When Homecoming is not Coming: 2021 Homecoming Ban Sentiment Analysis on Twitter Data Using Support Vector Machine Algorithm

1st Lidia Sandra
Doctor of Computer Sience
Universitas Bina Nusantara
Jakarta, Indonesia
lidia.sandra@binus.ac.id

2nd Ford Lumbangaol
Associate Professor
Universitas Bina Nusantara
Universitas Bina Nusantara Jakarta, Indonesia
fgaol@binus.edu

2021 International Conference on ICT for Smart Society (ICISS) | 978-1-6654-1697-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICISS53185.2021.9533255

***Abstract*—Homecoming, more traditionally known as *Mudik*, has become a trending topic on several social media platforms as soon as the 11-day homecoming ritual ban was announced on 7 April 2021. Opinions, varying from those in favor of and against the ban, start to rapidly appear. Twitter, a social media platform which is now considered to be an extension of oneself and often used to express ones' opinion, has become flooded with comments on the homecoming ritual ban. The swarm of opinions in the form of tweets were then used as a dataset for sentiment analysis in order to understand how people perceive the ban. The algorithm used in this research is the classification algorithm using the Support Vector Machine method. The dataset was classified into three sentiments: positive, negative, and neutral. The use of the Support Vector Machine algorithm yielded a 62% accuracy with this dataset. The sentiment analysis showed that the keyword "mudik" had a neutral sentiment for the most part. Meanwhile, results of engagement analysis show that the largest forms of engagements were retweets and liking tweets that had a neutral sentiment. When the neutral sentiment was removed, we found that the largest sentiment on the homecoming ritual ban was negative. This is likely due to the release of an addendum to the Covid-19 Handling Task Force Circular Number 13 of 2021 on 22 April 2021 that imposes more restrictions on and extends the effective dates of the restrictions related to the homecoming ritual ban; exactly one day before the data scraping of 5000 datasets on tweets from 23 April 2021 was carried out. The researcher had already sampled the tweets with the most engagements (those with the most retweets and likes). It was found that some tweets had a negative sentiment, but the model classified it as having a neutral sentiment. This may be affected by inaccuracies of dataset training as some of the tweets were in Malay rather than Indonesian. A challenge that needs to be overcome is the limited number of datasets for NLP training or sentiment analysis for the Indonesian language in comparison to that of the English language. On the other hand, this has become an opportunity for the researcher to develop a more appropriate training model.**

Keywords—Twitter Sentiment Analysis, Homecoming, Support Vector Machine

I. Introduction

The Covid-19 pandemic has given no space for the world to rest, even for just a moment. People can no longer enjoy long weekends, or partake in homecoming rituals (*mudik*) to rejoin families separated due to education or work, to prevent a spike in Covid-19 cases in Indonesia.

For two consecutive years, 2020 and 2021, the homecoming ritual ban has been in place. The ban is an effort made by the government to control and reduce the number of Covid-19 cases in Indonesia. Issued in the Circular Letter No.13 Year 2021, a nationwide ban is imposed from 6-17 May 2021 for homecoming activities that involve travelling. An addendum to this circular number was issued on 22 April 2021 which extends the restrictions for homecoming ritual from 22 April to 22 May 2021 [1]. This decision made by the government has garnered varying responses, both positive and negative.

Comments on the issue were diverse, both on their connotations and where they were made. The issue was the talk of the nation being brought up in street markets, radio stations and national television. Further, it also flooded social media platforms like Twitter and TikTok.

Twitter developed by Twitter, Inc. offers a new media in which users can expand their social networks through uploading and receiving texts with a limit of 140 characters. These texts are better known as 'tweet's. Twitter has its own format and characteristic writing style as well as symbols and certain rules [2]. Supported by the massive development in broadband use of above 45 GB each month in several countries including large cities in Indonesia, people have grown to intensely use virtual rooms on social media to express themselves and convey their opinions [3]. Indonesia ranks third in numbers of internet users in Asia. A record of 44.6 million users of Facebook and 19.5 million users of Twitter; making the country the fifth largest country in terms of Twitter users [4].

The homecoming ritual ban that had been imposed had become a hot issue. Twitter, a social media that works in real-time, allows users to express their opinions on many issues or problems. Those opinions can be utilized as material for sentiment analysis to discover whether people's perception of the homecoming ritual ban is positive or negative. Results of the sentiment analysis may help in the evaluation of the ban in the midst of Covid-19.

