International Journal of Software Engineering and Informatics
Vol. 1, No.1 March 2023 Pp 33-39.
ISSN: xxxx xxxxx,

Analisis Sentimen di Twitter Menggunakan Pendekatan Machine Learning
Sentiment Analysis on Twitter Using Machine Learning Approach

Panji Bintoro[1*], Tahta Herdian Andika[2], Aviv Fitria Yulia[3], Panggah Widiandana[4]

[1]Software Engineering, Aisyah University, Indonesia
[2]Software Engineering, Aisyah University, Indonesia
[3]Software Engineering, Aisyah University, Indonesia
[4]Institute of Science Technology and Mulia Health, Yogyakarta, Indonesia

[1]panjibintoro09@aisyahuniversity.ac.id, [2]tahta.herdian.a@aisyahuniversity.ac.id,
[3]avivfitria@aisyahuniversity.ac.id, [4]panggah_widiandana@istekmulia.ac.id

Article Info

Kata Kunci:

COVID-19
Analisis Sentimen
Twitter
Machine Learning

ABSTRAK

Di media sosial Twitter, Penanganan COVID-19 di Indonesia pernah menjadi isu yang sangat dibicarakan. Penanganan yang dilakukan oleh pemerintah Indonesia menimbulkan pro dan kontra di kalangan masyarakat. Opini publik yang terdapat di Twitter bisa digunakan sebagai sumber pendukung keputusan dalam membuat kebijakan yang tepat dalam mengevaluasi kinerja pemerintah. Dalam menghadapi fenomena ini, metode analisis sentimen dapat digunakan untuk menganalisis opini publik di Twitter. Tujuan dari penelitian ini adalah untuk memahami opini publik tentang COVID-19 di Indonesia. Data tweet mengenai COVID-19 di Indonesia diproses untuk mengklasifikasikan menjadi sentimen positif dan negatif. Preprocessing dilakukan untuk menghilangkan data duplikat dan tidak relevan. Selain itu, text feature extraction dibandingkan untuk melakukan analisis sentimen pada data yang baru. Algoritma pembelajaran mesin diuji menggunakan beberapa metode dan dilakukan 10-fold validasi. Hasil terbaik diperoleh pada skenario kedua dengan classifier terbaik adalah Support Vector Machine, yang memperoleh akurasi sebesar 84.03%.

Keywords:

COVID-19
Sentiment Analysis
Twitter
Machine Learning

ABSTRACT

On social media Twitter, the handling of COVID-19 in Indonesia has been a highly discussed issue. The handling carried out by the Indonesian government raises pros and cons among the public. Public opinion contained on Twitter can be used as a source of decision support in making appropriate policies in evaluating government performance. In dealing with this phenomenon, the sentiment analysis method can be used to analyze public opinion on Twitter. The purpose of this research is to understand public opinion about COVID-19 in Indonesia. Tweet data regarding COVID-19 in Indonesia is processed to classify sentiment into positive and negative. Preprocessing is done to eliminate duplicate and irrelevant data. In addition, text feature extraction is compared to perform sentiment analysis on the new data. The machine learning algorithm is tested using several methods and is subjected to 10-fold validation. The best results were obtained in the second scenario with the best classifier being the Support Vector Machine, which obtained an accuracy of 84.03%.

Correspondence Author:

Panji Bintoro,
Software Engineering,
Aisyah University, Indonesia
Email: panjibintoro09@aisyahuniversity.ac.id

1 INTRODUCTION

The high usage of the internet, especially social media, in Indonesia has been proven based on the results of the Hootsuite Social Wear research in January 2019. The data mentioned that the number of social media users in Indonesia reached 150 million, or around 56% of the total population, increasing by 20% from the previous survey [1]. Twitter is one of the popular and continuously growing social media platforms in Indonesia. Twitter users generally share opinions about government policies and socio-political issues, as well as reviews of new products (which can be utilized for business intelligence) [2]. However, the content shared on Twitter can be used as an effective source of information to assist in decision-making.

