PUBLIC OPINION ON COVID 19 REGULATION IN INDONESIA: PREDICTION USING NAÏVE BAYES AND TERM WEIGHTING TF-IDF

Supangat, Mohd Zainuri Bin Saringat, Mochamad Yovi Fatchur Rochman
Universitas 17 Agustus 1945 Surabaya

Abstract

There are many public responses about the implementation of government policies related to Covid-19. Some have positive and negative opinion, especially on the official social media portal of the government twitter where people are free to argue. This study aims to find out the opinion of sentiment analysis on Twitter in the implementation of government policies related to Covid-19 so as to classify public opinion. Several stages in the process of analyzing public sentiment are taken from the tweet data. The first step is to do the data mining process to get the tweets that will be analyzed later. Furthermore, the process of cleaning tweet data and equalizing tweet data into lowercase. After that perform the basic word search process of the tweet and calculate the frequency of its appearance. Then calculate using the naïve bayes method and determine the sentiment classification of the tweet. The results showed that public sentiment related to covid-19 prevention in Indonesia is neutral. The performance of the application shows an Accuracy value of 76.7%.

Keywords: Naïve Bayes, TF-IDF, COVID-19, Prediction

Abstrak

Banyak tanggapan masyarakat mengenai implementasi kebijakan pemerintah terkait Covid-19. Ada yang berpendapat positif dan negatif, terutama di portal media sosial resmi twitter pemerintah di mana orang bebas berpendapat. Penelitian ini bertujuan untuk mengetahui opini analisis sentimen di Twitter dalam implementasi kebijakan pemerintah terkait Covid-19 sehingga dapat mengklasifikasi opini publik. Beberapa tahapan dalam proses analisis sentimen publik diambil dari data tweet. Langkah pertama adalah melakukan proses data mining untuk mendapatkan tweet yang nantinya akan dianalisa. Selanjutnya dilakukan proses pembersihan data tweet dan menyamakan data tweet menjadi huruf kecil. Setelah itu melakukan proses pencarian kata dasar dari tweet tersebut dan menghitung frekuensi kemunculannya. Kemudian hitung menggunakan metode nave bayes dan tentukan klasifikasi sentimen dari tweet tersebut. Hasil penelitian menunjukkan bahwa sentimen publik terkait pencegahan covid-19 di Indonesia bersifat netral. Kinerja aplikasi menunjukkan nilai Accuracy sebesar 76,7%.

Kata kunci: Naïve Bayes, TF-IDF, COVID-19, Prediksi

Introduction

A sickness outbreak rocked the world at the start of the year 2020. This epidemic spread rapidly, infecting nearly every country on the planet. This is a coronavirus infection, also known as Coronavirus illness (COVID-19). WHO has classified the world in a global emergency about this virus since January 2020, according to the World Health Organization (Organization, 2020). The COVID '19 outbreak was widespread over the world in 2019. Within ten months, 38,085,762 infected people were found, with 28,628,813 (96%) recovering and 1,086,055 dying (4%). Even the transfer pace was lightning quick, affecting the entire world in a matter of seconds. The mortality rate was quite low. In response to the COVID-19 epidemic, practically all countries have enacted preventive policies such as social distance and remaining at home. (Supangat and Bin Saringat, 2020).