Sentiment analysis and opinion mining are activities that process, understand, and extract textual data to obtain information on sentiment from opinions. Sentiment analysis or opinion mining is a computational study of opinions where the sentiments and emotions expressed within a text are extracted through entity and attribute extraction using text mining. Sentiment analysis will classify the polarity of text in a sentence or document to help understand whether the opinion within that text or document is either positive, negative or neutral [5]. Prior to tweet preprocessing and cleaning, data crawling was first done on appropriate keywords [6].

978-1-6654-1697-9/21/$31.00 ©2021 IEEE

The Support Vector Machine (SVM) algorithm will classify the opinions. According to previous research [7-9], it has been found that the SVM method yields relatively high accuracy when used for sentiment analysis. This research aims to implement the Support Vector Machine (SVM) algorithm to conduct sentiment analysis on the homecoming ritual ban as well as restrictions on domestic travel two weeks prior to and following the ban period (22 April to 24 May 2021) on Twitter that was issued in Circular Letter No.13 Year 2021 by the Indonesian Covid-19 Task Force.

II. Theory

A. *Text Mining*

Text mining is a process done to extract information from text by looking at the pattern of a sentence. Text mining identifies the pattern in structured data, evaluates it and interprets an output. This process covers the categorization of text, clustering, attribute or entity extraction, granular taxonomy production, sentiment analysis, document inference and entity-relationship modelling [10]

B. *Sentiment Analysis*

Sentiment Analysis aims to categorize the polarity of texts in documents, sentences or features that are contained within opinions in a document. Sentiment analysis is often used to analyze issues and determine opinions or comments on th issue have positive, negative, or neutral tendencies. Results o sentiment analysis can be used to then evaluate and thus bette certain policies or the performance of products or service [11].

C. *Support Vector Machine (SVM)*

The Support Vector Machine (SVM) method, made by Vladimir Vapnik, is a classification method that uses machin learning (supervised learning) to predict classes based on patterns from the training process. Opinions are classified into either positive or negative by having them distinguished from each other. This algorithm allows for the linear separation of non-linear high dimensional input data [12].

Based on previous research, the use of SVM for sentiment analysis allows for great accuracy. The SVM method used with Term Frequency- Inverse Document Frequency (TF-IDF) for text feature extraction was found to have an accuracy of 86%, making it the most accurate method when compared to other methods used in sentiment analysis carried out by *Gojek* [13]. Moreover, this is confirmed through the findings of other research that consistently shows that the SVM method yields the highest accuracy [14].

D. *Term Frequecy- Inverse Document Frequency (TF- IDF)*

TF-IDF is a metric of how relevant a word is within a dataset; it is a word2vec based process used for text feature extraction. It first represents texts as vectors and then assigns a certain score to features based on how frequently that feature is found in a part of the data. These scores are then divided by the frequency of that feature in the data overall, giving the TF-IDF. This technique increases the accuracy of the sentiment analysis model as all features are weighted differently, providing insight on the most dominant feature in the data used [15].

E. *Twitter API*

Twitter API is a service provided by Twitter for developers. This service can be used to develop applications that require data from Twitter.

III. Methodology

A. *Data Extraction*

The data extraction from tweets on Twitter was done using the crawling text method. Twitter was chosen due its relatively access to API as well as the large, and diverse, amount of information available. After the data had been obtained, preprocessing was done.

B. *Data Processing*

Preprocessing is the first step in data cleaning. It is done by eliminating noise in the data and resolving missing information so that the data can be processed optimally. Stages of text preprocessing carried out in this research involved the removal of special characters, stop words, and changing the format of all text to lowercase.

