JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1357

SENTIMENT ANALYSIS DUE TO "MUDIK" PROHIBITED OF COVID-19 THROUGH TWITTER

Sabar Sautomo[1], Noor Hafidz[2], Yuni Eka Achyani[3], Windu Gata[4]

[1,2,3,4]Computer Science
[1,2,3,4]STMIK Nusa Mandiri Jakarta
[1,2,3,4]https://www.nusamandiri.ac.id
[1]14002304@nusamandiri.ac.id, [2]14002298@nusamandiri.ac.id, [3]yuni.yea@nusamandiri.ac.id, [4]windu@nusamandiri.ac.id

Abstract— "Mudik" is a habit every year for the people of Indonesia to return to their hometowns before the Eid. The existence of the Corona Virus pandemic (COVID-19) hit all over the world, including Indonesia, resulting in a ban from the government to do Mudik. Social media such as Twitter is often used as an expression of some people in commenting on something like the ban on Mudik. Comments on Twitter that are often known as tweets can be used as material for sentiment analysis. However, it is not easy to do sentiment analysis on Twitter, especially comments in Indonesian, because the text is not structured. This study uses data from Indonesian-language tweets containing the word "Mudik," the algorithm model used in this study, Naïve Bayes Classifier and Support Vector Machine, is compared to get accuracy, precision, recall, and F1-score values. From this research, it was concluded that the Naïve Bayes algorithm and Support Vector Machine performed well enough to predict the sentiment of tweets about Mudik on Twitter social media. Naïve Bayes with an accuracy of 82% and f1-score 0.8, while Support Vector Machine with an accuracy of 87% and f1-score 0.87.

Keywords: Mudik, Twitter, Naïve Bayes Classifier, Support Vector Machine.

Abstrak— *Mudik merupakan suatu kebiasaan setiap tahun bagi masyarakat Indonesia untuk pulang ke kampung halamannya menjelang perayaan Idul Fitri tiba. Adanya pandemi Corona Virus (COVID-19) melanda di seluruh dunia termasuk Indonesia, berakibat adanya larangan dari pemerintah untuk melakukan mudik. Media sosial seperti twiiter sering dijadikan ungkapan beberapa orang dalam mengomentari suatu hal seperti larangan mudik tersebut. Komentar pada twitter yang sering dikenal dengan istilah tweets dapat dijadikan bahan untuk dilakukan analisis sentimen. Namun tidak mudah dalam melakukan analisis sentimen pada twitter terutama komentar dalam bahasa Indonesia karena teks tersebut tidak terstruktur. Penelitian ini menggunakan data dari tweets berbahasa Indonesia yang mengandung kata "Mudik", model algoritma yang digunakan dalam penelitian ini yaitu Naïve Bayes Classifier dan Support Vector Machine dikomparasikan agar mendapatkan nilai accuracy, precision, Recall dan F1-score. Dari penelitian ini dihasilkan bahwa algoritma Naïve Bayes dan Support Vector Machine memiliki performa yang cukup baik untuk memprediksi sentimen dari suatu tweets mengenai mudik pada media sosial Twitter. Naïve Bayes dengan akurasi sebesar 82% dan f1-score 0.8, sedangkan Support Vector Machine dengan akurasi sebesar 87% dan f1-score 0.87.*

Kata Kunci*: Mudik, Twitter, Naïve Bayes Classifier, Support Vector Machine.*

INTRODUCTION

For most Indonesians, there is an annual routine that should not be missed, known as Mudik. Mudik is a culture that has become a habit for Indonesian to return to their hometowns once every year, usually done before the Eid celebration [1]. However, it seems that something different happens in 2020 because of Government policy, which bans Mudik. The policy is closely related to the global impact of pandemic COVID-19 in Indonesian.

In March 2020, the World Health Organization (WHO) has determined that the Coronavirus Novel (COVID-19) has become a global pandemic [2], and Indonesia is no exception. In Indonesia, the pandemic has a significant impact on all aspects of people's life, including both economic and social.

