Journal of Applied Data Sciences
Vol. 2, No. 3, September 2021, pp. 102-108

Application of the Vector Machine Support Method in Twitter Social Media Sentiment Analysis Regarding the Covid-19 Vaccine Issue in Indonesia

Riyanto [1,*], Abdul Azis [2]

Department Information Systems, Universitas Amikom Purwokerto, Indonesia
riyanto@amikompurwokerto.ac.id [1,*]; Abdazis9@amikompurwokerto.ac.id [2]
* corresponding author

(Received July 5, 2021 Revised August 13, 2021 Accepted August 28, 2021, Available online September 29, 2021)

Abstract

According to the Indonesian government, Indonesia has been afflicted by Covid-19 since March 2, 2020. Numerous countries, including Indonesia, have made efforts, but with the spread of perceptions, rumors, and a flood of information into the society regarding vaccines, there are both advantages and disadvantages to vaccines. government-led immunization campaigns. As a result, it is vital to examine public sentiment toward the government's immunization programs. The goal of this study is to ascertain the emotion toward the Covid-19 vaccination in Indonesia based on the classification results. The Support Vector Machine classification technique was employed in this investigation (SVM). The SVM classification method was chosen because it possesses the ability to generalize when it comes to identifying a pattern, excluding the data used in the method's learning phase. Classification with an SVM linear kernel and TF-IDF weighting, as well as data sharing via K-fold cross validation with a value of k=10. Positive and negative classifications are made. Following preprocessing and classification, we determined the f1 values, accuracy, precision, and recall to use as reference values when evaluating the classification. SVM performed well in classifying the data in this investigation, with f1 = 88.7%, accuracy = 84.4%, precision = 86.2%, and recall = 97%. This value is acceptable, and hence SVM is suitable for identifying sentiment data about the Covid-19 vaccine in Indonesia. Additional study can be conducted with richer tweet data, more thorough preprocessing, and comparison to other classification algorithms to obtain a higher categorization evaluation score.

Keywords: SVM, Data Mining, Sentiment Analysis, Vaccine Issue, Twitter

1. Introduction

At the start of 2020, the globe was stunned by the outbreak of an illness that quickly spread to over 190 countries and territories. This outbreak was dubbed coronavirus disease 2019 (Covid-19) and was caused by Coronavirus-2 Severe Acute Respiratory Syndrome (SARSCoV-2) [1]. The World Health Organization (WHO) has classified this virus outbreak a pandemic since March 12, 2020. The government's participation in policy formulation is critical in containing this pandemic. Therefore, rapid intervention is required not just in terms of adopting health protocols, but also in terms of other effective interventions to break the chain of disease transmission, specifically through vaccination initiatives [2]. Various countries, including Indonesia, have made attempts, but with the spread of perceptions, myths, and a deluge of information about vaccines, there are both advantages and disadvantages to the government's vaccination initiatives. As a result, it is vital to study attitude toward the government's immunization policies in order to ascertain the trend of sentiment against the Covid-19 vaccine in Indonesia [3].

According to Central Statistics Agency data issued in 2020, around 82.66 percent of urban and rural residents aged five and over use the internet for social media reasons. Social media is a method for swiftly and readily sharing information. Twitter is a social media platform that is distinguished by its unique traits and formats, which incorporate distinctive symbols or regulations [4]. Twitter users can only send and read blog entries with a maximum

character count of 140; these messages are referred to as tweets. Users' tweets take on a variety of forms, including opinions, facts, suggestions, and criticism of something.

Nisa [6] demonstrates that Twitter is the most effective way to gauge people's thoughts and emotional states. Nearly a billion people have accounts, and they tweet at a rate of approximately 6000 messages every second. This massive amount of mini-messages has generated a sea of data that scientists can utilize to gain a better understanding of human behavior [7]. The Support Vector Machine classification technique was employed in this investigation (SVM). The SVM classification method was chosen because it possesses the ability to generalize when it comes to identifying a pattern, excluding the data used in the method's learning phase [8]. SVM is a classification approach that has an advantage over other methods in that, in addition to employing distance as a determinant, it also employs vectors as a condition, resulting in a higher accuracy than other methods for entities on certain themes.