The impact of the COVID-19 pandemic has been felt by all layers of society and sectors in Indonesia. Based on data from covid.go.id on May 14, 2020, there were 16,006 confirmed cases, 3,518 recovered patients, and 1,043 deaths [3]. The Indonesian government's response to handling COVID-19 has been considered slow by some parties, causing the policies taken to become a popular topic on Twitter and sparking pros and cons [4], especially since the impact of this pandemic is very wide-ranging and involves many sectors [5].

To ensure that government policies can achieve their targets, the information conveyed must be clear and accurate. Public opinions scattered on Twitter can provide important input in the policy-making process. However, gathering public opinions from Twitter is not an easy task because the content is often unstructured and written in informal or non-standard language. However, if data collection can be done properly, the information obtained can help in making appropriate policies. Therefore, certain methods or techniques are needed to speed up the classification of existing opinions.

Sentiment analysis in text mining aims to understand the sentiment, emotion, and attitude contained in opinion texts. The basic approach in sentiment analysis is to identify the direction of the opinion tendency in the text, whether it is positive, negative, or neutral [6]. Therefore, in this study, sentiment analysis is used to obtain an understanding of public views on the government's performance in handling COVID-19 [7].

Research on sentiment analysis continues to evolve with the emergence of new approaches. One popular and effective approach in analyzing text-based data is machine learning-based sentiment analysis [8]. Previously, Prastyo et.al. [9] conducted research to analyze public sentiment on the government's performance in handling COVID-19 by comparing two algorithms, namely Support Vector Machine and Multinomial Naive Bayes. The results showed that Support Vector Machine (SVM) had a higher level of accuracy compared to Multinomial Naive Bayes (MNB).

G.Singh et. al. [10] conducted a comparison between Multinomial Naive Bayes (MNB) and Bernoulli Naive Bayes (BNB) in predicting the positive or negative sentiment of specific news articles. Based on the test results, MNB performed better than BNB. Meanwhile, A. Lutfi et. al. [11] conducted sentiment analysis on Indonesian market sales reviews by comparing the performance of Support Vector Machine (SVM) and Naive Bayes Classifier (NBC). The research showed that SVM outperformed NBC with an accuracy of 93.65%.

In this study, we conducted a comparison of machine learning algorithms by analyzing several algorithms, including SVM, MNB, and Random Forest to determine the best performance in conducting Indonesian sentiment analysis on the public's response to the government's handling of COVID-19.

2 RESEARCH METHOD

In this section, the research results are presented and a comprehensive discussion is provided. The results can be presented in the form of images, graphs, tables, and others that facilitate readers [2]. The discussion can be conducted in several sub-sections.

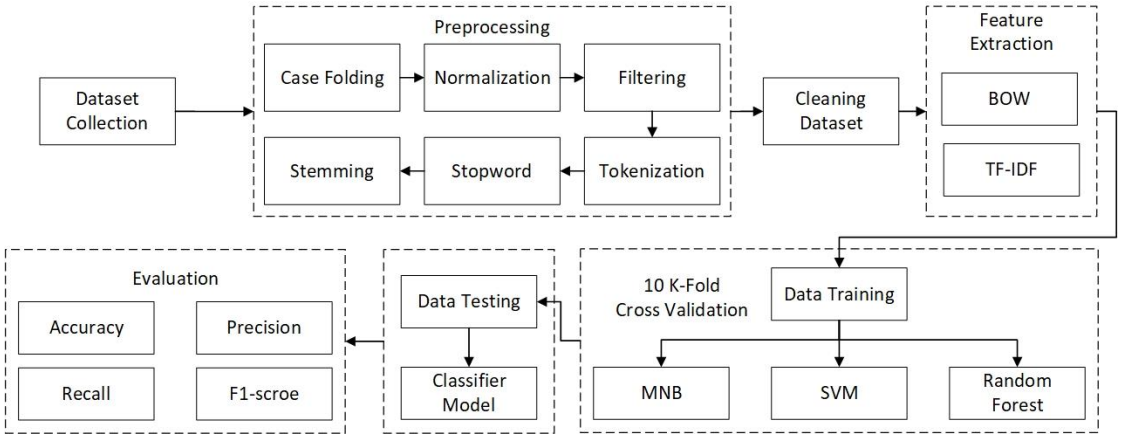

Figure 1. Proposed method.