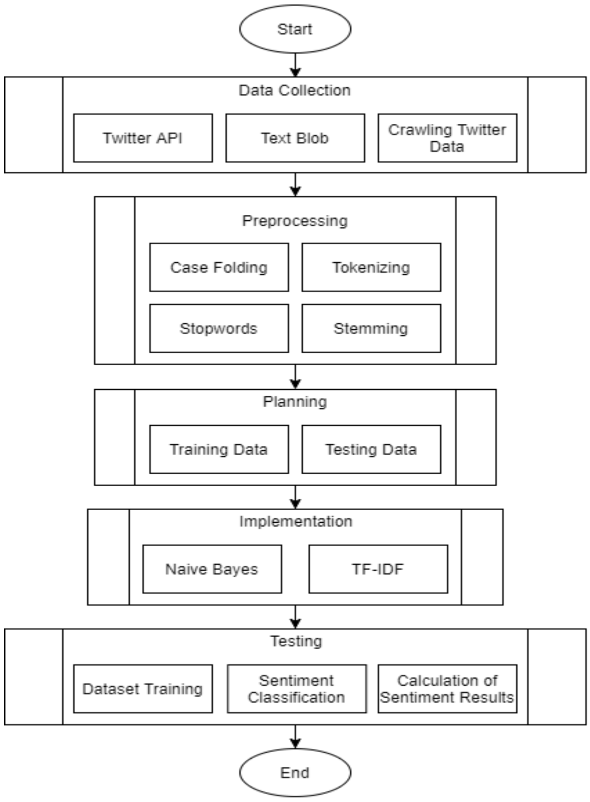
The Indonesian government has declared a catastrophe emergency in response to the virus epidemic, which will last from the end of February 2020. It continues and is being followed by the spread of countermeasures in different regions of Indonesia. The virus has spread to several areas until it occurred in May 2020. Other countermeasures implemented by the Indonesian government include the use of Physical Distancing, a concept that states that in order to reduce and even break the chain of Covid-19 infection, one must maintain a safe distance of at least 2 meters from other humans, avoid direct contact with other people, and avoid mass gatherings. (Ardan, Rahman, & Geroda, 2020).

Furthermore, the employment of multinomial discriminative methods of Naive Bayes and TF-IDF enhanced accuracy by 0.3%, according to the data. (Alsalman, 2020). Using the Nave Bayes technique, we observed a high categorization accuracy of 91% for short Tweets. We also discovered that with shorter Tweets, the logistic regression categorization algorithm provides 74% decent accuracy. (Samuel, Ali, et al., 2020). By displaying machine learning results using the Naive Bayes technique, we classify Twitter's sentiment data. Even though it takes longer than the listing technique, this algorithm can generate reasonably accurate estimations. (Alshaikh and Zohdy, 2020). The Naïve Bayes and decision tree approaches were compared in this study. With 73.59% accuracy, Nave Bayes outperformed Decision Tree. (Sari and Ruldeviyani, 2020). The Naive Bayes, Support Vector Machine and maximum entropy approaches were compared in this study. Max Entrophy had an accuracy value of 82.6% and a precision value of 84.0536512%, while Naive Bayes had an accuracy value of 86% and a precision value of 88.695952%, SVM had an accuracy value of 74.6% and a precision value of 75.88235235%, and SVM had an accuracy value of 74.6% and a precision value of 75.88235235%. According to the study, machine learning methods such as Naive Bayes have the best accuracy. They can be regarded as basic learning methods, while the Maximum Entropy method is helpful in some circumstances. (Mandloi and Patel, 2020). Based on statistical data, the goal of this work was to identify a machine learning strategy that was relatively better than SVM and Naive Bayes classifiers. The system achieves 82.853% precision, 82.884% recall, and 82.662% f1 score for SVM classifiers. Furthermore, Naive Bayes identified data with an accuracy of 83.990%, a recall of 83.997%, and an F1 score of 83.993%. As a result, when compared to traditional approaches, our model achieves SVM and Naive Bayes accuracy of 84% and 82.875%, respectively. (Dey et al., 2020). Naïve Bayes has been used widely for document classification since the 1950 (Pedregosa et al., 2011). However, NBCs are based on over-simplified assumptions of conditional probability and shape of data distribution (Liu et al., 2013; Ye et al., 2016; Kowsari et al., 2019; Alamoodi et al., 2020). Twitter data has also been extensively used for crisis situations analysis and tracking, including the analysis of pandemics (Fung et al., 2014; Kim et al., 2016; Samuel, Rahman, et al., 2020). The application is designed using Python and PHP languages. The accuracy value obtained in this study was 91.67%. (Sri Mulyani, Rohpandi and Rahman, 2019).

TF-IDF is one of the selection features that does not only contain the number of occurrences of words as in BOW. But TF-IDF also looks at important and less important words from the document. In a previous study (Nguyen et al., 2017) compared the feature selection using BOW with TF-IDF. The result is that TF-IDF is better than BOW, so in this study, the TF-IDF concept is used in the selection of features.

