Jurnal Pendidikan dan Konseling
Volume 4 Nomor 6 Tahun 2022
E-ISSN: 2685-936X dan P-ISSN: 2685-9351
Universitas Pahlawan Tuanku Tambusai

Sentiment Analysis of Twitter User's Opinions on Government's Performance in dealing with COVID-19 in Indonesia

Bagus Sujiwo[1], Antoni Wibowo[2]
Computer Science Department, BINUS Graduate Program- Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480[1,2]
Email: Bagus.sujiwo@binus.ac.id

Abstrak

Saat ini, cukup banyak masyarakat Indonesia yang menggunakan Twitter, sebuah jejaring sosial media yang menyediakan informasi berupa produk, iklan, dan promosi mengenai kritik, saran suatu isu, dan opini publik. Penelitian ini bertujuan untuk menyederhanakan dan meningkatkan pendeteksian suatu opini tanpa menggunakan metode yang memakan waktu, seperti kuesioner. Selain itu, ia membuat kumpulan data berdasarkan tweet pengguna berbahasa Indonesia. Label data dikumpulkan menggunakan metode k-fold cross-validation yang dibagi menjadi 10 bagian. Metode klarifikasi analisis sentimen dilakukan melalui studi banding antara tiga metode, yaitu Naive Bayes (NB), Support Vector Machine (SVM), dan Long Short-Term Memory (LSTM). Ketiga metode memberikan hasil yang sesuai untuk setiap sifat kepribadian tetapi SVM sedikit mengungguli yang lain.
Kata kunci: *Pemrosesan Bahasa Alami, Penambangan teks, pembelajaran mesin, COVID-19*

Abstract

Nowadays, quite a lot of Indonesians use Twitter, a social media network that provides information in the form of products, advertisements, and promotions regarding criticism, suggestions on issues, and public opinion. The study aims to simplify and improve the detection of an opinion without using time-consuming methods, such as a questionnaire. Also, it creates a dataset based on the tweets of Indonesian-speaking users. Data labels were collected using the k-fold cross-validation method divided into 10 parts. Sentiment analysis clarification method was performed through a comparative study between three methods, namely Naive Bayes (NB), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). The three methods provided suitable results for each personality trait but the SVM slightly outperformed the others.
Keywords: *Natural Language Processing, Text mining, machine learning, COVID-19*

INTRODUCTION

Social network is developing into a very popular communication tool, hence people use it as a medium to communicate. An example of such is Twitter, which is used for promoting products, advertising, political campaigns, as well as expressing opinions related to criticism, suggestions, issues, and public opinions. There are quite a lot of daily active Twitter users, and according to Twitter's 3rd quarter 2019 financial report, Indonesia is one of the countries with the largest growth of daily active Twitter users. Based on this development, it becomes one of the media used for conducting sentiment analysis on various topics. Therefore, this study performs sentiment analysis on a topic that is currently

trending on Twitter, namely "Corona". The analysis is a text mining used to classify tweets' polarities to observe whether the opinion given is positive, negative, or neutral (Ardiani et al., 2020).

The coronavirus or Covid-19 pandemic was first detected in Wuhan in December 2019 (Stokes et al., 2020), and within an hour of its spread, this virus was immediately discussed on Twitter. Currently, the coronavirus is becoming an international concern based on a large number of victims. The rapid spread has worried the people, specifically Indonesians, thereby causing various opinions to emerge. This opinion is then analyzed using sentiment analysis to determine its polarity. Sentiment analysis is part of text mining used for classifying text polarity to ascertain whether tweets are positive, negative, or neutral. Previous studies on sentiment analysis have been conducted smoothly for various purposes. Among them is the retweet analysis of television programs, which serves as a reference for rating a TV show (Berlian, 2019). Subsequently, a web crawler was used to collect tweets with the pre-processing text mining method. The study produces an application for converting the collected tweets into data that are processible as needed (Aditya, 2015). The difference between Berlian (2019) and Aditya (2015) was that Aditya (2015) employed a web scraping technique to extract all data or tweets with the keyword 'Corona', including replies, likes, and retweets. The scraped data is further analyzed to determine the public opinion about the "Coronavirus". This article explains how the public sentiment analysis of safe tweets shows the percentage of positive, negative, and neutral tweets. This helps to determine the extent to which the virus affected Indonesia based on public opinion via Twitter. In Saleh & Menai (2014), the Naive Bayes machine learning method was used to determine attributes.