Each country throughout the world applies its policies to minimize the spread of COVID-19, for example, limiting people gathering, implementing social distancing, and even implementing lockdowns in one or more regions of the country.

Restricting people's movement from one region to another is also one of the policies adopted by many countries. This includes the Indonesian government prohibiting its citizens from traveling far out of town, especially for Mudik on Eid al-Fitr [3]. According to the author, this policy is exciting to observe, especially on the aspect of the general public's sentiment related to the phenomenon of Indonesian Mudik tradition. The analysis of public sentiment on the policy of Mudik Ban will be done by using public comments on social media Twitter.

Twitter is one of the internet networks in the form of a social media platform with a lot of users in the world [4]. Twitter users in Indonesia are claimed to be one of the countries with the most significant growth of daily active Twitter users in the world where active users around the world have increased by 17 percent to 145 million [5].

The modeling method which is used in this study is the Naïve Bayes Classifier algorithm and Support Vector Machine, where the method has been used before in determining the sentiment analysis of the Indonesian Police Mobile Brigade Corps, with an accuracy rate of 86.96% [6]. Another related study is Sentiment Analysis of KPK's Capture Operations According to the Community, with the Weight Particle Swarm Optimize operator for Support Vector Machine (PSO) with the accuracy of 83.79% and AUC 0.910. In comparison, for Naïve Bayes (PSO), the accuracy is 80.21% and AUC 0.771. There is a 3.58% accuracy difference [7].

In research [8], it is known that the Support Vector Machine (SVM) classification method provides higher accuracy for the Indonesian Hate Speech tweet sentiment classification than the Naïve Bayes Classifier (NBC) classification method. While in research [9] on Twitter data sentiment analysis on Islamaphobia, it was concluded that the Naïve Bayes model with SMOTE accuracy performance is 92.75%. In comparison, the Support Vector Machine method with SMOTE accuracy performance is 93.2%.

The purpose of this study is to obtain the best method for analyzing sentiment from Twitter data on the Indonesian government's policy of the Mudik ban due to the COVID-19 pandemic. The method is by utilizing a lexical-based NLTK library in Python, namely Valence Aware Dictionary and Sentiment Reasoner (VADER), for automatic labeling of Twitter crawling data [10]. Using the tweet data that has been labeled and then two models will be made using the Naïve Bayes Classifier algorithm and Support Vector Machine. After that, it will be compared to the accuracy, precision, recall, and F1-score values to get the conclusion of the study.

MATERIALS AND METHODS

In this study, we will explain the steps in getting research results which consist of

a. Data Collecting Method

The data collection stage is taken from Indonesian social media Twitter comments or tweets with the word "Mudik" as a keyword. It is done by mining the data utilizing Twitter's Application Programming Interface (API) [11], then it is stored in a file with the JavaScript Object Notation (JSON) format.

Figure 1. Process of collecting data from twitter

b. Data Processing Method

The data processing or pre-processing stage is crucial because it can determine the efficiency of the next step [12]. There are several methods of pre-processing in Twitter data extraction that can be done [13]; this study uses several stages of data extraction:

a. Case Folding
Change all the letters in the data into lowercase letters.
Example
From: Larangan Mudik bagi Kita
Become: larangan mudik bagi kita

b. Cleaning
Cleaning is the process of cleansing characters other than letters, including removing URLs, usernames, hashtags, website links on the data.
Example
From: @ kita http://url.com
Become: kita

c. Tokenization
The Dividing Process
The process of breaking down a group of words (sentences/phrases) into words with a specific meaning.
Example
From: mudik hari raya
Become: mudik | hari | raya

d. Stemming
the process of returning words to their essential words. To clear a word based on its original spelling.

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1357

Example
From: mengatakan
Become: kata

e. Stop word Removal
Stop words are common words that usually appear in large numbers and are considered to have no meaning. Stop words are commonly used in information retrieval tasks.

