ATCSJ
Applied Technology and Computing Science Journal

Open Access

p-ISSN 2621-4458 | e-ISSN 2621-4474

Available online to journal2.unusa.ac.id

UNUSA

S4-Accredited – SK No.200/M/KPT/2021

Journal Page is available to https://journal2.unusa.ac.id/index.php/ATCSJ

Sentiment Analysis Pedulilindungi Tweet Using Support Vector Machine Method

Farhan Setiyo Darusman[1]*, Amalia Anjani Arifiyanti[2], Seftin Fitri Ana Wati[3]

[1,2,3] *Sistem Informasi,Fakultas Ilmu Komputer, UPN Veteran Jawa Timur, Indonesia*
Rungkut Madya No.1, Gunung Anyar, Surabaya, Indonesia
[1*]18082010044@student.upnjatim.ac.id, [2]amalia_anjani.fik@upnjatim.ac.id, [3]seftin.fitri.si@upnjatim.ac.id

Article history:

Received 9 April 2022
Revised 29 April 2022
Accepted 20 May 2022
Available online 30 May 2022

Keywords:

Classification
Pedulilindungi
Sentiment Analysis
SVM
Tweet

Abstract

Pedulilindungi application has many benefits but many controversies arise in the community. Various opinions in the form of tweets were expressed by the public, both positive and negative opinions. In this study, the objective is to make a classification model to classify tweets into two types of sentiment, namely positive and negative. The model is made in several stages, namely data retrieval, data filtering, data labeling, data preprocessing, splitting data train and data test, feature selection using Information Gain and Genetic Algorithm, and then classification using the SVM method. The model using two-stage feature selection and SVM method, obtained an accuracy value of 64.08% with 841 features and processing time of 0.033 seconds with 9.6% CPU usage. The model with two-stage feature selection is more efficient and effective than the one-stage feature selection model whose accuracy value is only 60.56% with 1800 features and a processing time of 0.044 seconds with 15.4% CPU usage.

I. INTRODUCTION

Since the first case on March 2, 2020, the COVID-19 pandemic has hit Indonesia [1]. After more than a year, variants of the COVID-19 virus emerged. Quoted from the Ministry of Health website, in May 2021, the Ministry of Health confirmed that three variants of COVID-19, namely the alpha, beta, and delta variants, had entered Indonesia[2]. The government will start re-implementing restrictions on community activities called the PPKM Darurat (Enforcement of Restrictions on Community Activities) in July 2021. This also coincides with the start of a vaccination program for the general public.

On August 23, 2021, the government will begin to allow activities outside the home on condition that the Pedulilindungi application is used as a screening and tracing tool. The results of the screening and tracing of Pedulilindungi will be used as a consideration for determining the level of PPKM in each region. Quoted from the official Pedulilindungi website, Pedulilindungi is an application developed by KOMINFO (Ministry of Communication and Information Technology) in collaboration with the Ministry of Health (Ministry of Health), KPCPEN (Covid-19 Handling Committee and National Economic Recovery), and the

[1*]Corresponding author

ISSN 2621-4474 (online) 2621-4458 (print) © 20XX The Authors. Published by UNUSA.

DOI: https://doi.org/10.33086/atcsj.v4i2.2836.

Ministry of BUMN where this application is an application for tracing Covid-19 which is applied in Indonesia to do contact tracing digitally [3].

Although the Pedulilindung application has many benefits both for the government and the community, many controversies arise in the community. Twitter is one of the platforms used by the public to express their opinion about Pedulilindungi. Twitter is an online social media and microblogging application where users can send and read messages called tweets[4]. Tweets are text-based messages on social media Twitter. In March 2006, Jack Dorsey founded Twitter along with several other social media in July. Since its launch, Twitter has been included in the ranks of the ten most visited sites by users on the Internet, and has earned the nickname "short messages from the Internet". On Twitter, users who have not registered can only view and read tweets, while registered users can write tweets via Twitter's website or application. Various opinions in the form of tweets are expressed by the public, both positive and negative opinions. These opinions discuss matters related to Pedulilindungi such as applications and regulations. The sentiments in this opinion[5] can be useful for the development of Pedulilindungi applications and regulations by related parties, so it is necessary to carry out sentiment analysis. In this study, the tweet will be classified into two classes of sentiment, namely positive and negative.