2.1 *Dataset Collection*

The dataset used in this study is the same as the dataset used in Prastyo et.al.'s [9] research, consisting of 2269 labeled data with details of 874 positive, 1044 negative, and 351 neutral. However, in this study, only two classes are used, namely positive and negative.

2.2 *Preprocessing*

The text data from Twitter was initially in a random state, therefore in this final project, preprocessing steps were carried out to obtain tweets with relevant and structured data, making it easier for the algorithms to be used later. Preprocessing steps include case folding, normalization, filtering, tokenization, stopword removal, stemming, and cleaning the dataset.

2.3 *Feature Extraction*

This research uses two different scenarios in the feature extraction stage. The first scenario only uses BOW (Bag of Word), while the second scenario uses TF-IDF (Term Frequency Inverse Document Frequency). The BOW approach divides each word in the document as a potential important keyword of the document. Each document is treated as a collection of individual words, and TF represents the frequency of each word to describe the relevance of the word to the document. IDF aims to measure how common a frequency is in a document. The purpose of these feature extraction methods is to improve performance, especially on recall and precision values. The TF equation can be seen in Equation 1.

$$TF_{(d,t)} = \frac{f\ t,d}{\sum t,d} \tag{1}$$

$f\ t,d$ represents the frequency of each word (t) in the document (d) and $\sum t,d$ epresents the total number of words contained in the document (d). The IDF equation can be seen in equation 2.

$$IDF_{(t)} = \log \frac{|D|}{f\ t,d} \tag{2}$$

$|D|$ represents the number of documents in the collection and $f\ t,d$ represents the frequency of each word (t) in the document (d). Then, the TF-IDF value can be calculated using equation 3.

$$W_{(d,t)} = TF_{(d,t)}.IDF_{(t)} \tag{3}$$

In the TF-IDF equation, the TF result is multiplied by IDF, and the final result of this calculation is called feature extraction.

2.4 *Classification Model*

After the preprocessing and feature extraction stage, the next step is to create a model for classification. In this stage, the data is divided into two types, namely training data and testing data with a 80:20 split ratio, where 80% is for training data and 20% is for testing data.

a. *Data Training*

After performing preprocessing and feature extraction, the next step is to create a model for classification using several machine learning algorithms such as Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), and Random Forest. In this experiment, only two classes of data are used, namely positive and negative, with four different classification algorithms. The performance of each algorithm is evaluated to obtain the best testing results using the prepared training data. The classification process uses the training data that is tested using 10-fold cross-validation to evaluate the model's performance. This is done to reduce computation time and avoid overfitting. The entire amount of training data is divided into two parts, namely training and testing data, then the value of k=10 is used to create 10 folds with the same or approximately the same size, so that it has 10 subsets of data to evaluate the model or algorithm's performance. Each subset of data will use cross-validation with 9-fold for training and 1-fold for testing. An illustration of k-fold cross-validation can be seen in Figure 2.

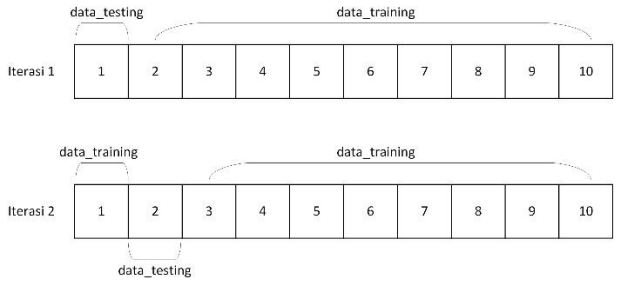

Figure 2. Illustration of k-fold cross validation.

