tn2

by Rina Mardiati

Sentiment Analysis of Covid-19 on Indonesian Twiter by Implementing the Naïve Bayes Method

Teguh Nurhadi Suharsono
Faculty of Engineering
Universitas Sangga Buana
Bandung, Indonesia
teguh.nurhadi@usbypkp.ac.id

Ahmad Fauzan
Faculty of Engineering
Universitas Sangga Buana
Bandung, Indonesia
Obhaz10@gmail.com

Rina Mardiati
Department of Electrical Engineering
UIN Sunan Gubung Djati Bandung
Bandung, Indonesia
r_mardiati@uinsgd.ac.id

***Abstract*—Twitter is one of the social media used in Indonesia to express opinions/opinions. One of them is the opinion about Covid-19 which is taking the world by storm. The government's provisions regarding Covid-19 itself reap many pros and cons on social media, one of which is Twitter. In this study, "Covid-19" will be used as a keyword to conduct sentiment analysis. Sentiment analysis is the process of understanding, extracting and processing textual data automatically to obtain information contained in an opinion sentence. The Naïve Bayes Classifier method is used to classify and calculate the total accuracy of the class that has been obtained. Based on the results from the kaggle dataset, there are a total of 2269 tweet documents with the keyword "Covid-19" on March 23 - May 14, 2020 which can be trusted because the data has been labeled by experts. The Naïve Bayes Classifier method has 2269 data sets, then divides it into 1815 training data, and 453 data testing data and produces an accuracy of 0.674.**

Keywords— sentiment analysis, social media, twitter, Covid-19, Naïve Bayes

I. Introduction

Social media is a collection of software that allows individuals and communities to gather, share, communicate, and in certain cases collaborate or play with each other [1]. One of the social media that is often used is Twitter, a new type of microblogging media that makes it easy to get news quickly and briefly [2]. Twitter users can express their opinion on a product or comment on a problem through tweets.

How to find out how people respond to COVID-19 on the twitter application can use text mining by applying sentiment analysis [3]. One of the purposes of using text mining is sentiment analysis. Sentiment analysis or opinion mining is a method used to analyze a person's opinions, judgments, attitudes and emotions towards a product, organization, individual, and event or topic to see a person's tendency to judge an object with positive or negative sentiments [4].

Problems arising from recognized and observed decisions may need to be resolved immediately, but the variability is so complex that the data cannot be obtained numerically [5]. There are many classification methods in completing sentiment analysis, including: K-Nearest Neighbors, Decision Tree, Naïve Bayes, Support Vector Machine and others [6] [7] [8]. In this study, the researcher used a dataset from Kaggle which had been labeled by a third party who was licensed in their field [9]. To test the data then use the Naive Bayes Classifier method as a classification process. The results of this study will be in the form of testing opinion classifications about "COVID-19 which are positive, neutral, and negative. Based on this, the research that will be carried out is sentiment analysis on public opinion in the form of sentiments obtained from social media Twitter using the Naive Bayes Classifier method to find out the pros and cons of the community regarding the analysis of COVID-19 sentiment on Indonesian Twitter using the Naive Bayes Classifier (NBC) method.

II. Research Method

The flow chart of this research methodology explains each flow or process in research which is conducted. In this study there is a scope so that this discussion is structured well. The stages of the research framework to be carried out can be seen in Figure 1 below:

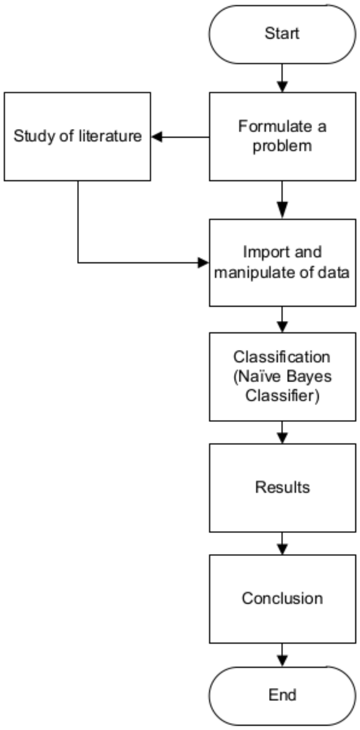
Start
Formulate a problem
Study of literature
Import and manipulate of data
Classification (Naïve Bayes Classifier)
Results
Conclusion
End

Fig. 1. Research model flow

The process in this study uses one method, namely the nave Bayes classifier. Sentiment analysis is to determine the categories of positive, negative, and neutral responses. The research flow in this sentiment analysis is a stage or general description that shows the stages of work being carried out. First, a literature study is carried out, this is done to get references on the theoretical basis related to the issues raised which are sourced from literature books. Based on the literature review from previous research, the authors found a solution for sentiment analysis using the Naïve Bayes method.