C. *Analysis Scheme*

Data extraction in this research was done using the keyword 'mudik' with research analysis scheme as follows:

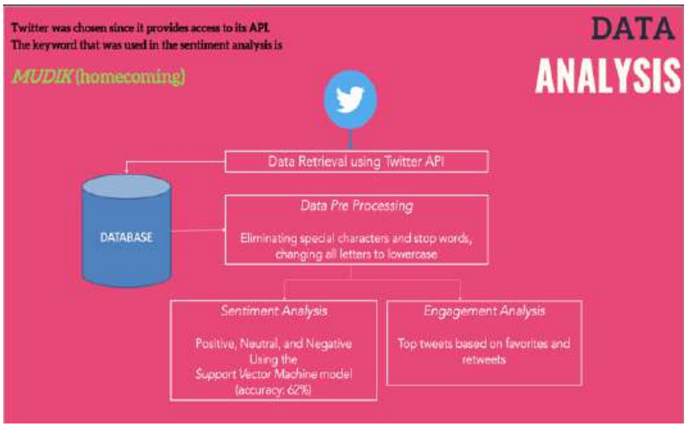

Twitter was chosen since it provides access to its API.
The keyword that was used in the sentiment analysis is
MUDIK (homecoming)
DATA
ANALYSIS
Data Retrieval using Twitter API
DATABASE
Data Pre Processing
Eliminating special characters and stop words, changing all letters to lowercase
Sentiment Analysis
Positive, Neutral, and Negative Using the Support Vector Machine model (accuracy: 62%)
Engagement Analysis
Top tweets based on favorites and retweets

Fig. 1. Data Analysis Scheme

D. *Evaluation Method*

The model used in this research was evaluated by obtaining an accuracy score for the Support Vector Machine algorithm used.

IV. Implementation

The training dataset used in this study uses research results from the UGM team in the following link: http://ridi.staff.ugm.ac.id/2019/03/06/indonesia-sentiment-analysis-dataset/. This dataset has been used to train models in Indonesian with 10,806 training data. The dataset includes a

mixture of Malay but because of the similarities, for example the word "no" in Indonesian and in Malay means the same, the dataset is considered quite valid. The limitation of the dataset in Indonesian is a challenge in itself in the sentiment analysis of Indonesian-language Twitter because of the unmet data needs and the very broad scope of sentiment analysis [16].

From the existing dataset, analysis of the bar plot is carried out as follows, and an image of the three dataset groups is obtained as follows.

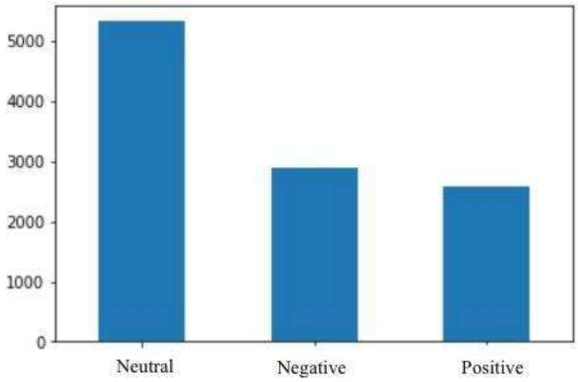
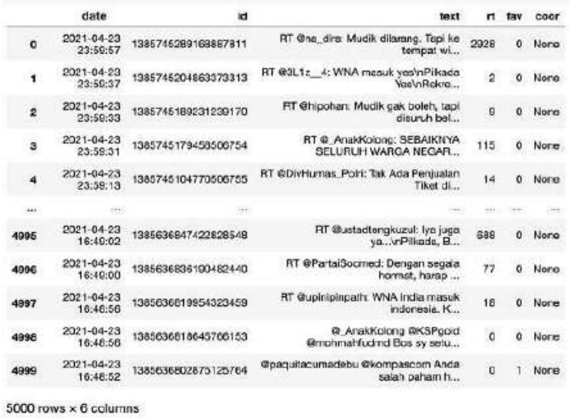