2. Literature Review

2.1. Sentiment Analysis

Sentiment analysis is a subfield of text mining research that is useful for identifying text documents as opinions based on their sentiment. Sentiment analysis can be used to provide a positive, negative, or neutral value to an individual's opinion on various issues contained inside a tweet [9]. Sentiment analysis, also known as opinion mining, is the process of analyzing, comprehending, processing, and extracting textual data in the form of views on entities in order to get information.

2.2. Term Frequency - Inverse Document Frequency (TF-IDF)

The TF-IDF method is a technique for quantifying the link between a word (term) and a document. This method combines two weighting principles, namely the frequency of occurrence of a term in a given document and the inverse frequency of documents containing that word. The frequency with which a term appears in a document shows the document's importance [10]. The frequency with which the word appears in papers demonstrates its popularity. Thus, the weight of the relationship between a word and a document will be greater if the term occurs frequently in the document and the document collection as a whole has few instances of the word [11]. The TF-IDF value is calculated by multiplying the two equations as follows:

$$tf = 0.5 + 0.5\, x \frac{tf}{max(tf)}$$

$$idf_t = log\left(\frac{D}{dft}\right)$$

$$W_{d.t} = tf_{d.t}\, x\, IDF_{d.t}$$

2.3. Support Vector Machine (SVM)

SVM is a machine learning (supervised learning) classification technique that predicts classes based on models or patterns learned during the training process [12,18]. Classification is accomplished by identifying a hyperplane or decision boundary that separates one class from another, which in this example helps distinguish positive sentiment tweets (labeled +1) from negative sentiment tweets (labeled -1). SVM utilizes support vectors and margin values to find hyperplane values [13].

Figure 2.1 illustrates the SVM approach by depicting the spread of data in red (box) and yellow (circle). The data in red is classified as belonging to class -1, whereas the data in yellow is classified as belonging to class +1. The primary classification difficulty is determining the dividing hyperplane between two classes [14]. This study uses input data that has been transformed into a vector representation via the weighting method. When data training is

applied to the SVM classification, a value or pattern is generated that is employed in the testing phase for the SVM process, which tries to classify the sentiment in a tweet [15].

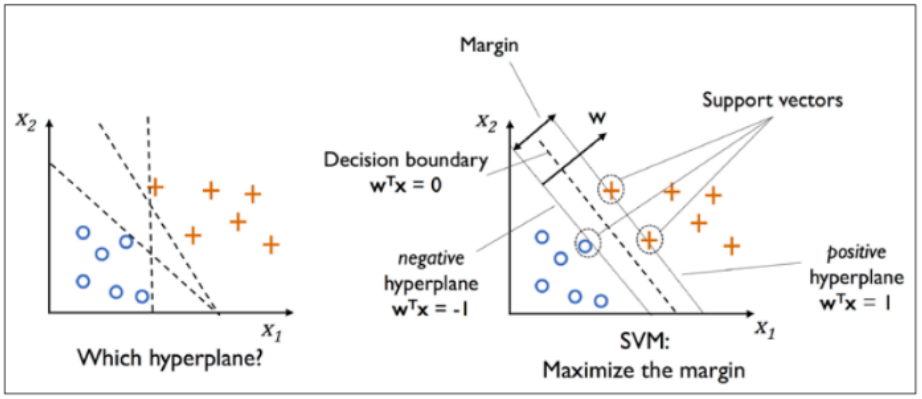

Figure. 1. SVM Model

2.4. K-Fold Cross Validation

K-fold cross validation is a data sharing approach that is frequently used to measure the performance or quality of a model. Its goal is to obtain a good training data model [16]. The dataset will be separated into k subsets of the same size and an experiment will be conducted with one subset of data serving as test data and the other serving as training data. K-Fold Cross Validation divides the data into partitions or folds with the same class size (K1;K2;:Kk). Training and submissions are repeated k times; throughout iterations, K1 becomes test data, K2 becomes training data, K3 becomes test data, and so on [17].