b. *Data Testing*

Dalam proses kedua, dilakukan pengujian menggunakan data testing dengan perbandingan 20% dari total dataset yang telah disiapkan. Tujuan dari proses ini adalah untuk mengevaluasi performa dari model yang telah dibangun dan mengetahui apakah model tersebut dapat bekerja dengan baik pada data baru. Data testing biasanya memiliki jumlah yang lebih sedikit dibandingkan data training, hal ini dikarenakan model yang diuji pada kasus yang banyak dan bervariasi cenderung lebih cerdas sehingga prediksi kesalahan pada pengujian dapat dikurangi.

2.5 *Evaluation dan Validation*

In the second process, testing is performed using 20% of the total prepared dataset. The purpose of this process is to evaluate the performance of the built model and determine whether the model can work well on new data. Testing data usually has a smaller amount compared to training data because a model that is tested on many diverse cases tends to be more intelligent, thus reducing prediction errors in testing.

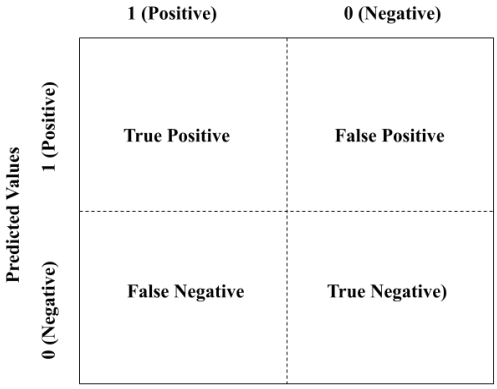

Figure 3. *Confusion matrix*

Based on the illustration in the above figure, there are four terms that represent the results of the classification process in the confusion matrix. These terms are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which are then used to calculate the values of accuracy, precision, recall, and f1-score. Accuracy indicates how accurate the system is in correctly classifying the data. Simply put, accuracy can be calculated by comparing the number of data that are classified correctly with the total amount of data. Precision measures the number of positive category data that are classified correctly compared to the total data that are classified as positive. Recall, on the other hand, indicates how much positive category data is successfully classified correctly by the system. F1-score describes the weighted average of precision and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (4)$$

$$Precision = \frac{TP}{FP + TP} * 100\% \quad (5)$$

$$Recall = \frac{TP}{FN + TP} * 100\% \quad (6)$$

$$F1 - score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (7)$$

where:

a. TP stands for True Positive, which is the number of positive data that are correctly classified by the system.
b. TN stands for True Negative, which is the number of negative data that are correctly classified by the system.
c. FP stands for False Positive, which is the number of positive data that are wrongly classified by the system.
d. FN stands for False Negative, which is the number of negative data that are wrongly classified by the system.

3 RESULTS AND ANALYSIS

In this research, the main focus is on Binary Classification task, where the data is divided into two classes, positive and negative. Several classifications are used to classify tweets and predict new unseen data. Three types of classifications used in this research are MultinomialNB, Support Vector Machine, and Random Forest. Experiments are conducted to determine which classification has the best performance.

The dataset used in this study, which was taken from Prastyo et al. [9], originally consisted of 3 classes: positive, negative, and neutral. However, in this study, only the binary classification task was focused on, so only 2 classes were used, namely positive and negative classes. Therefore, the neutral class was removed from the dataset and the dataset details can be seen in Table 1.

Table 1. The number of each class in the dataset.

No	Class	Total
1	Positif	874
2	Negatif	1044

Before training each classification algorithm, the tweet data was preprocessed through several stages including case folding, filtering, normalized alay words, tokenization, stopword removal, and stemming. In this study, two experiments were conducted:

a. Scenario 1

In the first experiment, preprocessing was performed on tweet data using Bag of Word feature extraction, which only takes into account the frequency of word occurrences or term frequency.

b. Scenario 2

In the second experiment (scenario 2), preprocessing was performed on tweet data using feature extraction in the form of Term Frequency Inverse Document Frequency (TF-IDF).