In this research process the data collection that will be used is taken from the responses to questionnaires distributed in the period March - May 2020. Furthermore, the data obtained is carried out by a cleaning process which aims to eliminate components that are not related to information and also correct the words that are used. not in accordance with the rules, besides that, a cleaning process is also carried out, namely case folding which aims to change capital letters to lowercase letters. When the cleaning and case folding processes are completed, the next step is to determine the sentiment score with the lexicon dictionary and the next is the nave Bayes classification using the confusion matrix test.

After the confusion matrix is formed, the classification determination is calculated using the proposed equation for accuracy, sensitivity, and specificity values with the following formula [10].

$$Accuracy = \frac{a+d}{a+b+c+d} \text{x100\%}$$

$$Specivity = \frac{d}{b+d}$$

$$Sensitivity = \frac{a}{a+c}$$

Information :
a = *true positive*
b = *false positive*
c = *false negative*
d = *true negative*

III. Results And Discussion

A. Data Source

This study uses secondary data which is typical of quantitative research, the dataset used is already available at Kaggle.com. The dataset used has been collected by a licensed 3rd party. This dataset was obtained from a collection of tweets with the keyword "Covid-19" in the time range from March 23 to May 14, 2020. The author chose this method to speed up the research, because it did not take long to access the dataset, and also did not need to go down and distribute questionnaires to the field. in the midst of this pandemic.

B. Data Preprocessing

The tweet data regarding the COVID-19 vaccination that has been collected and labeled is then preprocessed for the text including case folding, data cleansing, stemming, stopword and tokenizing. The following is a tweet data structure regarding the COVID-19 vaccination before preprocessing the text data

Table I. Data Preprocess

Tweet	Sentimen
Yuppp Because the government's debt is also the debt of the Indonesian nation. Means it's also my debt as a COVID 19 captain. If other people are up to them..but I am this...whatever the state of the government, THEY ARE STILL THE ARMY OF ALLAH SWT, which means that their good friends are negative.	Negative
Let's work together to help the government break the chain of covid-19 and break the chain of hatred. So that Indonesia is back to normal	Positive

[1] Case Folding

Case Folding is the process of equating cases in a document so that all text documents are consistent in the use of lowercase letters. Table is the process converting capital letters to lowercase, here is the process:

Table II. Case Folding

Text before Case Folding	Text after Case Folding
Yuppp Because the government's debt is also the debt of the Indonesian nation, it also means that it's my debt as a COVID 19 captain. If other people are up to them.. but it's me..whatever the state of the government, THEY ARE STILL THE ARMY OF ALLAH SWT, which means that they are good friends.	yuppp, because the government's debt is also the debt of the Indonesian nation, it also means that I owe it as the captain of covid 19. If other people are up to them.. but I'm right..whatever the state of the government is, they are still the army of Allah swt, which means they are good friends.
Let's work together to help the government break the chain of covid-19 and break the chain of hatred. So that Indonesia is back to normal.	let's work together to help the government break the chain of covid-19 and break the chain of hatred. So that Indonesia is back to normal.

[2] Data Cleansing

After doing case folding, then data cleansing is done. Data Cleansing functions to clean data or non-standard characters that can interfere with data processing. In data cleansing there are several stages of the process, as follows:

a. The process of changing the page URL to [URL] is as follows:

Table III. Data Cleansing

The text before the URL change	Text after URL change
imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) http:///content/detail/33735/penanganan-sebaran-konten-hoaks-covid-19-rabu-06042021/0/infografis	imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) [URL]

b. The process of deleting Hashtags is as follows:

Table IV. Removing Hashtags

Text before hashtag removal	Text after hashtag removal
"Let the government and the Indonesian population still provide insight and information to the general public regarding Covid 19, so that there will be no more discrimination against patients, families & corpses of COVID-19" @divisihumaspolri #kapoldasulsel #poldasulsel #opsyustisipoldasulsel	"Let the government and the Indonesian population still provide insight and information to the general public regarding Covid 19, so that there will be no more discrimination against patients, families & corpses of COVID-19" @divisihumaspolri

c. The process of changing username to [username] is as follows:

Table V. Changing Username

The text before changing the username	The text after changing the username
imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) [URL]	imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) [URL]

d. The process of removing redundant spaces is as follows:

Table VI. Removing Spaces

Text before removing excess whitespace	Text after removing excess whitespace
imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) [URL]	imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) [URL]

e. The process of deleting digits is as follows:

Table VII. Process of Deleting Digits

Text before deleting digits	Text after deleting digits
imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday (06/02/2020) [URL]	imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday [URL]

f. The process of removing punctuation is as follows:

Table VIII. Process of Removing Punctuation

The text before the punctuation is removed	Text after punctuation is removed
imagine that the data from the government is a lie, even though in Indonesia no one has the covid-19 virus. Wednesday [URL]	imagine that the data from the government is a lie even though in Indonesia no one has the covid-19 virus on Wednesday [URL]

g. The process of changing negative emoticons to [neg] and positive emoticons to [pos] is as follows:

Table IX. Changing Emoticons

The text before the emoticon deletion	Text after emoticon deletion
imagine that the data from the government is a lie even though in Indonesia no one has the covid-19 virus on Wednesday [URL]	imagine that the data from the government is a lie even though in Indonesia no one has the covid-19 virus on Wednesday [URL]

h. The process of converting non-standard or slang words into standard words is as follows:

Table X. Changing non-standard words

The text before it is converted into a standard word	The text after being converted into a standard word
imagine that the data from the government is a lie even though in Indonesia no one has the covid-19 virus on Wednesday [URL]	imagine that the data from the government is a lie even though in Indonesia no one has the covid-19 virus on Wednesday [URL]

i. The process of changing the negation to [not] is as follows:

Table XI. Changing the Negation

The text before changing the negation	The text after changing the negation
imagine that the data from the government is a lie even though in Indonesia no one has the covid-19 virus on Wednesday [URL]	imagine that the data from the government is a lie even though in Indonesia [not] there are those who got the covid-19 virus on Wednesday [URL]

[3] Stemming

After performing data cleansing, the next step is stemming process. Stemming is the process of getting basic words by removing prefixes, suffixes, insertions and confixes (a combination of prefixes and suffixes) and also removing keywords.

Table XII. Stemming

Text before stemming	Text after stemming
imagine that the data from the government is a lie even though in Indonesia [not] there are those who got the covid-19 virus on Wednesday [URL]	imagine that the government data is a lie even though in Indonesia [not] there is a covid-19 virus on Wednesday [URL]

[4] Tokenizing

Finally do tokenizing. Tokenizing is the process of cutting text into words, so that it becomes a token that can be analyzed.

Table XIII. Stemming

Text before tokenizing	Text after tokenizing
imagine that the government data is a lie even though in Indonesia [not] there is a covid-19 virus on Wednesday [URL]	Shadow, data, government, lies, Indonesia, [not], virus, covid [URL]

3.1 Classification

Classification is the process of entering data whose class is unknown based on data whose class has been defined previously. The classification process is divided into two phases, namely the learning step and the classification step.

3.2 *Term Frequency Inverse Document Frequency* (TF-IDF)

After preprocessing, the TF-IDF calculation is then carried out with the aim of obtaining important terms as keywords for each of these documents. In this study, the weighting was carried out by calculating the TF-IDF. TF (Term Frequency) represents the number of occurrences of a term in a document, while DF (Document Frequency) is the number of documents containing a term. The IDF shows the relationship between the availability of a term in all documents. This study displays 4438 words from 2981 tweets.

3.3 Test Data and Training Data

The Naïve Bayes classification algorithm uses training data (80%) to form a classification model, while test data (20%) is used to make predictions based on the classification model that is formed so that the value of the goodness of classification can be evaluated.

Table XIV. Training Data and Test Data

Classification	Positive	Negative	Total
Training Data	656	783	1815
Test Data	218	261	453
Total	874	1044	2269

3.4 Naïve Bayes Classifier

Nave Bayes classifier is a statistical-based classification method based on Bayes' theorem to classify data into predetermined classes. The data testing process is carried out by testing the data obtained from the results of Twitter sentiment using R by calculating training data and testing the performance of the Naive Bayesian Classifier algorithm.