Fig. 2. Sentiment Classification Dataset Analysis



	date	id	text	rt	fav	coor
0	2021-04-23 23:59:57	1385746289168887811	RT @ne_dira: Mudik dilarang. Tapi ke tempat wi...	2928	0	None
1	2021-04-23 23:59:37	1385746204863373313	RT @3L1z__4: WNA masuk yes\nPilkada Yes\nRokre...	2	0	None
2	2021-04-23 23:59:33	1385745189231239170	RT @hipohan: Mudik gak boleh, tapi disuruh bel...	9	0	None
3	2021-04-23 23:59:31	1385745179458506754	RT @_AnakKolong: SEBAIKNYA SELURUH WARGA NEGAR...	115	0	None
4	2021-04-23 23:59:13	1385745104770506755	RT @DivHumas_Polri: Tak Ada Penjualan Tiket di...	14	0	None
...	...	...	...	...	...	...
4995	2021-04-23 16:49:02	1385636847422828548	RT @ustadtengkuzul: Iya juga ya...\nPilkada, B...	688	0	None
4996	2021-04-23 16:49:00	1385636836190482440	RT @PartaiSocmed: Dengan segala hormat, harap ...	77	0	None
4997	2021-04-23 16:46:56	1385636619954323459	RT @upinipinpath: WNA India masuk indonesia. K...	16	0	None
4998	2021-04-23 16:46:56	1385636618645766153	@_AnakKolong @KSPgoid @mohmahfudmd Bos sy selu...	0	0	None
4999	2021-04-23 16:46:52	1385636802875125764	@paquitacumadebu @kompascom Anda salah paham h...	0	1	None

5000 rows × 6 columns

Fig. 3. Scraping Data

The dataset for training has removed unnecessary characters, such as stop words and words that do not contain sentiments and converts words into numbers. The feature used is a count vectorizer, cut to 1500 words. The number of words that are 2000, for example, will be cut to 1500. Meanwhile, from 50 words, it will be changed to 1500 by filling in the remaining words (word difference) with the number 0.

The dataset is divided into two, namely training data and testing data, with a composition of 20% for data test and 80% for training data. By using random state 42. The next implementation step is to define the model. The kernel parameter chosen in this study was rbf as the optimal result of trial and error after trying various other kernels.

The next step is fitting the model using the training data. Data scrapping was then carried out by taking 5000 tweets with a duration from April 16-24 with 1500 features. It was found that 5000 tweets were netted from tweets only on 23 April 2021 (see Figure 3). The next step is to calculate the accuracy and precision results.

Data collection in this study was carried out by scraping the site www.search.twitter.com. The scraping process is carried out using the Python Tweepy library http://tweepy.readthedocs.io/en/v3.5.0/api.html#tweepy-api-twitter-api-wrapper and the Twitter API. Tweet scraping is done automatically to retrieve tweet data with the data period specified above. Before scraping is possible, the system checks the consumer key and access token so that the system can retrieve data from Twitter. The following is the consumer key and access token used in this study:

consumer_key = 'cEcWpJKUzM043X1RO2WFn6QJb'
consumer_secret =
'7OBpCTMblS1trVhXUH2vMG9oUtmuURnRIbIghKaL060E H8u5Ha'

with access_token =
'61717767-

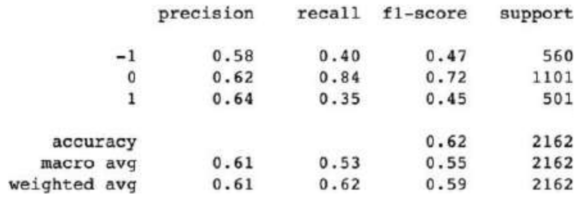
I4auSvDEIdVt3KIdeOPS0QaaHigKt6i4b7wBHKX3y'
access_token_secret =
'DnwdboAFqOWemztgEOHrOEnvlhR3SImn3AKSkhhURE1 KL'

The access authentication process above uses the tweepy API library.

The next step in the preprocessing process is done by removing the special characters, single characters, single characters from the start/initial special characters, replacing multiple spaces with single spaces and changing the uppercase to the lower case.

V. Result and Discussion

The Support Vector Machine algorithm can be implemented to see the sentiment of the 2021 homecoming



```
              precision    recall  f1-score   support

          -1       0.58      0.40      0.47       560
           0       0.62      0.84      0.72      1101
           1       0.64      0.35      0.45       501

    accuracy                           0.62      2162
   macro avg       0.61      0.53      0.55      2162
weighted avg       0.61      0.62      0.59      2162
```

Fig. 4. Level of Accuracy

The results of the sentiment classification can be divided into three, namely positive, negative and neutral. The biggest sentiment is neutral sentiment. If we eliminate neutral sentiment, the biggest sentiment is negative sentiment. From the results of the sentiment analysis of the keyword “mudik” data that has been obtained from the scrapping process, using the homecoming or so-called “mudik” keyword, the count taken is 5000 tweets from April 16-24, 2021

The results of the homecoming sentiment analysis on Twitter show that the most sentiments are neutral and the second most negative. If we look only at the negative and positive sentiments, the keyword ‘mudik’ since April 23, 2021 has reaped negative sentiment. This is likely influenced by the issuance of an addendum to Circular Letter No.13 of 2021 from the Covid-19 Task Force which tightens the regulation of the ban on going home from April 22, 2021 to May 24, 2021.