In each scenario, the dataset was divided into two parts, namely training data and testing data with a ratio of 80%:20%. The training data is used as a training sample for each classifier, while the testing data is a sample that has never been seen by the classifier. The testing data is tested to evaluate the performance of each classifier in performing classification tasks.

3.1. Scenario 1

In scenario 1, after the dataset was preprocessed and feature extraction was performed using the CountVectorizer function from the Sklearn library, training was carried out for each classifier algorithm on the training dataset. This training process was carried out by considering a k-fold cross-validation scheme with a value of k equal to 10, and the accuracy value was calculated. Details of the training process in scenario 1 for each classifier can be seen in Table 2.

Table 2. Scenario 1.

10-Fold	Accuracy		
	MNB	SVM	RF
1	0.77037	0.822222	0.82963
2	0.807407	0.777778	0.77037
3	0.843284	0.835821	0.828358
4	0.80597	0.820896	0.80597
5	0.843284	0.850746	0.850746
6	0.798507	0.791045	0.791045
7	0.858209	0.820896	0.798507
8	0.843284	0.791045	0.80597
9	0.828358	0.783582	0.783582
10	0.880597	0.835821	0.813433
Avg	0.827927	0.812985	0.807761
Min	0.77037	0.777778	0.77037
Max	0.880597	0.850746	0.850746

It can be seen that MultinomialNB has a higher average accuracy than the other classifiers in 10-fold validation, reaching 82.79%. After training, the next step is to perform hyperparameter tuning using Grid Search CV from Sklearn to find the best parameter from each parameter. After the best parameter is found, predictions are made on the testing data. The results of the experiment using the testing data are shown in Table 3.

Table 3. Result of scenario 1.

No	Classifier	Accuracy
1	SVM	82.64%
2	Random Forest	81.25%
3	MultinomialNB	80.90%

From the results of scenario 1, it can be seen that SVM is the classifier algorithm that has the best performance in predicting testing data, with an accuracy of 82.64%.

3.2. Scenario 2

In scenario 2, the training dataset has been preprocessed and feature-extracted using the TF-IDF Vectorizer function from the Sklearn library. Then, each classifier algorithm was trained by considering a 10-fold cross-validation scheme, and its accuracy score was computed. The details of the training process in scenario 2 for each classifier are presented in Table 4.

Table 4. Scenario 2.

10-Fold	Accuracy		
	MNB	SVM	RF
1	0.770337	0.822222	0.807407
2	0.777778	0.785185	0.762963
3	0.798507	0.850746	0.798507
4	0.843284	0.828358	0.828358
5	0.828358	0.858209	0.865672
6	0.783582	0.80597	0.776119
7	0.813433	0.820896	0.798507
8	0.828358	0.813433	0.791045
9	0.850746	0.835821	0.80597
10	0.798507	0.843284	0.813433
Avg	0.809292	0.826412	0.804798
Min	0.77037	0.785185	0.762963
Max	0.850746	0.858209	0.865672

Based on the experiment results in scenario 2, it is seen that the SVM classifier shows the best average accuracy in 10-fold validation compared to other classifiers, reaching 82.57%. After the training process is completed, hyperparameter tuning is performed using the Grid Search CV function from Sklearn to find the best parameters for each classifier. Once the best parameters are found, predictions are made on the testing data. The experiment results using the testing data are shown in Table 5.

Tabel 5. Hasil skenario 2.

No	Classifier	Accuracy
1	SVM	84.02%
2	Random Forest	80.9%
3	MultinomialNB	80.55%

From the results of the experiment in scenario 2, it can be seen that the SVM classifier has a very good accuracy, reaching 84.02%, compared to other classifiers in the same scenario. This indicates an improvement in performance compared to scenario 1, where the SVM classifier only reached 82.64%. The details of the performance evaluation results of the SVM classifier with TF-IDF BOW feature extraction are shown in Table 6.

Tabel 6. Hasil metric evaluasi SVM.