Examples: “di”, ”oleh”, “pada”, ”sebuah”, ”karena” and etc.

f. Translate into English
In this stage, the data will be translated into English because the original data is in Bahasa Indonesia.

Example
From: tahun ini
Become: this year

g. Labeling
The next stage is the process of data labeling on the data that has been processed before. So, it will be more comfortable in the process of calculating the polarity of the tweets taken. Labeling can be positive, neutral, or negative.

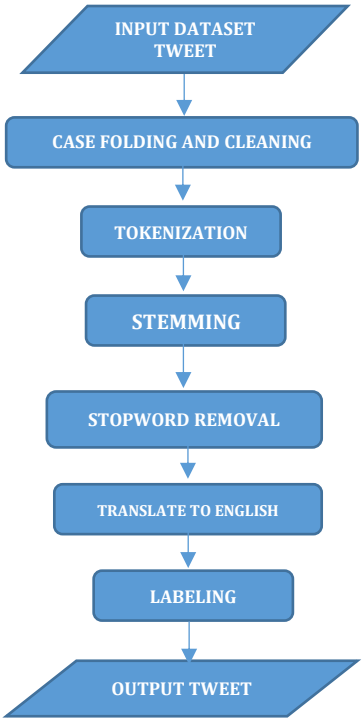

Figure 2. Data Processing Stages

c. *Modeling*
After processing the data in the previous stage, the next step is to split the data into two categories, Data Training and Data Testing. This process is to get data extraction, which will be used for Machine Learning modeling in predicting sentiments analysis of Twitter data.

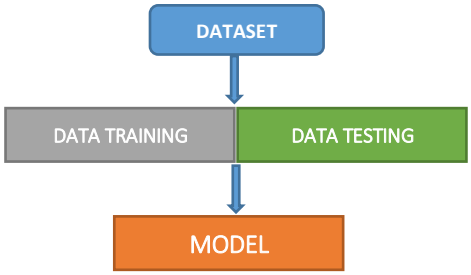

Figure 3. Dataset split process for modeling

d. Evaluation Model
The result of sentiment prediction analysis from the previously created model will be evaluated to get the value of accuracy, precision, recall, and F1-score.

RESULT AND EXPLANATION

This section will explain the results of the study based on the process of the method that has been used. The author uses Jupyter as a tool for data processing, modeling, and evaluating to get the result of this study.

The following is the result and the explanation:

1. Data Collecting
The data used in this study is based on a real-time tweet in Bahasa Indonesia from the Twitter web, obtained by the author on May 12, 2020, at 16.32 WIB. 5,437 tweets are collected successfully.

Table 1 is an example of the results of the tweet data that was successfully obtained.

Table 1. Data *Tweet* from Twitter Web

No	*Tweets*
1	RT @TweetPolitikID: Larangan mudik untuk memut.....
2	@TweetPolitikID Yok tahun ini tidak mudik dul...
3	RT @DilaNasution6: Ada beberapa hal yg boleh b...
4	RT @lebahquu: ADA SALDO 50K BUAT 2 ORANG\n\nRT...
5	RT @wisnu_prasetya: Doublespeak: Bukan mudik t...

No	*Tweets*
6	RT @ayubsr: biasanya udah rame berita arus mud...
7	RT @zlametabidin: Ketegasan Aparat Polisi di j...
8	Ayo patuhi pemerintah, untuk tidak mudik demi k...
9	@TweetPolitikID Putus rantai covid19 mulai dar...
10	Untuk yang masih bandel, nih simak bagus2 #Apa...

2. Data Processing

The tweet data that was successfully collected, as in table 1, is an example of the results of the twitter data mining that was successfully obtained by using the Twitter API. Then the next stage is called pre-processing, which consists of Case Folding, Cleaning, Tokenization, Stemming, Stop-word Removal, Translating, and labeling.

