Sentiment Analysis on Covid-19 Vaccination in Indonesia Using Support Vector Machine and Random Forest

I Made Sumertajaya[1], Yenni Angraini[2], Jamaluddin Rabbani Harahap[3], Anwar Fitrianto[4]

[1,2,3,4]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia*

[1]imsjaya@apps.ipb.ac.id, [2]y_angraini@apps.ipb.ac.id,
[3]jamaluddin_rabbani@apps.ipb.ac.id, [4]anwarstat@gmail.com

Abstract - World Health Organization (WHO) stated Covid-19 as a global pandemic in March, 2020. This pandemic has influenced people's life in many sectors such as the economy, health, tourism, and many more. One way to end this pandemic is to make herd immunity obtained through the vaccination program. This program still raises pros and cons at the beginning of its implementation in Indonesia. Many people doubt the safety and side effects of the vaccine. There are also pros and cons to vaccination programs in social media such as Twitter. This platform generates a huge amount of text data containing people's perceptions about vaccines. This research aims to predict sentiment using supervised learning such as support vector machine (SVM) and random forest and capture sentiment about vaccines in Indonesia in the first two weeks of the program. The result shows SVM was a better model than random forest based on the precision and F1-score metrics. The SVM approach produces a precision value of 0.50, a recall of 0.64, and an F1-score of 0.52. In the study, it was also found that tweets with neutral sentiment dominated the twitter user sentiment in the study period. Tweets with negative sentiment decreased after the first week of the COVID-19 vaccination program.

Keywords: Coronavirus, Vaccination, Support Vector Machine, Random Forest, Twitter, Linear Discriminant AnalysisDA

I. INTRODUCTION

A new type of coronavirus was discovered in Wuhan City, China, in December 2019. This virus causes coronavirus disease 2019 (Covid-19) by the World Health Organization (WHO). Coronavirus causes respiratory disease in humans and the severity varies for each individual [1-3]. This virus has been increasingly spreading to various regions in China and other countries due to many residents from Wuhan City traveling during the Chinese New Year. The spread of the virus is very fast and increasingly widespread, causing WHO to designate Covid-19 as a global pandemic in early March 2020.

Various countries worldwide set restrictions on people's mobility to suppress the spread of the virus. These restrictions impact various sectors such as tourism, the economy, and education. All countries worldwide are competing to be free from the pandemic to restore life in various sectors. One way to stop a pandemic is to establish community immunity. This can be formed if most of the population already has resistance to the coronavirus. Immunity can be created through natural infection with viruses or through immunization with vaccines, [4]. Indonesia is one of the countries that launched the Covid-19 vaccination program. This program still raises pros and cons because many still doubt the safety and side effects of the vaccine. One of the vaccines used in Indonesia is Sinovac, which has been clinically tested in Bandung, involving 1,600 volunteers. The Sinovac vaccine has an efficacy of 65.3%. This figure is still lower than other vaccines such as Pfizer and Moderna.

Pros and cons also occur on social media, one of which is Twitter. The CEO of Twitter stated that the number of daily active users of this platform worldwide in May 2020 reached 166 million users. This makes Twitter a potential data source for analysis [5-6], one of which is sentiment analysis. It is used to observe people's opinions, emotions, and attitudes in the form of texts. The sentiment analysis on Twitter has been applied in many applications, such as in [7-10].