2.5. Confusion Matrix

The confusion matrix, alternatively referred to as the classification matrix, is a visual aid used in the process of supervised learning. The confusion matrix contains the number of correctly identified cases and the number of wrongly classified cases. Evaluation or testing is performed on test data derived from previously learned data and represented as a confusion matrix [18]. The accuracy value of the evaluation indicates the classification system's effectiveness. The higher the accuracy score, the more accurate the classification system, since in this study, which uses two classes, positive and negative, the confusion matrix table is in the form of a 2x2 matrix with predict and actual classes.

Table. 1. Confusion Matrix 2x2

		Prediction	
		Class A	Class B
Actual Class	Class A	AA	AB
	Class B	BA	BB

Class A is a representation of sentiment with a positive class and class B for sentiment with a negative class. Accuracy is a value that describes how accurately the model can classify correctly. Accuracy is the ratio of correct

predictions (positive and negative) to the overall data, in other words, accuracy is the degree to which the predicted value is close to the actual (true) value [19]. The following is an equation for calculating the accuracy value:

$$Accuracy = \frac{AA + BB}{AA + AB + BA + BB}$$

Classification evaluation can also be done by looking at the value of precision and recall. Precision is the probability that a selected item is relevant and describes the level of accuracy between the requested data and the prediction results provided by the model [20]. Precision is the ratio of positive correct predictions to the overall positive predicted outcome. Of all the positive classes that have been correctly predicted, how many data are truly positive. The precision value of each class can be obtained as follows:

$$Precision\ i = \frac{Ai}{Ai + Bi}$$

The amount of precision can be obtained by the number of precisions for each class, then divided by the number of classes. While recall is the ratio of the selected relevant items to the total number of relevant items. Recall describes the success of the model in retrieving information [21]. Thus, recall is the ratio of true positive predictions compared to the overall data that are true positive. The recall value can be obtained by the following equation:

$$Recall\ i = \frac{iA}{iA + iB}$$

The same thing as precision, to calculate the overall value of recall is the total recall of each class divided by the number of classes (Iskandar and Suprapto, 2015). Then there is the f1 value which is the ratio of the average precision and recall that is weighted, with the following equation:

$$f1 = \frac{2\ x\ (recall\ x\ presicion)}{(recall\ x\ precision)}$$

3. Method

This research is a case study on Twitter data regarding the Covid-19 vaccine in Indonesia to produce a classification of public sentiment. The entire analysis process is carried out using the Python programming language. The following are the steps that will be carried out in this research.

3.1. Data Preparation

The data used is secondary data through the Application Programming Interface (API) provided by Twitter, then the data is processed manually by labeling (positive, negative, and neutral). The data collected and ready to be processed in the form of tweets amounted to 1675 data tweets which were tweets by Twitter users in Indonesia related to the Covid-19 vaccine with a data collection time span of January 10, 2021 - April 10, 2021.

3.2. Pre-Processing

Before the data is classified, it is necessary to do preprocessing to change the shape of the document into structured data according to its needs so that it can be further processed in the text mining process. The preprocessing stage of text in classification aims to improve the accuracy of data classification. Preprocessing in text mining is quite complicated because in Indonesian there are various rules for writing sentences and forming words with affixes. The sequence of stages in the preprocessing includes: data cleaning by removing duplicate tweets. Duplicate Tweets on Twitter are usually retweeted with the “RT” symbol, then the tweet will be deleted. In addition, tweets in foreign languages will be deleted, because the data used in this study is Indonesian Twitter data, removing URL and username links, doing case folding, which is changing all text to lowercase (non-capital) and removing punctuation marks, whitespace (space, tab, newline) is used as a separator between words to be cut, deletes words in tweets that are in the stopwords list (terms or words that are not related to the document even though the word often appears in

the document, but if it is deleted it does not changing the meaning of tweets), doing steeming to remove affixes and getting basic words, changing tweet data into the form of frequency of word occurrences. Data that has gone through preprocessing will then be labeled manually according to its class. This manual determination of sentiment is subjective, because it is determined according to the opinion of the researcher based on pre-determined sentiment classes, namely positive, negative, and neutral. This process is carried out in order to find out the grouping of sentiments in the previously obtained data.

3.3. K-Fold Cross Validation

The data sharing technique is done by K-fold cross validation using the value of k=10.