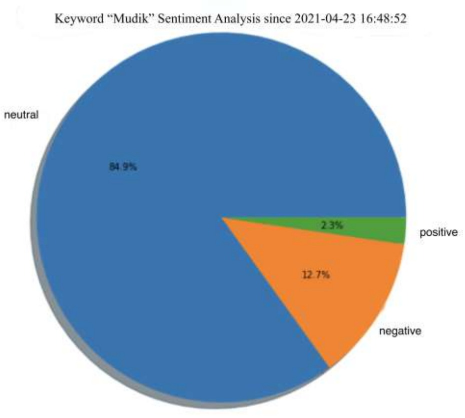


Fig. 5. Results of "mudik" Sentiment Analysis

Furthermore, the researcher conducted an engagement analysis based on the number of retweets and likes and obtained the following graph using the plot bar

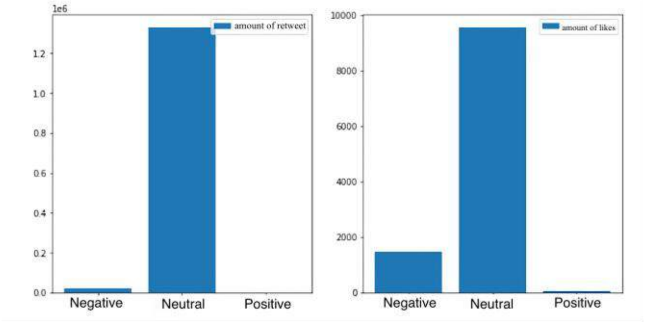
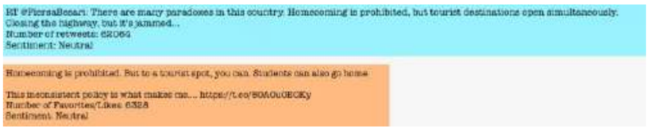


Fig. 6. Results of "mudik" Sentiment Analysis

It can be seen that the number of retweets and favorite tweets (which have the most likes) is neutral sentiment. This is of course related to the results of the sentiment analysis above which shows that tweets have a neutral sentiment that dominates.

But by taking a random sample of tweets that have high engagement below and have a neutral sentiment, when it is being viewed from the whole sentence, it can be seen subjectively that the sentiment below is more appropriate in the negative sentiment category than neutral.



RT @FiersaBesari: There are many paradoxes in this country. Homecoming is prohibited, but tourist destinations open simultaneously. Closing the highway, but it's jammed...
Number of retweets: 62064
Sentiment: Neutral

Homecoming is prohibited. But to a tourist spot, you can. Students can also go home

This inconsistent policy is what makes me.... https://t.co/SOAOuOBCKy
Number of Favorites/Likes: 6328
Sentiment: Neutral

Fig. 7. Example of Neutralized Tweets with the Biggest Engagement.

The possibility of the above happened because the accuracy rate of the model is 62% which still allows a large chance of misclassification, especially training data which also uses several Malay languages.

What is also very interesting to observe from tweets with the highest engagement is that tweets from influencers, for example in Twitter circles, the accounts used for data scraping are from FiersaBesari (Figure 7), have a very large opportunity to form other opinions and transmit opinions with many retweets and likes. This is very interesting for further research. Contagious retweets appear to be from "influencers" giving an indication that "tweets" and opinions are contagious. The tendency for conformity to friends in the Twitter, and the occurrence of polarization in opinion according to the circle is very high. It is interesting to do further research to see the circle of influence and how big the chance is of a tweet or opinion spreading in this infinite virtual space.

VI. Conclusion

The results of the implementation of the Support Vector Machine algorithm to analyze the sentiment of the 2021 homecoming ban due to the Covid-19 outbreak dividing sentiment into three major classifications, namely positive, negative and neutral sentiments. The results show that the majority of the keyword "mudik" has a neutral sentiment followed by the greatest engagement. By eliminating neutral sentiment, the biggest sentiment is negative sentiment. Seeing the data scraping obtained is from April 23, 2021, which is exactly one day after the announcement of the extension of the homecoming ban, this is thought to have contributed to the negative sentiment regarding the homecoming ban.