Meanwhile, several studies have been conducted for comparing several methods in classification problems [11-12]. Related to Covid-19, a study was conducted about a sentiment analysis on Covid-19 vaccination using Twitter data [13]. They used naive Bayes and support vector machine (SVM) with 845 tweets. The sample size is an essential factor in the case of sentiment analysis so that the conclusions generated can represent actual conditions. This study will use as many as 14,307 tweets about the Covid-19 vaccination in Indonesia. A comparison of the SVM and Naive Bayes methods to classify product review sentiment in a marketplace was carried out by [14] As a result, the SVM method has better performance than Naive Bayes. In addition, the algorithm that can be used in sentiment analysis is random forest. Research conducted by [15] on political orientation in Indian elections shown random forest has better performance than the other three algorithms used. This study aims to compare the performance of the SVM and random forest methods and see an overview of sentiment trends on Covid-19 vaccination in Indonesia. Similar research on haven been conducted

II. METHOD

A. Data

An account to get API access from Twitter via the https://apps.twitter.com/ link needs to be created before extracting data, a. The keywords used at the time of data collection were “vaksinasi Covid”, “vaksin sinovac”, “vaksin pfizer”, “vaksin astrazeneca”, “vaksin covax”, dan “vaksin novavax”. The data collection period starts from January 15, 2021 to January 28, 2021. The reason for choosing this timeframe is because January 13, 2021 is the first day of the COVID-19 vaccination program in Indonesia, so the pros and cons of vaccines are still a hot issue being discussed on Twitter. Details of the variables during data collection are in Table I.

B. Analysis Procedures

The data analysis procedure is described in the flow chart of Fig. 1.

Below is the description of each procedure in the flow chart:

1) Data Extraction: This step includes extracting tweets related to the covid vaccination from the web.

2) Data Sorting: The data sorting process involves deleting tweets in non-Indonesian languages, deleting accounts suspected of being bots, deleting redundant tweets, and deleting irrelevant tweets. Irrelevant tweets include news of vaccination programs in other countries, tweets that only contain links about vaccines, and usernames that contain keywords but do not tweet about vaccines. Factors that characterize an account as a bot or not are the age of the account [16] and a username that contains a mixture of numbers and letters. This process reduces the data so that it leaves 9,587 tweets.

3) Data Labeling: It is conducted manually on 3,000 tweets selected randomly and proportionally according to the number of tweets each day. The reason for manual labeling of data is because there are many tweets containing news of optimism about vaccines. This tweet is classified as neutral, not positive. In addition, some tweets contain sarcasm, so the author's consideration is needed in the data labeling process. There are three types of sentiment: positive, negative, and neutral. Positive sentiments are tweets that express support, negative are tweets that express contra, while neutral only contains news or information on vaccination programs.

4) Data Pre-processing: The process includes word normalization, replacing outlier values on numeric variables with quartile 1 or quartile 3 values, lowercasing (making all letters non-capital), stemming (converting words into basic words), removing links, mentions, hashtags, numbers, stopwords, and punctuation marks. The list of words defined as stopwords is obtained from a combination of the Sastrawi and NLTK modules in the python software. Next is forming a document term matrix where the *j*-th word will be worth 1 if it is contained in the *i*-th tweet, whereas if it is not it will be 0.

TABLE I
DETAILS OD THE DATA EXTRACTION VARIABLES

No.	Variables	Notes
1	Tweet	Tweet content
2	Username	Unique username
3	Followers	The number of followers at the time of data extraction
4	Verified	Binary variables that indicate the account has been verified by Twitter or not
5	Account creation time	Account creation time
6	Tweet creation time	Tweet creation time