Several ideas have emerged over the years on how to achieve quality results from web classification systems. This led to the use of different approaches to analyze sentiments, such as Naive Bayes and Bayesian Networks, NNs, DTs, support vector machines (SVM), etc. The Naive Bayes model is popular in machine learning applications because it is easy to make each attribute contribute to the same final decision and is independent of other attributes (Xhemali et al., 2009).

METHODS

The Naive Bayes Classifier method requires two stages in the text classification process, namely the training and classification. In the training stage, a process was performed on a sample that served as a data representation. Furthermore, the prior probabilities for each category based on the sample data were determined. In the second phase, the data category value was obtained based on the terms appearing in the classified data. The Naive Bayes classification assumes the presence or absence of certain class characteristics has nothing to do with that of other classes [9]. The Naïve Bayes theorem is expressed as, Where Posterior states (Probability Xk in Y) can be calculated from prior states (Probability Y in Xk divided by the sum of all probabilities Y in all Xi).

Several Indonesian-language tweets' data about Corona found on Twitter were used and they mainly contain the expressions of joy, love, anger, sadness, and fear. The tweets' numbers taken per emotion was 500, and it sums up to 2,500 pieces. A table was created in the MySQL database, called corpus_tweets to store tweets, while Tala's stopword and rootword tables were imported from the internet which was later used for stemming and stopword removal processes. Tweets' data were searched and retrieved with the Twitter API using the keyword "Corona Indonesia" and five emotional hashtags. Furthermore, the system uses the user ID and consumer key ID to access and retrieve the tweets in question. The data obtained were then sorted manually to ensure the tweets used are pure Indonesian text without images, which are stored in the corpus_tweets table.

Table 1. Example of Training Tweet Data

Class	Tweet
Cinta	@B_Zaenuri terimakasi kepada pemerintah atas bantuan kepada rakyat yang membutuhkan #coronaindonesia @muyanneni semoga corona bisa cepet hilang #corona indonesia
Senang	@KalidYanuar salah satu hal positive dari corona adalah makin dekat dengan keluarga #coronaindonesia @mibahwiyono Para HRD, Personalia agar bergerak cepat. Jika perlu karyawan tidak punya REKENING kasih KASBON buat BUKA REKENING. terima Kasih @BPJSTKinfo #coronaindonesia #Corona
Marah	@SaaeBunglon Ditengah hoax dan fitnah yang membuat resah Mari kita jangan marah Jokowi saja tabah Yuk Mending kita #GoyangJempol ajaaaahh https://t.co/GdN7Wi0UrU
Sedih	@coronareport_id Tanggal: 2020-8-12 Terkonfirmasi: 130718 Meninggal: 5903 Sembuh: 85798 #CoronaOutbreak #CoronaIndonesia @detikcom "Ingat dalam kondisi ekonomi sulit seperti ini, pasti ada dampak kepada kejahatan, tapi jangan kambinghitamkan semua pada napi asimilasi," kata Yasonna. #coronaindinesia #wabahviruscorona#yasonnalaoly#Napi
Takut	@yonsas888 sample frozen chicken wings yg dikirim dari Brazil ke China setelah di test hasilnya positif ada corona virusnya! wow bahaya jg yah bisa sampe dimakanan gitu!#corona#coronaindonesia @CNNIDdaily Bahaya Covid-19 ada di mana-mana, hindari penyebarannya saat masuk ke dalam rumah dengan langkah-langkah berikut. Ayo lawan penyebaran virus corona! #CoronaIndonesia #MediaLawanCovid19

Preprocessing

In this stage, the indexing process was applied to Information Retrieval that the data need to pass through for it to represent the information needed by the user. Pre-processing is also needed in the classification stage because it requires more specific data characteristics, such as word frequency. The procedures in this pre-processing include:

Tokenizing

At this stage, the words in the tweets were tokenized and all punctuation marks, as well as symbols that do not represent the document's contents, were removed. The steps involved in tokenizing were as follows:

a. Read the entire text line as a single-sentence tweet.
b. Take each token in the sentence as a separator between tokens and perform case-folding.
c. Remove all kinds of non-text symbols, punctuation marks, mentions e.g. @jokowi, and hashtags such as #prabowo.
d. Save the tokens in an ArrayList as one tweet.