3.5 *Confusion Matrix Naïve Bayes*

Furthermore, to measure the determination of the classification of the test data is done by forming a Confusion Matrix based on the prediction results. In the test results, 20% of the available data predicts tweets regarding the response of the Indonesian people to Covid-19.

Table XV. Prediction Results

Actual Class	Prediction Class		
	Positive	Netral	Negative
Positive	165	35	47
Netral	20	23	15
Negative	20	11	118

After the confusion matrix is formed in the next table, the classification determination is calculated using the following equation:

$$Accuracy = \frac{165 + 118}{165 + 118\ + 20 + 11}\text{ x100\%} = 0.674$$

$$Specivity = \frac{118}{118 + 20} = 0.8550$$

$$Sensitivity = \frac{165}{165 + 47} = 0.7783$$

The following is the calculation of the g-mean and AUC based on the equation:

$$G - Mean = \sqrt{0.8550 \text{ x } 0.7783} = 0.8822131$$

$$AUC = \frac{1}{2}(\ 0.8550 + \ 0.1183\ ) = 0.48665$$

The results of the classification determination on the Supporting Vector Machine obtained that the accuracy is 0.674, the g-mean is 0.8822131 and the AUC is 0.48665.

IV. Conclusion

Based on the results of the analysis and descriptions that have been put forward in previous chapters, the authors can make decisions for the conclusions of the problems in the study as follows:

1. Implementation of the Naïve Bayes Classifier method successfully carried out the sentiment analysis process well for the "Covid-19" tweet from the existing dataset, the reason was 2269 data, then divided into 1815 training data and 453 testing data or 80%: 20% data which exists.
2. The results of the classification of testing data show that the accuracy of testing is 95% where negative tweets from the Indonesian people are dominant towards Covid-19.

References

[1] R. Nasrullah, Teori dan Riset Media Siber (Cybermedia), vol. 28, Jakarta: Kencana Prenadamedia Group., 2014, p. 2–12.

[2] A. S. Cahyono, "Pengaruh media sosial terhadap perubahan sosial masyarakat di Indonesia," *Jurnal Publiciana,* vol. 9, no. 1, pp. 140-157.

[3] T. Yusuf, "Gaya hidup orang percaya berlandaskan Mazmur 91 : 1-16 dalam menyikapi masalah virus corona (Covid-19) masa kini," Institut Agama Kristen Negeri Toraja, Toraja, 2021..

[4] B. Liu , Opinion Mining & Summarization - Sentiment Analysis, Chicago: University of Illinois at Chicago, 2008.

[5] T. N. Suharsono, Kuspriyanto, F. A. Yulianto and B. Rahardjo, "Candidate Recommendations for Voting System Using Modified AHP," in *2020 14th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Bandung, 2020.

[6] G. Vinodhini and R. Chandrasekaran, "A comparative performance evaluation of neural network based approach for sentimen classification of online reviews," *Journal of King Saud University - Computer and Information Sciences,* vol. 28, no. 1, p. 2–12, 2016.

[7] V. Parkhe and B. Biswas, "Sentimen analysis of movie reviews: finding most important movie aspects using driving factors," *Soft Computing,* vol. 20, no. 9, pp. 3373-3379, 2016.

[8] M. Subhan, A. Sudarsono and A. R. Barakbah, "Classification of Radical Web Content in Indonesia using Web Content Miningand k-Nearest Neighbor Algorithm, Emit," *Int. J. Eng. Technol,* vol. 5, no. 2, p. 328, 2018.

[9] Kaggle, "Kaggle," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/search?q=Covid-19. [Accessed 12 July 2021].

[10] R. Morton, J. Hebel and R. McCarter, Panduan Studi Epidemiologi dan Biostatistika. Editor Bahasa Indonesia: Ferna Solekhah, Jakarta: EGC, 2009, pp. 54- 57.

tn2

ORIGINALITY REPORT

15%	9%	12%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

2%

★ S Sunarti, Irawan Dwi Wahyono, Hari Putranto, Djoko Saryono et al. "Lexicon-Based Features on Naive Bayes Modification for Classification of Chinese Film", 2022 International Seminar on Application for Technology of Information and Communication (iSemantic), 2022

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		