The tweets that will be analyzed are tweets in September 2021 to October 2021. Because in that timeframe the topic of Pedulilindungi is being discussed hotly, it is often included in the trending topic column on Twitter. This may be because at the beginning of September the government began to require the use of Pedulilindungi so people began to use and discuss the application.

To perform sentiment analysis, a machine learning approach is used using supervised learning with the Support Vector Machine (SVM) method which is included in the linear classifier category. At the preprocessing stage, the two-stage feature selection method will also be used. SVM is a method to separate two classes in the input space by identifying the best hyperplane. SVM is classified as linear classification. SVM applies the basic principles of linear classification, which was later developed by adding the concept of a trick kernel on a high-dimensional space into a non-linear classifier. In general, real problems have a non-linearly separable form, so the hyperplane cannot classify the two classes perfectly. Therefore, the kernel principle is applied to SVM. The concept of non-linear SVM is to apply the (x) function to map changing x data to vector space with higher dimensions[6]. To increase the effectiveness and efficiency of the model, a two-stage feature selection will be carried out with Information Gain (IG) in the first stage and Genetic Algorithm (GA) as proven by Uguz[7]. Although this method has increased the effectiveness and efficiency of the model, but the data used to make the model is dummy data. In this study, this method will be tested using real data which are tweets about Pedulilindungi. Model obtained in this study can later be used to classify tweets, especially about Pedulilindungi.

II. METHODS

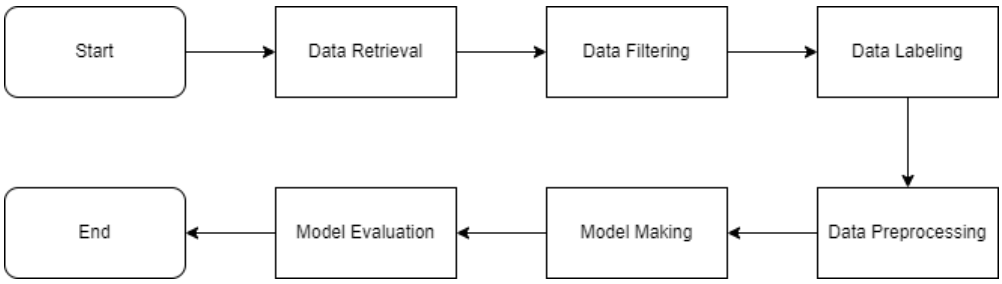

Figure 1. Methodology

A. Data retrieval[8]

The first stage in classifying is collecting data to be classified. To retrieve tweet data using a scraping technique[9]. In doing scraping, the snscrape library in the python programming language is used where this library does not require an API from Twitter and has no limit on the amount of data obtained.

B. Data Filtering

Then the data obtained will be filtered[10]. This is because not all of the data obtained are sentiments[11] about Pedulilindungi. The tweets obtained may only contain the keyword Pedulilindungi but not discuss Pedulilindungi.

C. Data Labeling

The tweet data will then be manually labeled with its sentiment class[12]. This is necessary because classification is included in supervised learning, meaning that a label/class is needed to be used in training the classification model created.

D. Data Preprocessing

After the data is label/class, then the data will go through the preprocessing stage[13]. In this stage, the data will be standardized and simplified. This is done to improve model performance and overcome noise

E. Model Making

The data that has been processed is then divided into train data and test data. Then the train data features will be selected in two stages using Information Gain and Genetic Algorithm. With the selected features, classification will be carried out using the support vector machine method. Classification [14] is carried out using two data, the first data is data selected for features using Information Gain only, and the second data is data selected for features using Information Gain and Genetic Algorithm. This is done to see a performance comparison between the two models made.