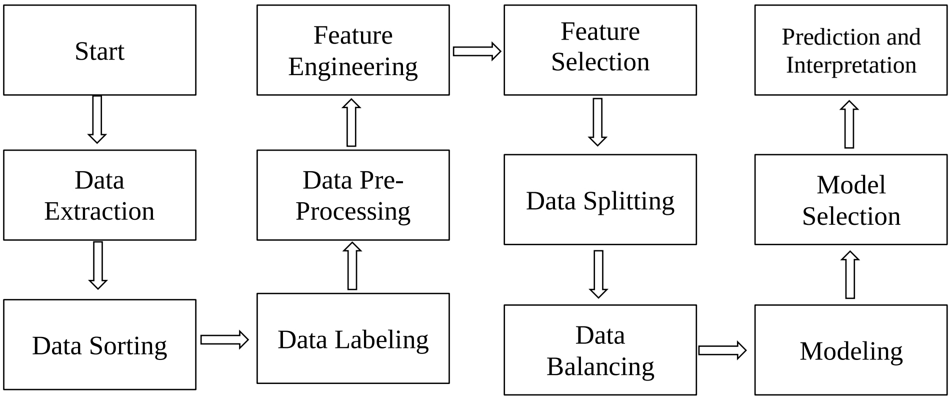


Fig. 1 Flow chart of the data analysis procedures

5) Feature Engineering: This stage adds new relevant variables and is expected to improve model performance. Six variables are added: the number of hashtags, mentions, links in each tweet, the proportion of numbers in the username, account age, and topics grouped using the Latent Dirichlet Allocation (LDA) method.

6) Feature Selection: There are 26,001 explanatory variables after the feature engineering process. Large data dimensions will slow down the computational process and not all variables are relevant in predicting sentiment. This study selects variables that have information value (IV) > 0.02.

7) Data Splitting: 67% of the data is used as training data while the remaining 33% is used as test data. The machine learning algorithm will study the data using training data, and evaluate its performance using test data.

8) Data Balancing: This process is carried out because the number of observations in one class is very small compared to observations in other classes. This method is applied to training data so that the number of observations is evenly distributed for each sentiment. Synthetic Minority Oversampling Technique (SMOTE) is used in this process.

9) Modeling: The step uses 5-fold stratified cross-validation to select the best hyperparameters for the SVM and random forest methods.

10) Model Selection: Selecting the best models among a few alternatives that have been founf in the previous.

11) Prediction and Interpretation.

III. RESULTS AND DISCUSSION

All the beginning of the paragraph is written indented. All sentences in a paragraph must be aligned right and left.

A. Feature Selection

Information value (IV) is used to select the variables used in the modeling. The IV will measure the predictive power of the variable in distinguishing tweets with neutral sentiment from tweets that have sentiment (negative and neutral). According Table II, variables having IV < 0.02 is categorized as having no predictive power on the response variable, [17]. Therefore, IV > 0.02 is used as a standard to select the variables to be used in modeling.

A total of 178 of 26,001 variables have IV> 0.02. The list of variables with these criteria is in Table III.

TABLE II
CRITERIA FOR THE PREDICTIVE POWER OF EXPLANATORY VARIABLES

Information Value (IV)	Predictive Strength
< 0.02	None
0.02 ≤ IV < 0.1	Weak
0.1 ≤ IV < 0.3	Medium
≥ 0.3	Strong

TABLE III
INFORMATION VALUE OF THE INDEPENDENT VARIABLES

No	Variables	Information Value
1	*Kita*	0.24
2	*Halal*	0.21
...	....	....
26000	*Tidak benar*	0.00
26001	*Panjang*	0,00

B. *Data exploration*

Fig. 2 is a pie chart of 3,000 manually labeled tweets. The chart shows unbalanced data in which most tweets have neutral sentiment (81.4%), while negative and positive sentiments are 4.2% and 14.4%, respectively.

A density plot was used to see the data distribution with a numerical scale. The *y*-axis indicates the probability that a value will appear in the data. The higher the peak, the more concentrated the value is in that range of values. Fig. 3 (d) shows no difference in the distribution of the value of the proportion of digits in the accounts of each sentiment. Whether the accounts provide positive, negative, or neutral sentiments, most accounts have a digit percentage on their account names below 7.5%. Tweets with negative sentiments tend not to include hashtags in their tweets, while positive and negative sentiments tend to include hashtags more, as shown in Fig. 3 (a). The number of mentions was different for negative sentiments, where tweets mention more to several accounts than positive or neutral sentiments, as shown in Fig. 3 (b). Fig. 3 (c) shows no accounts list more than one link on negative sentiment, while tweets with positive or neutral sentiment are the opposite.