Table 2. shows an example of a standard word dictionary:

Normalization

At this stage, non-standard words were changed into standard forms using a standard word dictionary to avoid duplicating words with the same meaning. The standard word dictionary was

Standard Word Dictionary			
Non Standard Form (1)	**Non Standard Form (2)**	**Non Standard Form (3)**	**Standard Form**
Knp	napa	-	Kenapa
Bpk	Bp	bokap	Bapak
Anjenk	asu	anjir	Anjing
Elo	ente	lo	Kamu

obtained from the internet with various additions by the author. The steps include:

a. Read the tokens in the ArrayList and match them with non-standard words in the standard dictionary.
b. When the token is similar to a non-standard word in the dictionary, the system replaces the word with the standard form corresponding to the dictionary.
c. Otherwise, it continues with the stemming process.

Stemming

In this stage, words are formed from affixes to unaffixed or basic words using a root word dictionary, such as Tala's Indonesian root word. The stemming steps were described as follows:

a. Check each token with the Indonesian root word dictionary.
b. When the word token is similar to the one in the root word dictionary, then the token is the root word and does not change.
c. When the token is not similar to the word in the root word dictionary, delete all its prefixes and suffixes.

Stopword Removal

At this stage, words occurring too often with no meaning, such as prepositions, conjunctions, etc., appeared through Tala's Indonesian stopword dictionary. The steps for removing stopword in this study include:

a. Read all stemming data in ArrayList.
b. Check each token with Tala's stopword dictionary.
c. When the token is a stopword, the system automatically removes the token from the ArrayList.
d. Otherwise, the token is stored in the database.

It is important to note that stemming was performed before stopword removal because there were several Indonesian stopword words with affixes. For example, the word "permissible" having the root word "permissible" is also a stopword that needs to be removed. Therefore, when stemming is conducted last, the word "may" that has been stemmed is not likely to be deleted by the system, thereby becoming a residue.

Calculating Term Frequency (Word Frequency)

In this phase, the words resulting from the four processes above are counted by the number of their occurrence or frequency per class.

Feature extraction.

The Term Frequency-Inverse Document Frequency (TF-IDF) is a method of assigning weight to the relationship between keywords or terms in a document. This method combines two concepts for calculating weights, namely Term frequency (TF), which is the occurrence of words in a sentence, and

Document Frequency (DF), defined as the number of sentences in which a word appears (Priatna, 2019).

Classification process.

In this section, a comparative study was conducted between the two classification methods of sentiment analysis to determine the one with the higher accuracy. The sentiment analysis methods employed include:

1. Naïve Bayes

   In Suara.com, the Naive Bayes Algorithm was employed when obtaining the students' estimated study time using data mining techniques, namely classification, to predict the timeliness of studies based on existing training data. Based on the Naive Bayes method, data with an undetermined sentiment label was predicted using the previously trained classifier.

2. Support Vector Machine (SVM)

   Support Vector Machine or SVM is a good prediction technique in classification and regression Luqyana (2018). Based on the method, data with an undetermined sentiment label was predicted according to the previously trained classifier.

3. Evaluation method

   A comparative study of SentiWordNet-based accuracy was performed on the Support Vector Machine classification method with different kernels. This was conducted by comparing the accuracy, precision, recall, and F Measure obtained from the comparison between the sentiment prediction and class label determination results based on SentiWordNet using variables.

RESULTS AND DISCUSSIONS

Preprocessing result. In this section, the results of collecting, processing tweets, and data cleaning were discussed.

1. Case Conversion

   In this Phase, the sentence goes through capitalization changes from uppercase to lowercase. Examples of the case conversion results were shown in Table 3.

Table 3. Example Case Conversion

No	Previous Tweet	The Next Tweet
1	Sejak awal pemerintah menyepelekan covid-19, hingga akhirnya kita disuruh berdamai dengannya...	Sejak awal pemerintah menyepelekan covid-19, hingga akhirnya kita disuruh berdamai dengannya...

2. Data Cleaning

   In the data cleaning process, words having hashtag elements, Twitter usernames, URL links, and other symbols were cleansed from tweets. This is conducted with a query regex as follows:

$$(@[A-Za-z0-9]+)|([\^0-9A-Za-z\ t])|(\w+:\/\/\S+)\ \#(4.2)$$

Table 4. Example Case Conversion

No	Previous Tweet	The Next Tweet
1	Sejak awal pemerintah menyepelekan covid-19, hingga akhirnya kita disuruh berdamai dengannya...	Sejak awal pemerintah menyepelekan covid hingga akhirnya kita disuruh berdamai dengannya

3. Sentence Translation

All tweets were first translated to English using Google Translate because the SentiWordNet lexicon is entirely composed of English. Table 5 shows examples of tweets obtained and their translation results in English.