F. Model Evaluation

After the model has been created, it will be evaluated. This evaluation is carried out to measure the performance of the model created by calculating the data correctly or incorrectly predicted by the model. The comparison between the number of correctly predicted data and the amount of test data is called accuracy. Meanwhile, the ratio of the number of incorrectly predicted data to the amount of test data is called the error rate. Evaluation of the model uses the confusion matrix method. Confusion Matrix is a tool for measuring classifier performance in identifying tuples from different classes. The following is a description of the confusion matrix in Table 1.

Table 1. Mean and Standard Deviation of Question from Usability testing

		Predicted Class	
		False	Positive
Actual Class	False	TN	FN
	Positive	FN	TP

III. RESULTS AND DISCUSSIONS

A. Data retrieval

The data retrieval stage utilizes the snscrape library by applying filters according to the research limits, namely tweets with the keywords Pedulilindungi, in Indonesian and the date of tweet creation in the period September 1, 2021 to October 31, 2021. The results obtained are tweet data of 33075 lines of data in Comma Separated Value (CSV) file extension. The file contains several columns, namely the name of the user who made the tweet, the date and time of the tweet, and the text of the tweet.

B. Data Filtering

Not all data obtained at the data collection stage is sentiment[15]. Therefore, it is necessary to do a filter to select data that is truly a sentiment with to increase the effectiveness of the model to be made. After the data was filtered, 2131 data were obtained which were truly sentiments

C. Data Labeling

Classification is a method that is included in supervised learning so that a label/target/class is needed on the data to be modeled. At this stage, each data line is labeled according to its sentiment class manually by one person. Sentiment class consists of positive and negative classes.

D. Data Preprocessing

At this stage the data will be uniformly shaped through several stages, namely the translation of emoji and emoticons, case folding, cleansing, changing slang to their standard forms, changing non-standard words to their standard forms, foreign language translation, stopword removal, and stemming.

E. Model Making

a) Splitting Data Train and Data Test

After the data has gone through the preprocessing stage, the next step is to divide the data into training data and test data. The training data will be used to create the model and the test data will be used to test or evaluate the model. Data Splitting is done using the KFold method.

b) Two-Stage Features Selection

The selected training data will then select what features will be used in model training. The data that was previously in the form of a string is converted to vector form by using the TfidfVectorizer function so that it can be processed by the classifier. In the first stage with Information Gain, 1800 features were selected and in the second stage with Genetic Algorithm 841 features were selected.

c) Classification Using Support Vector Machine

Classification[16] is done by running the fit function where this function is used to train the classifier to classify. In addition to classification with two-stage feature selection using 841 features, a classification with single-stage feature selection with 1800 features will also be carried out to compare the performance results of the two. Both classifications use the same kernel, namely the linear kernel.

F. Model Evaluation

```
Penggunaan CPU:  15.4
Waktu Proses : --- 0.04488205909729004 seconds ---
Nilai Akurasi :
0.6056338028169014

Confussion Matric :
[[195  59]
 [109  63]]

Classification Report :
              precision    recall  f1-score   support

     negatif       0.64      0.77      0.70       254
     positif       0.52      0.37      0.43       172

    accuracy                           0.61       426
   macro avg       0.58      0.57      0.56       426
weighted avg       0.59      0.61      0.59       426
```

Figure 2. One-Stage Feature Selection Model Evaluation

```
Penggunaan CPU:  9.6
Waktu Proses : --- 0.033905982971191406 seconds ---
Nilai Akurasi :
0.6408450704225352

Confussion Matric :
[[176  78]
 [ 75  97]]

Classification Report :
              precision    recall  f1-score   support

     negatif       0.70      0.69      0.70       254
     positif       0.55      0.56      0.56       172

    accuracy                           0.64       426
   macro avg       0.63      0.63      0.63       426
weighted avg       0.64      0.64      0.64       426
```

Figure 3. Two-Stage Feature Selection Model Evaluation

Based on the result experiment, that shown in Figure 2 and Figure 3 during stage by using model evaluation with a single-stage feature selection, an accuracy of 60.56% was obtained using 1800 features and a processing time of 0.044 seconds with a CPU usage of 15.4%. Meanwhile, with two-stage feature selection, accuracy increased by 3.52% to 64.08% using 841 features and processing time of 0.033 seconds with 9.6% CPU usage, so it can be concluded that with the two-stage feature selection, the effectiveness and efficiency increase. Although the increase in accuracy is not very high, the efficiency increase is quite significant because the dimensionality of the dataset is reduced by more than half so that the complexity of the model is reduced and has an impact on increasing CPU resource efficiency and processing time.