With an accuracy rate of 62% and after sampling the tweets with the most engagement, retweets and likes, some sentiments can be categorized as negative but in the model are categorized as neutral sentiments. This is possible because the training data used contained several Malay languages.

The results of the accuracy of the use of the Support Vector Machine method for sentiment analysis are influenced by several things, including in this study the training dataset is less precise, the number of datasets used and the composition of the number of positive and negative data. The obstacle that needs to be overcome is the limited dataset for NLP training / sentiment analysis in Indonesian which is still minimal compared to English. This is also an opportunity for researchers to continue to develop more appropriate training models.

References

[1] Satuan Tugas Penanganan Covid-19, "Surat Edaran Peniadaan Mudik Hari Raya Idul Fitri Tahun 1442 Hijriah dan Upaya Pengendalian Penyebaran Corona Virus Disease 2019 Selama Bulan Suci Ramadhan 1442 Hijriah," Surat Edaran no. 13, 2020.

[2] 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, 2019, p. 0-12.

[3] A. Novantirani et al., "Analisis Sentimen pada Twitter Mengenai Penggunaan Transportasi Darat Dalam Kota Dengan Menggunakan Metode SVM" E Proceeding of Engineering, vol. 2, no 1, 2015, p. 1177.

[4] GATRAnews, "Indonesia Peringkat Lima Pengguna Twitter," 2012. http://www.gatra.com/iltek/internet/20244-indonesia- peringkat-lima-pengguna-twitter.html, accessed Apr. 25, 2021.

[5] B. Liu, "Sentiment Analysis and Subjectivity," NLP Handbook, no. 1, 2010, p 138.

[6] C. Gu and A. Kurov, "Informational role of social media: Evidence from Twitter sentiment," J. Bank. Financ., vol. 121, 2020, p. 105969.

[7] Imamah et. al "Text Mining and Support Vector Machine for Sentiment Analysis of Tourist Review in Bangkalan Regency" Jurnal of Physics, vol. 1447, 2020.

[8] W.A. Luqyana, I. Cholissodin and R.S. Perdana "Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Universitas Brawijaya, vol 2, no 11, 2018, p. 4704-4713.

[9] A.M. Pravina, I. Cholissodin, P.P Adikara "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter menggunakan

Algoritma SVM," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. vol 3, no 3, 2019, p 2789-2797.

[10] C. Bridge, "Unstructured Data and the 80 Percent Rule," https://breakthroughanalysis.com/2008/08/01/unstruct ured-data-and-the-80-percent-rule/, accessed Apr 24, 2021.

[11] T. Adilah,, Y. Alkhali, N.A. Mayangky and W. Gata, "Analisis Sentimen Opini Publik Mengenai Larangan Mudik pada Twitter Menggunakan Naiva Bayes", Jurnal CoreIT, vol. 6, no 2, 2020, p.85-88.

[12] M.A. Maulana, A. Setyanto and M.P. Kurniawan"Analisis Sentimen Media Sosial Universitas Amikom", Seminar Nasional Teknologi Informasi dan Multimedia 10 February 2018, Amikom Yogyakarta, 2018, p 7-12.

[13] I.P Windasari et al "Sentiment Analysis on Twitter Posts: An Analysis of Positifve or Negative Opinion on Gojek," Proceeedings 2017 4th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2017, January 2018, p. 366-369.

[14] I. Santoso, W. Gata and A.B Paryanti, "Penggeunaan Feature Selection di Algoritma Support Vector Machine," Jurnal RESTI., vol. 1, no 10, 2019, p. 5-11.

[15] V.I.Santoso, G. Virginia, Y. Lukito, "Penerapan Sentiment Analysis pada Hasil Evaluasi Dosen Dengan Metode Support Vector Machine," Jurnal Transformatika., vol. 14, no 2, 2017, p. 72-76.

[16] R. Ferdiana et.al. "Dataset Indonesia untuk Analisis Sentimen", JNTETI., vol. 8, no 4, 2019, p. 334-339.