The LDA method has been used to find topics contained in tweets. Word cloud plots have been used as a medium for interpreting the topic being discussed. Fig. 4 (a), Fig. 4 (b), and Fig. 4 (c) show that the first topic discusses a lot about the vaccination program in Indonesia, the second topic discusses the results of clinical trials and side effects of vaccines, and the third topic talks a lot about the implementation of vaccination. The topic of clinical trial results and vaccine side effects was the most discussed topic with a percentage of 42%. Another topic, namely the implementation of vaccination and vaccination programs in Indonesia, received 33% and 35% of the talk.

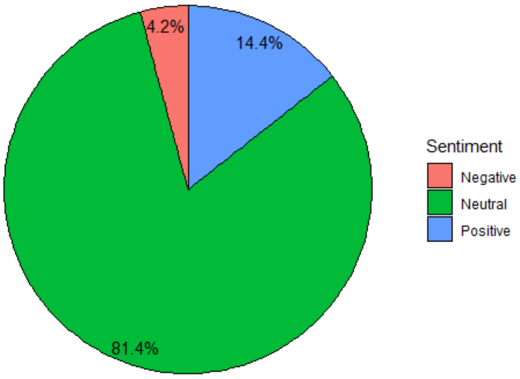
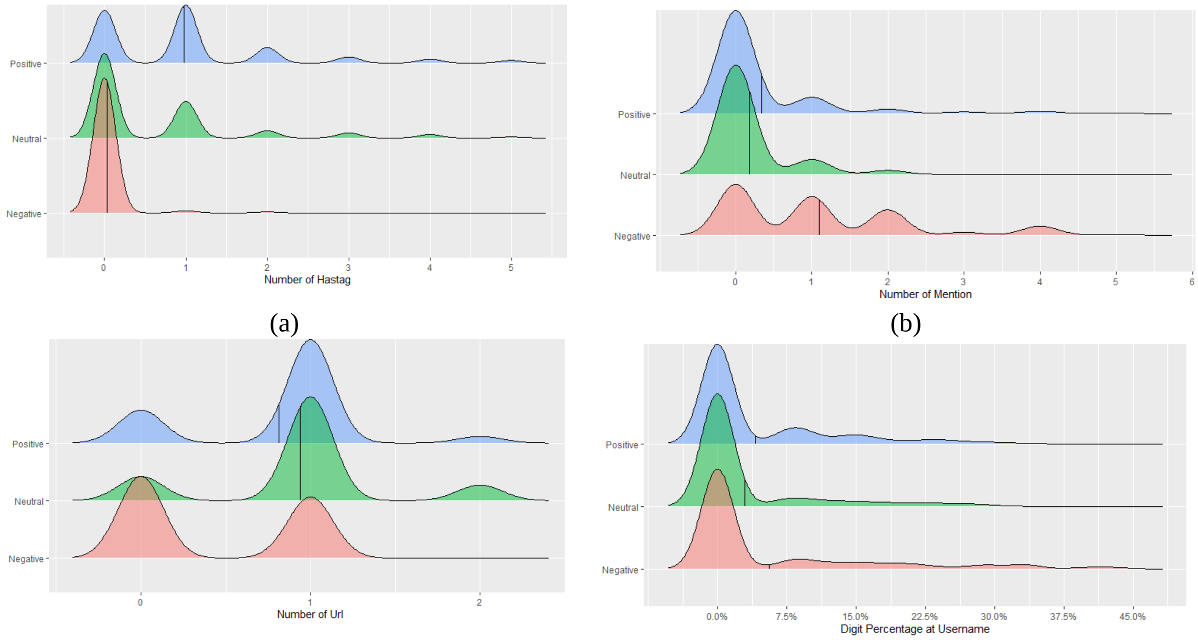


