. P. Mehta, M. A. Sanghvi, D. K. Shah, and A. Singh, "Sentiment analysis of tweets using supervised learning algorithms," in *Advances in Intelligent Systems and Computing*, **1045**, 323–338 (2020).

9. D. S. Sisodia, S. Bhandari, N. K. Reddy, and A. Pujahari, “A Comparative Performance Study of Machine Learning Algorithms for Sentiment Analysis of Movie Viewers Using Open Reviews,” Springer, 107–117 (2020).
10. M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, and G. Fortino, “A hybrid deep learning model for efficient intrusion detection in big data environment,” *Information Sciences*, **513**, 386–396 (2020),
11. M. M. Hassan, A. Gumaei, S. Huda, and A. Almogren, “Increasing the Trustworthiness in the Industrial IoT Networks Through a Reliable Cyberattack Detection Model,” *IEEE Transactions on Industrial Informatics*, **16**, 6154–6162 (2020),
12. M. Alqahtani, A. Gumaei, H. Mathkour, and M. M. Ben Ismail, “A genetic-based extreme gradient boosting model for detecting intrusions in wireless sensor networks,” *Sensors (Switzerland)*, **19**, (2019),
13. A. Gumaei, M. M. Hassan, M. R. Hassan, A. Alelaiwi, and G. Fortino, “A Hybrid Feature Extraction Method With Regularized Extreme Learning Machine for Brain Tumor Classification,” *IEEE Access*, **7**, 36266–36273 (2019),
14. A. Gumaei, M. M. Hassan, A. Alelaiwi, and H. Alsalman, “A Hybrid Deep Learning Model for Human Activity Recognition Using Multimodal Body Sensing Data,” *IEEE Access*, **7**, 99152–99160 (2019),
15. A. Gumaei, R. Sammouda, A. M. S. Al-Salman, and A. Alsanad, “An Improved Multispectral Palmprint Recognition System Using Autoencoder with Regularized Extreme Learning Machine,” *Computational Intelligence and Neuroscience*, **2018**, (2018),
16. A. Gumaei, R. Sammouda, A. M. S. Al-Salman, and A. Alsanad, “Anti-spoofing cloud-based multi-spectral biometric identification system for enterprise security and privacy-preservation,” *Journal of Parallel and Distributed Computing*, **124**, 27–40 (2019),
17. A. Gumaei, R. Sammouda, A. Al-Salman, and A. Alsanad, “An Effective Palmprint Recognition Approach for Visible and Multispectral Sensor Images,” *Sensors*, **18**, 1575 (2018),
18. A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 1320–1326 (2010),
19. R. I. Permatasari and M. A. Fauzi, “Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes,” *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, 92–95 (2018).
20. Imamah and F. H. Rachman, “Twitter sentiment analysis of Covid-19 using term weighting TF-IDF and logistic regresion,” *Proceeding - 6th Information Technology International Seminar, ITIS 2020*, 238–242 (2020),
21. D. A. Nurdeni, I. Budi, and A. B. Santoso, “Sentiment Analysis on Covid19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer,” *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, 122–127 (2021),
22. M. Wongkar and A. Angdresey, “Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter,” *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 1–5 (2019),
23. S. Ernawati, “Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies,” *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1–5 (2018),
24. M. Lestandy, L. Syafa, and A. Faruq, “Klasifikasi Pendonor Darah Potensial Menggunakan Pendekatan K-Nearest Neighbors dan Naïve Bayes Classification of Potential Blood Donors Using K-Nearest Neighbors and Naïve Bayes Approach,” 2–7,
25. V. L. Nguyen, D. Kim, V. P. Ho, and Y. Lim, “A new recognition method for visualizing music emotion,” *International Journal of Electrical and Computer Engineering*, **7**, 1246–1254 (2017),
26. Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, “Sentiment classification of Internet restaurant reviews written in Cantonese,” *Expert Systems with Applications*, **38**, 7674–7682 (2011),
27. E. W. Sandi Fajar Rodiyansyah, “Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification,” *Indonesian Journal of Computing and Cybernetics Systems*, **2**, 3–33 (2010),