With the implementation of government policies related to Covid-19, many public responses, some have positive opinions, and some have negative opinions, especially on Twitter, where people are free to express their opinions. Based on this discussion, the author tries to conduct research on opinion sentiment analysis on Twitter in implementing government policies related to Covid-19 using Naive Bayes because in previous studies the Naive Bayes algorithm showed better results compared to other algorithms and feature weighting using TF-IDF. to support the naive bayes algorithm in increasing accuracy so that it can classify opinions from the public on Twitter social media which is built using python language.

Method Research [size 11, spacing 1,15]



Picture 1.
Research Framework

The stages of the framework of thought have the following explanation:

1. Data Collection
   Data collection is carried out at this stage to support how the system will be built in it. Before collecting twitter data, it is necessary to first prepare the twitter API. Twitter data collection is done by twitter data scraping method. To perform textual data processing at the next stage, a textBlob library is needed.
2. Preprocessing
   Case folding, tokenizing, stopword, and stemming are some of the preprocessing phases that are performed on tweet data before processing. The use of capital letters is not uniform across all text sources. As a result, case folding is required to convert the document's complete text into a standard format (uppercase or lowercase). For example, users who type "Berita," "BERITA," or "Berita" to acquire information about "BERITA" get the same retrieval result, namely "Berita." Case folding is the process of converting all letters in a document to lowercase. The letters 'a' to 'Z' are the only acceptable ones. Other than letters, all characters will be eliminated. The tokenizing stage is then utilized to break down the sentences in the string into single-word chunks.
   At this step, less powerful or frequently appearing words (stopwords), such as connecting words and adverbs that are not unique words, such as "an," "by," "on," and so on, are discarded. The stemming stage removes affixes, prefixes, and suffixes to change the

words back to their original form. The stop words in this investigation were generated using a modified Sastrawi library (Surya, Seetha and Subbulakshmi, 2019). Word weighting is used in news classification to determine a category. TF-IDF (Term Frequency–Inverse Document Frequency) is one of the weighting methods. Its weight value expresses the importance of a word (term) in representing the title. The weight will be more significant in the TF-IDF weighting if the frequency of occurrence of the term is higher. However, it will be lower if the word appears more frequently in other news.

3. Planning
   At this stage, a design for the distribution of training data and test data will be made based on the dataset that has been obtained. Because in this study the dataset obtained was 1000 tweets, this study will try to compare three comparisons of datasets, namely 70%-30%, 80%-20%, 90%-10% based on references in previous studies.
4. Implementation
   At this stage, based on the data that has been collected, it is made into a web-based application using the Naïve Bayes algorithm with the TF-IDF weighting feature using the python language.
5. Testing
   In the last stage, the training dataset was tested by looking at the level of accuracy generated by the training from each experiment. Then perform sentiment analysis based on available data and calculate the level of precision, recall, and accuracy using a confusion matrix.

Results and Discussion

The object of this research is community Twitter data related to the community's response to the Covid-19 response in Indonesia. There are three classification classes, namely negative, positive and neutral sentiments. The dataset is taken through a web scraping process to get data from Twitter social media. The following is a labeled dataset used in this study which can be seen in table 1.

TABLE 1. OBJECT RESEARCH

Class	Total Data	Resources
Positive	314	twitter
Negative	321	twitter
Neutral	365	twitter

In this study, the analysis was carried out by dividing the training data and testing data into three categories to test a good level of accuracy for sentiment classification with 1000 datasets, namely 70%-30%, 80%-20%, 90%-10%.

A total of 1000 data points are split into three categories: positive, negative, and neutral. Using the confusion matrix method, the data that has been normalized before being entered into the classification engine is separated into three experiments, namely

700 training data and 300 test data using the confusion matrix method.

800 training data and 200 test data were collected using the confusion matrix method

900 training data and 100 test data using the confusion matrix method.