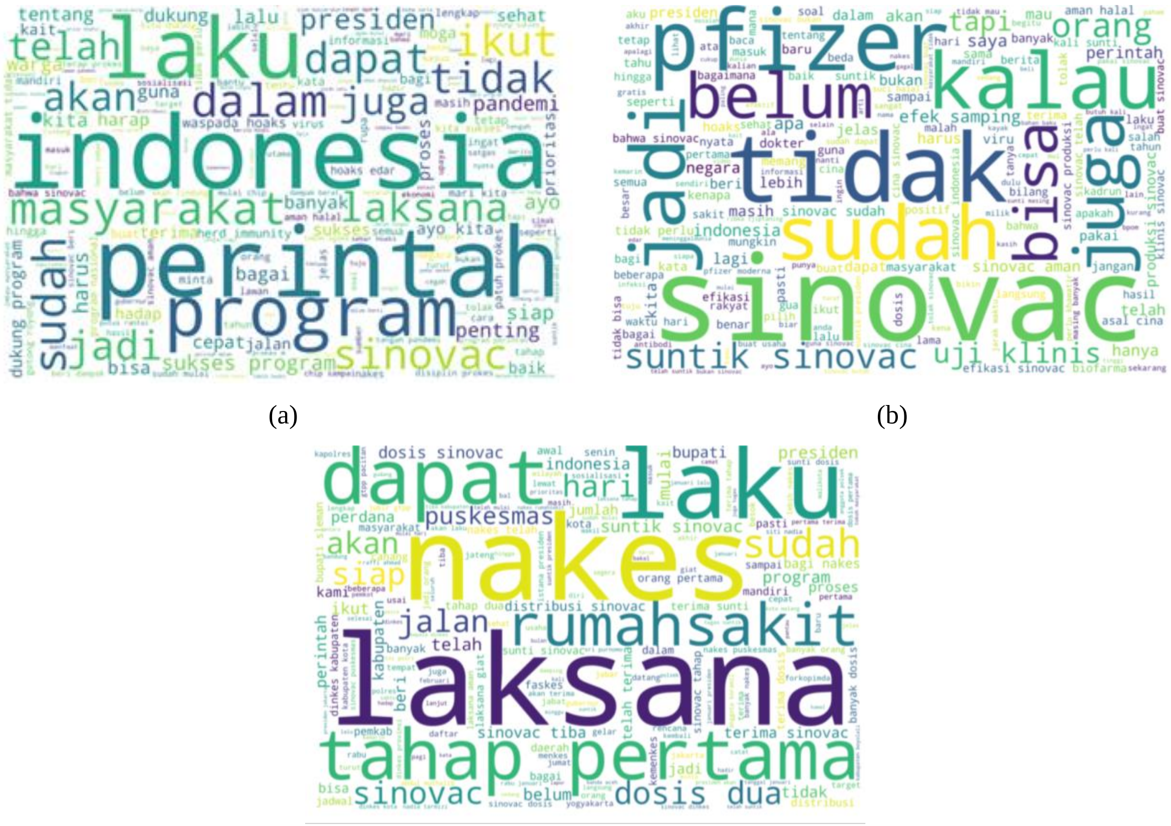
Fig. 2 Pie chart of percentages of vaccination sentiment



(a) (b)

(c) (d)

Fig. 3 (a) The number of hashtags (b) the number of mentions (c) the number of links (d) the percentage of digits in the account name



(a)

(b)

(c)