Table 5. Example of Translated Tweet Results

No	Tweet Orisinil	Tweet in Inggris
1	memaksa mudik ditengah pandemi jeruji dunia dilanda pandemi covid terlepas indonesia menekan angka penyebaran covid pemerintah resmi melarang masyarakat lakukan mudik	forcing going home in the middle of a pandemic with the world being hit by a covid pandemic despite indonesia suppressing the spread of covid the government officially prohibits people from going home
2	pandemi covid indonesia kebijakan pemerintah menyesuaikan negara	covid pandemic indonesian government policy of adjusting the country
3	pemerintah daerah berusaha maksimal menangani covid pemerintah pusat bikin amburadul	the local government is trying its best to deal with covid the central government is making chaos
4	indonesia orang kayak orang dermawan pemerintah anggaran mengulur terputus rantai penyebaran covid harap pemerintah memikirkan buruknya jangka pendek jangka keputusan	indonesia people like generous people government budget stalled the distribution of covid distribution hope the government think of the bad short-term decision
5	slawi pemerintah kabupaten masyarakat disiplin mematuhi protokol kesehatan pencegahan covid	slawi district government community discipline comply with covid preventive health protocol

4. Tokenization

The main purpose of tokenization was to divide a text into smaller parts called tokens. The results obtained from this were processed sequentially.

5. Removal of Stop Words

Stopword refers to the words ignored in the NLP agreement because they do not provide a better sentiment value. Words including 'is', 'the', 'with', as well as 'and' were removed from the processed sentence.

Table 6. Example of Stop Words Removal Results

No	Before Tweet	After Tweet
1	Slawi district government community discipline comply **with** covid preventive health protocol	Slawi district government community discipline comply covid preventive health protocol

6. Lemmatization

Lemmatization was passed by the tweets' data to convert the words into the most basic. Examples of tweets obtained and the results after the lemmatization stage were shown in Table 7:

Table 7. Lemmatization Results

No	English Tweet	Tweet After Lemmatization
1	forcing going home in the middle of a pandemic with the world being hit by a covid pandemic despite indonesia suppressing the spread of covid the government officially prohibits people from going home	force go home middle pandemic world hit covid pandemic despite indonesia suppress spread covid government officially prohibit people go home
2	covid pandemic indonesian government policy of adjusting the country	covid pandemic indonesian government policy adjust country
3	the local government is trying its best to deal with covid the central government is making chaos	local government try best deal covid central government make chaos
4	indonesia people like generous people government budget stalled the distribution of covid distribution hope the government think of the bad short-term decision	indonesia people like generous people government budget stall distribution covid distribution hope government think bad short term decision
5	slawi district government community discipline comply with covid preventive health protocol	slawi district government community discipline comply covid preventive health protocol

Table 8. SVM Kernel Method Comparison

No	Methods	Accuracy	Precision	Recall	F1-Score
1	SVM Linear	88.5%	84.8%	74.9%	78.4%
2	SVM RBF	81%	40.5%	50%	44.7%
3	SVM Polynomial	81%	40.5%	50%	44.7%
4	Gaussian Naive Bayes	39.8%	21.85%	83.83%	34.67%
5	Multinomial Naive Bayes	82.11%	54.56%	36.54%	43.76%
6	LSTM	87.8%	70.9%	60.99%	65.57%

The results showed that the SVM model with a linear kernel has the greatest accuracy value of 88.5%. This accuracy was due to the number of positive sentiment datasets not corresponding with that of negative, thereby causing classification bias. This was observed from the results of SVM RBF, SVM Polynomial, and Multinomial Naive Bayes, in which all three have similar accuracies. The specified accuracy also corresponds to the scenario that all train datasets were negatively classified.

The Gaussian Naive Bayes model recorded the highest recall value of 83.83% but has low precision. This simply indicates the Gaussian Naive Bayes model was not good at classifying sentiment text. The LSTM model has a fairly good score but was still inferior to the linear kernel SVM. Since data were retrieved manually and processed through SentiWordNet values, there are potential errors when labeling the data. This error caused the LSTM model not being able to converge to the optimal value. It was also observed that SVM was better because it has less overfitting tendency than LSTM (Mondal et al., 2012). Meanwhile, in the study conducted by M. A. Nurrohmat and Azhari SN, entitled

"Sentiment Analysis of Novel Review Using Long Short-Term Memory Method", LSTM produces high accuracy compared to other methods (Nurrohmat, 2019).