TABLE 2. TRAINING DATA & TESTING DATA

Data Comparison		**Amount of Data**		**Naïve Bayes Accuracy**
Training Data	***Testing Data***	***Training Data***	***Testing Data***	***Sentiment Analysis***
70%	30%	700	300	76.70%
80%	20%	800	200	72.93%
90%	10%	900	100	74.36%

From table 2, the results of the experiment of several dataset distributions show that the comparison of 70%-30% datasets shows the greatest accuracy value compared to the distribution of 80%-20% and 90%-10% datasets. Therefore, in this study, what will be used for calculating the confusion matrix, precision, recall and f-score is a dataset with a division of 70% as training data and 30% as test data.

Table 3. describes the results of data acquisition and then, through preprocessing 1000 existing data, divided into three sentiments, namely category one is positive, category 2 is neutral, and category 3 is harmful. The data that has been normalized before being entered into the classification engine is divided into two, namely training data and test data using a confusion matrix.

TABLE 3. CONFUSION MATRIX SENTIMENT ANALYSIS

Actual class	**Predicted class**		
	Negative	***Neutral***	***Positive***
Negative	229	63	29
Neutral	24	317	24
Positive	34	59	221

Based on calculations using a confusion matrix, this study resulted in Sentiment Analysis using DMNB and TF-IDF on Twitter regarding the Covid-19 response into 3 categories, namely positive, neutral, and negative with positive sentiment as much as 28.7%, neutral as much as 43.9%, and negative as much as 27.4%.

Then it will use the formula and look for precision, recall, accuracy and f-1 measure values to determine the classification results with the following formula (Alsalman, 2020):

$$Precission = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+TN} X100\% \quad (3)$$

$$F1 - Measure = 2x\frac{precission+recall}{precission+recall} \quad (4)$$

Table 4. demonstrates the results of other evaluation metrics for negative and positive tweets.

TABLE 4. RESULT OF EVALUATION MATRIX

Category Name	Precision	Recall	F-Score	Accuracy
Negative	0.79	0.71	0.75	
Neutral	0.72	0.87	0.79	
Positive	0.81	0.70	0.75	76.7%
Weighted Avg.	0.77	0.76	0.76	

From table 3. and 4., we can notice that the Naïve Bayes classifier achieve 0.71 of recall metric for the negative tweets, 0.87 of recall metric for the neutral tweets and 0.70 of recall metric for the positive tweets. In addition, the experiment attains 0.77 of weighted average precision, 0,76 of weighted average recall and 0,76 of weighted average f-score. This study shows the accuracy for sentiment analysis is 76%.

Conclusion

This paper combines the multinomial nave Bayes (DMNB) discriminatory method with case folding, stop words, stemming, and term frequency-inverse document frequency (TF-IDF) approaches. The three primary phases are tweet data preparation and normalization, sentiment categorization and precision computation, and f-score recall and accuracy. The experiment was carried out on a local website written in the Python programming language. The dataset consists of 1000 Indonesian-language tweets that have been classified into three categories (negative, neutral, and positive). According to the testing data, the proposed approach has an average precision class of 77%, recall of 76%, and f-score of 76%. Moreover, it is 76.7% accurate. Sentiment Analysis utilizing DMNB and TF-IDF on Twitter divided the Covid-19 response into three categories: favorable, neutral, and negative, with positive sentiment accounting for 28.7%, neutral for 43.9%, and negative for 27.4%.

References

1. Organization, W. H. (2020). Situation Report-12. *Novel Coronavirus(2019-nCoV)*, 7. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/330777/nCoVsitrep01Feb2020-eng.pdf?sequence=1&isAllowed=y
2. Supangat and Bin Saringat, M. (2020) 'Development of e-learning system using Felder and Silverman's index of learning styles model,' *International Journal of Advanced Trends in Computer Science and Engineering*, 9(5), pp. 8554–8561. DOI: 10.30534/ijatcse/2020/236952020.
3. Ardan, M., Rahman, F. F. and Geroda, G. B. (2020) 'The influence of physical distance to student anxiety on COVID-19, Indonesia', *Journal of Critical Reviews*, 7(17), pp. 1126–1132. DOI: 10.31838/jcr.07.17.141.
4. Alsalman, H. (2020) 'An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media,' *ICCAIS 2020 - 3rd International Conference on Computer Applications and Information Security*, pp. 23–26. DOI: 10.1109/ICCAIS48893.2020.9096850.
5. Samuel, J. *et al.* (2020) 'COVID-19 public sentiment insights and machine learning for tweets classification', *Information (Switzerland)*, 11(6), pp. 1–22. doi: 10.3390/info11060314.