Fig. 4 Word cloud (a) first topic (b) second topic (c) third topic

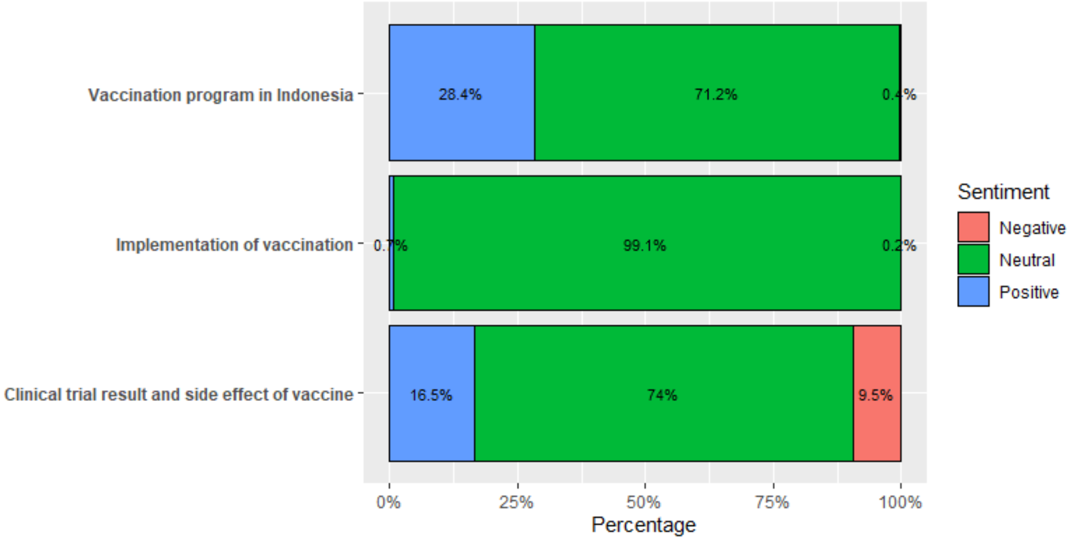


Fig. 5 Sentiment percentage for each topic

Each topic has different sentiments as shown in Fig. 5. The topic of clinical trial results and vaccine side effects have become the topic with the highest percentage of negative sentiment compared to other topics. While the topic of support and the implementation of vaccination, the majority of his tweets have a neutral sentiment.

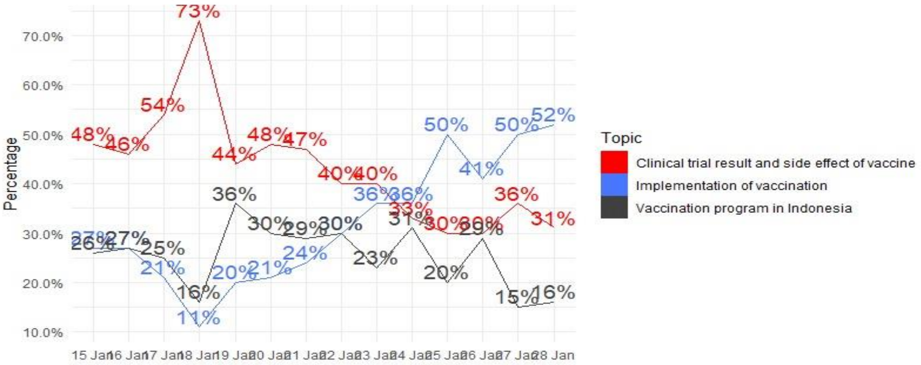


Fig. 6 The development of the topic of vaccination

Fig. 6 shows the most discussed topics in the first week of the vaccination program, namely the results of clinical trials and the side effects of vaccines. The percentage of conversations on this topic has been steadily decreasing after the first week of the vaccination program (January 20, 2021). On the other hand, the implementation of vaccination has an increasing trend because news has been filled with news regarding the distribution of vaccines and the injection of the first dose of vaccine to health workers.

C. Modeling

Two models were compared, namely SVM and random forest to produce the best model. Machine learning models have issues when dealing with imbalanced data. The resulting model has good performance in the majority class but not in the minority class. According to [18], balancing the training data will positively impact the model's performance on the test data. The technique used to perform data balancing in this study is SMOTE.