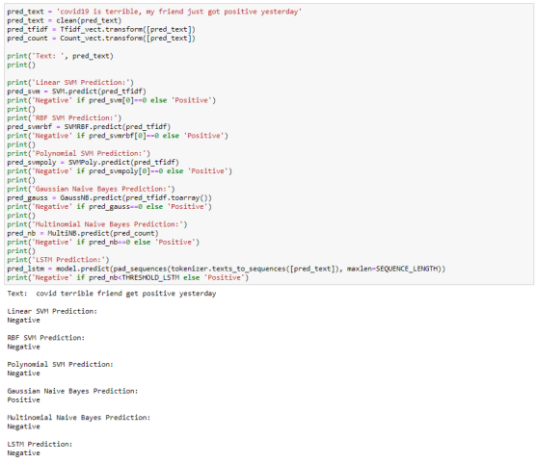

Figure 1. All Negative Prediction Scenario

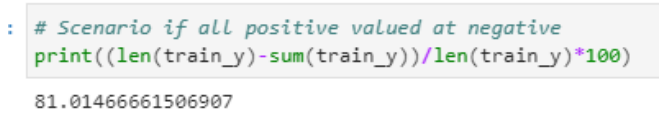

Figure 2. SVM Model Manual Prediction Results

CONCLUSION

Based on the processed and analyzed data, this study answered the questions previously stated in the problem formulation. It was discovered that the majority of Indonesians have negative sentiments about COVID-19 as evidenced by 29,883 processed tweets. This classification was quite effective in measuring the public's positive and negative opinion regarding COVID-19, with an accuracy of 88.5%. In conclusion, it was observed that the SVM model is the most effective in predicting sentiment sentences from Twitter.

REFERENCES

Aditya, B. R. (2015). Penggunaan Web Crawler untuk Menghimpun Tweets dengan Metode Pre-Processing Text Mining. *JURNAL INFOTEL - Informatika Telekomunikasi Elektronika*, *7*(2), 93. https://doi.org/10.20895/infotel.v7i2.35

Ardiani, L., Sujaini, H., & Tursina, T. (2020). Implementasi Sentiment Analysis Tanggapan Masyarakat terhadap Pembangunan di Kota Pontianak. *Jurnal Sistem Dan Teknologi Informasi (Justin)*, *8*(2), 183. https://doi.org/10.26418/justin.v8i2.36776

Berlian, T. ., Herdiani, A., & Astuti, W. (2019). *Analisis Sentimen Opini Masyarakat terhadap Acara Televisi pada Twitter dengan Retweet Analysis dan Naive Bayes Classifier*. *Vol.6 No.2*, 8660–8669.

Fadrial, Y. E. (2021). Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa. *INTECOMS: Journal of Information Technology and Computer Science*, *4*(1), 20–29. https://doi.org/10.31539/intecoms.v4i1.2219

Fauzi, M. A., & Adinugroho, S. (2018). *Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking Image Processing and*

Computer Vision View project Food Image Classification, Retrieval, and Analysis View project. *February*.

Luqyana, W. A. (2018). *Instagram dengan Metode Klasifikasi Support Vector Machine*.

Mondal, A., Kundu, S., Chandniha, S. K., Shukla, R., & Mishra, P. K. (2012). Comparison of Support Vector Machine and Maximum Likelihood Classification Technique using Satellite Imagery. *International Journal of Remote Sensing and GIS*, *1*(2), 116–123.

Nurrohmat, M. A., & SN, A. (2019). Sentiment Analysis of Novel Review Using Long Short-Term Memory Method. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(3), 209. https://doi.org/10.22146/ijccs.41236

Stokes, E. K., Zambrano, L. D., Anderson, K. N., Marder, E. P., Raz, K. M., El Burai Felix, S., Tie, Y., & Fullerton, K. E. (2020). Coronavirus Disease 2019 Case Surveillance — United States, January 22–May 30, 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*(24), 759–765. https://doi.org/10.15585/mmwr.mm6924e2

Suara.com. (2019). *Pengguna twitter indonesia terbesar ke empat di dunia*. Suara.Com.

W, W., Priatna, W., & Hidayat, J. S. (2019). *A Implementasi Term Frequency – Inverse Document Frequency (TF-IDF) dan Vector Space Model (VSM) untuk Pencarian Berita Bahasa Indonesia*.

Xhemali, D., J. Hinde, C., & G. Stone, R. (2009). Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science*, *4*(1), 16–23.