Table 2. Data Tweets after pre-processing before translating and labeling

No	*Tweets*
1	rt larangan mudik untuk memut.....
2	tahun ini tidak mudik dul...
3	rt ada beberapa hal yg boleh b...
4	rt ada saldo 50k buat 2 orang rt
5	rt doublespeak: bukan mudik t...
6	Rt biasanya udah rame berita arus mud...
7	rt ketegasan aparat polisi di j...
8	ayo patuhi pemerintah, untuk tidak mudik demi k...
9	putus rantai covid19 mulai dar...
10	untuk yang masih bandel, nih simak bagus

The process of labeling tweets is done by using the VADER library in Python, but it requires the text to be in English before being labeled. Therefore, for all the tweet data that was successfully taken must be translated into English first. The process of translating Indonesian texts into English itself uses the googletrans library in Python [14].

After the data is successfully translated into English, then the compound value is calculated for labeling.

Table 3. Data *tweet* after pre-*processing translated* and labeled

No	*Tweets*	*Compound*	Label
1	rt prohibition forth to break the chain of Cov...	-0.02304	negative
2	this year's first homecoming for the ...	-0.4404	negative
3	rt, several things could operate...	0.6124	positive
4	rt 50k fo two people rt	-0.1363	negative
5	rt doublespeak not going home but returned hom...	-0.3491	negative
6	The household usually has much news about mud ...	0.0000	neutral
7	rt firmness police officers on the road prohib...	0.0000	neutral
8	let's obey the government, not to go home for the sake of	0.2376	positive
9	disconnect COVID chain 19 from now with no hom...	-0.2960	negative
10	for those who are still stubborn, this is good	0.4602	positive

Figure 4 below illustrates the distribution of data that has been labeled positive, neutral, and negative.

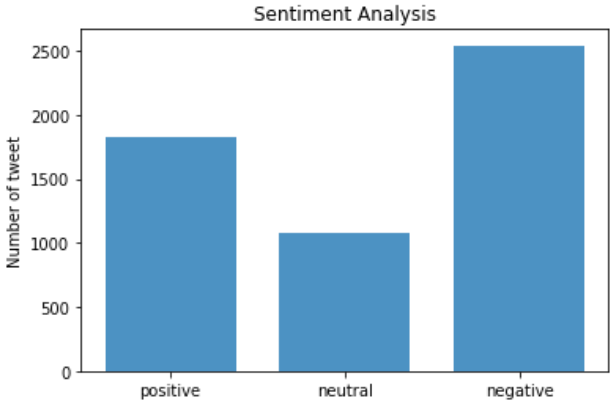

Figure 4. Histogram of Sentiment Analysis base on the number of tweets

3. Modeling

At this stage, a modeling process is carried out on the data that has been previously pre-processed. From the amount of data available, the process is split into two categories, data training, and data testing with a ratio of 70%: 30%. Furthermore, the modeling process using the Naïve Bayes Classifier algorithm and Support Vector Machine to produce accuracy, precision, recall, and F1-score values.

4. Evaluating Model

Here are reports of accuracy, precision, recall, and F1-score which are obtained from the

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1357

model with two algorithms, Naïve Bayes Classifier and Support Vector Machine,

a. Naïve Bayes Classifier
Table 4 is the result of an experiment using the Naïve Bayes Classifier algorithm.

Table 4. The value of accuracy, precision, Recall dan F1-score on Naïve Bayes algorithm

	Precision	Recall	F1-score	Support
Negative	0.89	0.67	0.77	325
Neutral	0.87	0.83	0.79	754
Positive	0.72	0.89	0.79	553
Accuracy			0.82	1632
Macro avg	0.83	0.79	0.80	1632
Weighted avg	0.82	0.81	0.81	1632

b. Support Vector Machine
Table 5 shows the results of the experiments using the Support Vector Machine algorithm.