The SVM and random forest methods have settings that must be set first, called hyperparameters. The right combination of hyperparameters will improve the performance of the model. The 5-fold stratified cross-validation method was used to select the best hyperparameters for each method. Four hyperparameters were used in the random forest method, including the depth of each tree, the minimum number of samples at the node, the minimum number of samples for separation, and the number of variables used for each tree. A total of 250 combinations of hyperparameters were tried in the random forest method with the optimal value occurring at tree depth of 6, the maximum number of variables used for each tree was 13, the minimum number of samples for separation of 5, and the minimum number of samples at node 2. In the research, a linear kernel with one parameter was used in the SVM method, namely the regularization parameter (C). This value sets the amount of regularization when forming margins, where a smaller C value will make the margins on the hyperplane wider. This study uses as many as 40 possible C values , with the most optimal value found when the C value is 0.00001.

D. Model Selection

Accuracy is a commonly used measure of model performance. However, this measure is not suitable for use in imbalanced data. The model performance will be good in the majority class and the accuracy calculation will give greater weight to the majority class. This situation may lead to misinterpretation and misclassification. This study used precision, recall, and F1-score to choose the best model.

Table IV shows that the SVM method performs better than random forest on precision and F1-score sizes. Therefore, the method that will be used to predict tweets that do not yet have a label is SVM. The SVM model is good at predicting neutral sentiment. However, it is still not good at predicting negative or positive sentiment. There were many misclassifications for negative and positive sentiments in the confusion matrix Table V).

E. Prediction

There were 6,587 tweets that do not yet have a label will be predicted using the SVM method. The line chart has been used to see the sentiment trend on January 15, 2021-January 28, 2021 after the prediction process (Fig. 7). Neutral sentiment tweets will dominate the tweets throughout prediction period. Meanwhile, positive sentiment tweets were slightly higher than negative tweets.

TABLE IV
PERFORMANCE OF RANDOM FOREST AND SVM

Sentiment	Random forest			SVM		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Negative	0.20	0.76	0.32	0.16	0.57	0.25
Neutral	0.93	0.67	0.78	0.92	0.72	0.81
Positive	0.34	0.60	0.43	0.43	0.62	0.51
Mean	0.49	0.68	0.51	0.5	0.64	0.52

TABLE V
CONFUSION MATRIX OF THE SVM ON TESTING DATA

Actual class	Predicted class		
	Negative	Neutral	Positive
Negative	24	15	3
Neutral	110	588	116
Positive	18	36	90