No	Classifier	Accuracy
1	Accuracy	84.02%
2	Precision	79.15%
3	Recall	84.36%
4	F1-measure	81.67%

4 CONCLUSION

In this study, we developed a sentiment classification model that analyzes tweets about COVID-19 in Indonesia and classifies them into positive and negative classes. We also compared the use of feature extraction with or without TF-IDF for BOW representation, and tuned hyperparameters for each classifier. The results showed that SVM with RBF and a C value of 10 and gamma of 1 exhibited the best performance, with an accuracy of 84.02%, precision of 79.15%, recall of 84.36%, and F1-measure of 81.67%. This represents a 2% improvement over the model developed by Prastyo[9], which did not use slang normalization and stemming as we did. We hope to use more data and other features in the future to find an even better model.

REFERENCE

[1] Katadata, "Berapa pengguna media sosial Indonesia?" katadata.id, 2019 [Online]. Tersedia: https://databoks.katadata.co.id/datapublish/2019/02/08/berapa-pengguna-media-sosial-indonesia (accessed May 05, 2020).

[2] AntaraNews, "Pengguna Twitter indonesia tumbuh pesat di 2018," AntaraNews.com, 2019 [Online]. Tersedia: https://www.antaranews.com/berita/839825/pengguna-twitter-indonesia-tumbuh-pesat-pada-2018 (accessed May 05, 2020).

[3] Gugus Percepatan Penanganan COVID-19, "Data Covid-19 di Indonesia," covid19.go.id, 2020 [Online]. Tersedia: https://covid19.go.id/ (accessed May 14, 2020).

[4] Suara, "Pemerintah Indonesia dinilai lambat mengantisipasi covid-19 sejak dini," suara.com, 2020 [Online]. Tersedia: https://www.suara.com/news/2020/04/10/025500/pemerintah-indonesia-dinilai-lambat-mengantisipasi-covid-19-sejak-dini?page=all (accessed May 14, 2020).

[5] DetikNews, "Pemerintah dinilai lambat tangani corona, Jokowi: Kita tak ingin grusa-grusu," news.detik.com, 2020 [Online]. Tersedia: https://news.detik.com/berita/d-4987368/pemerintah-dinilai-lambat-tangani-corona-jokowi-kita-tak-ingin-grasa-grusu (accessed May 14, 2020).

[6] I. Zulfa and E. Winarko, "Sentimen analisis tweet berbahasa Indonesia dengan deep belief network," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 11, no. 2, pp. 187, 2017, doi: 10.22146/ijccs.24716.

[7] J. Michie, "The covid-19 crisis – and the future of the economy and economics," Int. Rev. Appl. Econ., vol. 34, no. 3, pp. 301–303, 2020, doi: 10.1080/02692171.2020.175604

[8] R. Shahid, S. T. Javed, and K. Zafar, "Feature selection based classification of sentiment analysis using biogeography optimization algorithm," in 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), 2017, pp. 1–5, doi: 10.1109/ICIEECT.2017.7916549

[9] Prastyo, Sumi, Dian, & Permanasari , "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel", Journal of Information Systems Engineering and Business Intelligence, 2020, 6 (2), 112-122

[10] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," in 2019 International Conference on Automation, Computational and Technology Management, ICACTM, 2019, pp. 593–596, doi: 10.1109/ICACTM.2019.8776800

[11] F. Ratnawati and E. Winarko, "Sentiment Analysis of Movie Opinion in Twitter Using Dynamic Convolutional Neural Network Algorithm," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 12, no. 1, pp. 1, 2018, doi: 10.22146/ijccs.19237.

[12] N. D. Putranti and E. Winarko, "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine*," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 8, no. 1, pp. 91, 2014.ISSN: 1978-1520

[13] L. B. Ilmawan and E. Winarko, "Aplikasi Mobile untuk Analisis Sentimen pada Google Play" IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 9, no. 1, pp. 53, 2015.ISSN: 1978-1520