Table 5. The value of accuracy, precision, Recall dan F1-score on Support Vector Machine algorithm

	Precision	Recall	F1-score	Support
Negative	0.88	0.81	0.84	325
Neutral	0.87	0.92	0.89	754
Positive	0.87	0.86	0.87	553
Accuracy			0.87	1632
Macro avg	0.87	0.86	0.87	1632
Weighted avg	0.87	0.87	0.87	1632

CONCLUSION

From the results of sentiment analysis on Indonesian tweets containing the word "Mudik," it can be concluded that twitter as a social media platform with a large number of users, providing APIs for parties to access their data for research purposes. Twitter data that is accessed can be stored in the JSON file format, which can then be processed to create a prediction on sentiment analysis of particular topics, in this case, the sentiment analysis related to Mudik. In this study, the Naïve Bayes algorithm and Support Vector Machine has excellent performance to predict sentiment analysis from tweet data containing the word "Mudik" on Twitter social media. The predictive value of each algorithm is 82 % with f1-score 0.8 and 87% with f1-score 0.87, respectively.

In addition to the conclusions above, this study also tries to provide suggestions for further research, such as testing using different algorithms from this research or by adding more data and more mature data preparation so that better results can be obtained.

REFERENCES

[1] B. B. Soebyakto, "Mudik Lebaran," *J. Ekon. Pembang. J. Econ. Dev.*, vol. 9, no. 1829–5843, pp. 61–67, 2015.

[2] D. Cucinotta and M. Vanelli, "WHO declares COVID-19 a pandemic," *Acta Biomed.*, vol. 91, no. 1, pp. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.

[3] "Tok! Pemerintah Larang Mudik Lebaran Mulai 24 April 2020," *https://news.detik.com/*.

[4] "Pengguna Aktif Harian Twitter Indonesia Diklaim Terbanyak," *Kompas.Com*, 2019.

[5] "Twitter Klaim Pengguna Harian Terbanyak Berasal dari Indonesia," *https://wartakota.tribunnews.com/2019/10/30/twitter-klaim-pengguna-harian-terbanyak-berasal-dari-indonesia*, 2019.

[6] B. Pratama *et al.*, "Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and NB Methods," *J. Phys. Conf. Ser.*, vol. 1201, no. 1, pp. 0–12, 2019, doi: 10.1088/1742-6596/1201/1/012038.

[7] M. I. Komputer and K. J. Pusat, "Sentimen Analisis Operasi Tangkap Tangan KPK Menurut Masyarakat Menggunakan Algoritma Support Vector Machine , Naive Bayes Berbasis Particle Swarm Optimizition," vol. 12, no. 3, pp. 230–243, 2019, doi: 10.30998/faktorexacta.v12i3.4992.

[8] G. A. Buntoro, "ANALISIS SENTIMEN HATESPEECH PADA TWITTER DENGAN METODE NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE," vol. 5, no. September, p. 1939, 2016.

[9] A. Bayhaqy, "Analisa sentimen tentang islamophobia," no. August 2019, 2020.

[10] S. Prayoginingsih and R. P. Kusumawardani, "Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode Support Vector Machine (SVM," *J. Sisfo*, vol. 06, no. 03, pp. 347–382 Sistem, 2017.

[11] J. Pfeffer, K. Mayer, and F. Morstatter, "Tampering with Twitter's Sample API," *EPJ Data Sci.*, vol. 7, no. 1, 2018, doi: 10.1140/epjds/s13688-018-0178-0.

[12] S. Sagar, "Twitter Sentiment Analysis Using Vader," *IJARIIT (Volume 4)*, vol. 4, no. 1, pp. 485–489, 2018, [Online]. Available: https://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r/.

[13] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[14] S. Thakare, A. Kamble, V. Thengne, and U. R. Kamble, "Document Segmentation and Language Translation Using Tesseract-OCR," in *2018 13th International Conference on Industrial and Information Systems, ICIIS 2018 - Proceedings*, 2018, doi: 10.1109/ICIINFS.2018.8721372.