3.4. Data Classification

The classification technique used is SVM. After the classification results are obtained, an evaluation of the classification results will be carried out, until finally interpreting them and drawing conclusions.

3.5. Evaluation

The next step is to conduct an evaluation. The evaluation aims to determine the level of accuracy of the results of data classification using the SVM method. Evaluation can be measured by calculating the accuracy value of the classification results. The accuracy value of the classification can be obtained by using the confusion matrix.

4. Results and Discussion

Prior to starting the analysis, it is necessary to prepare the data for processing. The data was acquired using the Indonesian Twitter API using the Python tool tweepy and the term "vaccine" without include retweet data. The gathered data is saved as comma separated values (csv) files for further manual sorting and tagging, until 1675 tweets are collected and ready to be processed further. The following are the findings of data labeling:

Table. 2. Distribution of Data for each Class

Label	Amount of data
Negative	421
Netral	767
Positive	478

The data that has been formed into the three classes, is converted into two classes by combining tweets labeled neutral with data labeled positive, and making new data forms to be processed are 1245 data labeled 1 (positive) and data labeled 0 (negative). Data that has been processed previously still contains components that are not needed or will reduce the accuracy of the classification results. Therefore it is necessary to clean the data in the form of:

1) do case folding, which changes all text to lowercase (non-capital)
2) remove punctuation
3) remove ASCII (American Standard Code for Information Interchange) and Unicode which is a form of character code remove punctuation marks, whitespace (spaces, tabs, newlines)
4) delete words in tweets that are in the stopwords list (terms or words that are not related to the document even though the word often appears in the document, but if deleted does not change the meaning of the tweet)
5) and do stemming to remove affixes and get basic words. After cleaning the data, the test data and training data will be divided with a test data size of 0.1.

Before doing the classification, the text data needs to be converted into vectors and word weighted with TF-IDF. Classification is done using the SVM method with a linear kernel and using the help of a package from Python, namely sklearn. In the data classification process, it is tested using the 10 fold cross validation method. The results of the classification analysis using the SVM technique with the use of 1675 data are then evaluated. Classification evaluation is carried out using f1 values, accuracy, precision, and recall, the following table contains the values of the classification evaluation components:

Table. 3. Value of Classification Evaluation Results

f1	Accuracy	Precision	Recall
88.7%	84.4%	86.2%	97%

The results of the classification evaluation show a good value in the classification process. The classification evaluation can also be visualized in the form of a confusion matrix as follows:

Table. 3. Confusion Matrix

		Prediction	
		Class A	Class B
Actual Class	Class A	121	26
	Class B	7	7

Class A is a representation of sentiment with a positive class and class B for sentiment with a negative class. Next, we will try to enter new sentiment data and see if the data is classified correctly. The first sentence to be tested is "let's make the vaccine program a success" and the second sentence is "still afraid of the side effects of the vaccine". The output results show that the input data in the form of the first sentence is a weight value of 1 and the second sentence is 0 which indicates the first sentence is a sentence with positive sentiment and the second sentence is included in a sentence with negative sentiment.

5. Conclusion

Sentiment analysis for the Covid19 vaccination was conducted on Twitter data using the linear kernel SVM classification approach with TF-IDF weighting and data sharing via K-fold cross validation with k=10. Positive and negative classifications are made. Following preprocessing and classification, we got f1, accuracy, precision, and recall as reference values for evaluating the classification, with f1 = 88.7%, accuracy = 84.4%, precision = 86.2%, and recall = 97%. This value is acceptable, and hence SVM is suitable for identifying sentiment data about the Covid-19 vaccine in Indonesia.

References

[1] S. Lestari and S. Saepudin, “Analisis Sentimen Vaksin Sinovac Pada Twitter Menggunakan,” 2021.

[2] 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,” pp. 530–536, 2021.

[3] M. D. Mulyawan, M. D., & Slamet, I. (2021). Analisis Sentimen Terkait Vaksin Covid-19 Pada Data Twitter Menggunakan Support Vector Machine. 133–139.Mulyawan and I. Slamet, "Analisis Sentimen Terkait Vaksin Covid-19 Pada Data Twitter Menggunakan Support Vector Machine," pp. 133–139, 2021.