6. Alshaikh, M. and Zohdy, M. (2020) 'Sentiment Analysis for Smartphone Operating System: Privacy and Security on Twitter Data', *IEEE International Conference on Electro Information Technology*, 2020-July, pp. 366–369. doi: 10.1109/EIT48999.2020.9208303.
7. Sari, I. C. and Ruldeviyani, Y. (2020) 'Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers', pp. 23–28. doi: 10.1109/iwbis50925.2020.9255531.
8. Mandloi, L. and Patel, R. (2020) 'Twitter sentiments analysis using machine learninig methods', *2020 International Conference for Emerging Technology, INCET 2020*, pp. 1–5. doi: 10.1109/INCET49848.2020.9154183.
9. Dey, S. *et al.* (2020) 'A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews', *2020 International Conference on Contemporary Computing and Applications, IC3A 2020*, pp. 217–220. doi: 10.1109/IC3A48958.2020.233300.
10. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830. 35.
11. Liu, B.; Blasch, E.; Chen, Y.; Shen, D.; Chen, G. Scalable sentiment classification for big data analysis using naive bayes classifier. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013 ; pp. 99–104. 36.
12. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. Information 2019, 10, 150. [CrossRef]
13. Ye, X.; Li, S.; Yang, X.; Qin, C. Use of social media for the detection and analysis of infectious diseases in China. ISPRS Int. J. Geo-Inf. 2016, 5, 156. [CrossRef] 26.
14. Fung, I.C.H.; Yin, J.; Pressley, K.D.; Duke, C.H.; Mo, C.; Liang, H.; Fu, K.W.; Tse, Z.T.H.; Hou, S.I. Pedagogical Demonstration of Twitter Data Analysis: A Case Study of World AIDS Day, 2014. Data 2019, 4, 84. [CrossRef] 27.
15. Kim, E.H.J.; Jeong, Y.K.; Kim, Y.; Kang, K.Y.; Song, M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. J. Inf. Sci. 2016, 42, 763–781. [CrossRef] 28. Samuel, J.; Ali, N.; Rahman, M.;
16. Samuel, Y.; Pelaez, A. Feeling Like it is Time to Reopen Now? COVID-19 New Normal Scenarios Based on Reopening Sentiment Analytics. arXiv 2020, arXiv:2005.10961
17. Sri Mulyani, E. D., Rohpandi, D. and Rahman, F. A. (2019) 'Analysis of Twitter Sentiment Using the Classification of Naive Bayes Method about Television in Indonesia', *2019 1st International Conference on Cybernetics and Intelligent System, ICORIS 2019*, 1(August), pp. 89–93. doi: 10.1109/ICORIS.2019.8874896.
18. Alamoodi, A. H. *et al.* (2020) 'Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review', *Expert Systems with Applications*, p. 114155. doi: 10.1016/j.eswa.2020.114155.
19. V. L. Nguyen, D. Kim, V. P. Ho, and Y. Lim, "A New Recognition Method for Visualizing Music Emotion," in International Journal of Electrical and Computer Engineering (IJECE), 2017, vol. 7, no. 3, pp. 1246–1254.
20. P. P. M. Surya, L. V. Seetha, and B. Subbulakshmi, "Analysis of user emotions and opinion using Multinomial Naive Bayes Classifier," in Proceedings of the 3rd International

Conference on Electronics and Communication and Aerospace Technology, ICECA 2019, 2019, pp. 410–415, doi: 10.1109/ICECA.2019.8822096

Acknowledgment

I am deeply indebted to my mentor, Mr. Supangat, Mr. Mohd Zainuri bin Saringat, for warm support, inspiration and thoughtful guidance.