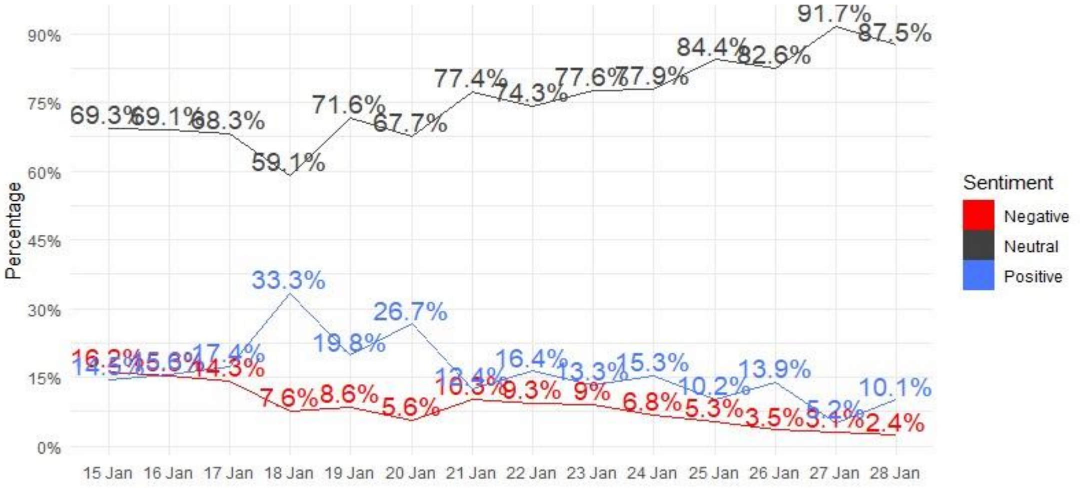


Fig. 7 Line chart of vaccination sentiment percentage

IV. CONCLUSION

The best-performing method in classifying Twitter users' sentiments towards the COVID-19 vaccination program was SVM. This method produces a precision value of 0.50, a recall of 0.64, and an F1-score of 0.52. The trend of twitter user sentiment in the range of January 15, 2021-January 28, 2021 was dominated by tweets with neutral sentiment. Tweets with negative sentiment decreased after the first week of the COVID-19 vaccination program. The topic widely discussed in the negative sentiment was the outcome of the Sinovac vaccine clinical trial and the recommendation to the government to use a vaccine with higher efficacy such as Pfizer. The downward trend in negative sentiment was accompanied by growing news that Pfizer vaccine requires storage. It is less practical because it must be stored below -70°C and Pfizer vaccine manufacturers are asking for legal freedom if there are side effects after using the vaccine. Meanwhile, tweets with positive sentiments fluctuated during the first two weeks of the vaccination program.

REFERENCES

[1] F. He, Y. Deng, and W. Li, "Coronavirus disease 2019: What we know?," *Journal of medical virology*, vol. 92, no. 7, pp. 719–725, 2020.

[2] S. Law, A. W. Leung, and C. Xu, "Severe acute respiratory syndrome (SARS) and coronavirus disease-2019 (COVID-19): From causes to preventions in Hong

Kong," *International Journal of Infectious Diseases*, vol. 94, pp. 156–163, 2020.

[3] N. A. Bakar and S. Rosbi, "Effect of Coronavirus disease (COVID-19) to tourism industry," *International Journal of Advanced Engineering Research and Science*, vol. 7, no. 4, pp. 189–193, 2020.

[4] H. E. Randolph and L. B. Barreiro, "Herd immunity: understanding COVID-19," *Immunity*, vol. 52, no. 5, pp. 737–741, 2020.

[5] C. Rosales Sánchez, M. Craglia, and A. K. Bregt, "New data sources for social indicators: the case study of contacting politicians by Twitter," *International journal of digital earth*, vol. 10, no. 8, pp. 829–845, 2017.

[6] J. M. Banda *et al.*, "A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration," *Epidemiologia*, vol. 2, no. 3, pp. 315–324, 2021.

[7] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine learning-based sentiment analysis for twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, 2018.

[8] S. A. El Rahman, F. A. AlOtaibi, and W. A. AlShehri, "Sentiment analysis of twitter data," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–4.

[9] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *Journal of Computational Science*, vol. 36, p. 101003, 2019.

[10] M. R. Huq, A. Ali, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 19–25, 2017.

[11] A. Afiyati, A. Azhari, A. K. Sari, and A. Karim, "Challenges of Sarcasm Detection for Social Network: A Literature Review," *JUITA: Jurnal Informatika*, vol. 8, no. 2, pp. 169–178, 2020.

[12] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial intelligence review*, vol. 52, no. 3, pp. 1495–1545, 2019.

[13] B. Laurensz and E. Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 10, no. 2, pp. 118–123, 2021.

[14] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, "A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews," in *2020 International Conference on Contemporary Computing and Applications (IC3A)*, 2020, pp. 217–220.

[15] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of political sentiment orientations on twitter," *Procedia Computer Science*, vol. 167, pp. 1821–1828, 2020.

[16] P. G. Pratama and N. A. Rakhmawati, "Social Bot Detection on 2019 Indonesia President Candidate's Supporter's Tweets," *Procedia Computer Science*, vol. 161, pp. 813–820, 2019.

[17] G. Saposnik *et al.*, "Factors associated with the decision-making on endovascular thrombectomy for the management of acute ischemic stroke," *Stroke*, vol. 50, no. 9, pp. 2441–2447, 2019.

[18] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.