[4] R. Yanuarti, "Jurnal Sistem dan Teknologi Informasi Analisis Media Sosial Twitter Terhadap Topik Vaksinasi Covid-19," vol. 6, no. 2, pp. 121–130, 2021.

[5] S. K. S. Kom, "Implementasi Algoritma Latent Dirichlet Allocation Untuk Topic Modeling Terhadap Data Twitter Terkait Pandemi Covid-19," 2021.

[6] Khoirun Nisa Aulia Sukmani, "Analisis Postingan Di Twitter Mengenai Vaksinasi Covid-19: Perilaku Sosial Terhadap Vaksinasi Covid-19 Guna Pencegahan Penularan Covid-19," HUMAYA J. Hukum, Humaniora, Masyarakat, dan Budaya, vol. 1, no. 1, pp. 30–42, 2021, doi: 10.33830/humaya.v1i1.1802.2021.

[7] R. Sistem, M. Lestandy, A. Abdurrahim, and L. Syafa, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent," vol. 5, no. 10, pp. 802–808, 2021.

[8] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid - 19 Menggunakan Algoritma Naïve Bayes Classifier," vol. 2, no. 2, pp. 1–9, 2021.

[9] P. S. Informatika, F. Teknik, and U. M. Malang, "Analisis Sentimen Pengguna Twitter Terhadap Vaksin Covid-19 Menggunakan Metode Naïve," no. 201710370311009, 2021.

[10] M. A. N. Febriansyach, F. Rashif, G. I. P. Nirvana, and N. A. Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Tweet Akun Bot Twitter bertagar #covid-19," CogITo Smart J., vol. 7, no. 1, p. 170, 2021, doi: 10.31154/cogito.v7i1.299.170-181.

[11] 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.

[12] R. A. Widyanto, "Data Mining Predicts the Need for Immunization Vaccines Using the Naive Bayes Method," vol. 2, no. 3, pp. 93–101, 2021.

[13] L. Jen and Y. Lin, "A Brief Overview of the Accuracy of Classification Algorithms for Data Prediction in Machine Learning Applications," vol. 2, no. 3, pp. 84–92, 2021.

[14] H. J. Pambudi, A. Lukito, A. Nugroho, L. Handoko, and F. E. Dianastiti, "Buzzer Di Masa Pandemi Covid-19 : Studi Analisis Wacana Kritis Kicauan Buzzer Di Twitter Buzzers During The Covid-19 Pandemic : Study Of Critical Discourse Analysis Of Buzzer ' S Tweet On," vol. 23, no. 1, pp. 75–89, 2021, doi: 10.14203/jmb.v23i1.1265.

[15] R. Yasmin, "Covid-19 Menggunakan Metode Naive Bayes Classifier Pada Media Sosial Twitter Covid-19 Menggunakan Metode Naive Bayes," 2021.

[16] E. Nufa, "Analisis Klasifikasi Sentimen Tentang Pro Dan Kontra Masyarakat Indonesia Terhadap Vaksin Covid-19 Pada Media," no. May, p. 2, 2021.

[17] U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," vol. 5, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.

[18] F. Fitriana, E. Utami, and H. Al Fatta, "Analisis Sentimen Opini Terhadap Vaksin Covid-19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes," vol. 5, no. 1, pp. 19–25, 2021.

[19] D. Hernikawati, "Kecenderungan Tanggapan Masyarakat Terhadap Vaksin Sinovac Berdasarkan Lexicon Based Sentiment Analysis The Trend of Public Response to Sinovac Vaccine Based on Lexicon Based Sentiment Analysis," vol. 23, no. 1, pp. 21–31, 2021.

[20] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," Heal. Inf. Manag. J., vol. 8, no. 2, pp. 100–109, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175.

[21] A. Sasmito Aribowo, "Analisis Sentimen Publik pada Program Kesehatan Masyarakat menggunakan Twitter Opinion Mining," Semin. Nas. Inform. Medis, vol. 0, no. 0, pp. 17–23, 2018, [Online]. Available: https://journal.uii.ac.id/snimed/article/view/11877.