IV. CONCLUSIONS AND RECOMMENDATIONS

Based on the discussion of the previous chapters, it can be concluded that the model is made in several stages, namely data retrieval, data filtering, data labeling, data preprocessing, splitting data train and data test, feature selection, and classification. The model made using the two-stage feature selection method and the Support Vector Machine got an accuracy value of 64.08% with 841 features and a processing time of 0.033 seconds with 9.6% CPU usage. The model with two-stage feature selection is more efficient and effective than the one-stage feature selection model whose accuracy value is only 60.56% with 1800 features and a processing time of 0.044 seconds with 15.4% CPU usage.

V. REFERENCES

[1] A. I. Almuttaqi, "Kekacauan respons terhadap Covid-19 di Indonesia," *Thc Insigjts*, vol. 13, 2020.
[2] "Waspada 3 Varian Baru Covid-19 di Indonesia," May 11, 2021. https://promkes.kemkes.go.id/waspada-3-varian-baru-covid-19-di-indonesia (accessed Jun. 01, 2022).
[3] "PeduliLindungi," Dec. 23, 2021. https://id.wikipedia.org/wiki/PeduliLindungi (accessed Jun. 01, 2022).
[4] "Twitter," May 16, 2022. https://id.wikipedia.org/wiki/Twitter (accessed Jun. 01, 2022).
[5] L. Asri *et al.*, "Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm," *ejournal.nusamandiri.ac.id*, vol. 13, no. 1, 2017, Accessed: Jun. 01, 2022. [Online]. Available: http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/153
[6] M. A. Fauzi, "Analisis Sentimen Review Barang Berbahasa Indonesia Dengan Metode Support Vector Machine Dan Query Expansion Automatic Essay Scoring View project Twitter Sentiment Analysis View project," 2018.

[7] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[8] D. J. Hand, "Principles of data mining," in *Drug Safety*, 2007, vol. 30, no. 7. doi: 10.2165/00002018-200730070-00010.

[9] Gifa Delyani, "Kenali Web Scraping, Salah Satu Teknik Pengumpulan Data Sekunder!," Mar. 03, 2021. https://www.dqlab.id/kenali-web-scraping-salah-satu-teknik-pengumpulan-data-sekunder (accessed Jun. 01, 2022).

[10] "Text Classification in Python. Learn to build a text classification… | by Miguel Fernández Zafra | Towards Data Science." https://towardsdatascience.com/text-classification-in-python-dd95d264c802 (accessed Jun. 01, 2022).

[11] A. Fauzia, F. H.-S. N. Hukum, and undefined 2021, "Pendekatan Socio-Cultural dalam Pelaksanaan Vaksinasi Covid-19 di Indonesia: Socio-Cultural Approach in the Implementation of Covid-19 Vaccination in," *proceeding.unnes.ac.id*, vol. 7, no. 1, p. 2021, doi: 10.15294/snhunnes.v7i1.709.

[12] H. Mosioi, E. M.-P. SISFOTEK, and undefined 2021, "Analisa Sentimen Publik Terkait Otonomi Khusus (OTSUS) di Papua dengan Pendekatan Sains Data," *seminar.iaii.or.id*, Accessed: Jun. 01, 2022. [Online]. Available: http://seminar.iaii.or.id/index.php/SISFOTEK/article/view/275

[13] E. Wahyuni, … A. A.-N. S. and, and undefined 2021, "Feature Extraction for Sentiment Analysis in Indonesian Twitter," *nstproceeding.com*, vol. 2020, doi: 10.11594/nstp.2021.0913.

[14] E. Dharmawan, … E. W.-J. I. dan, and undefined 2020, "Klasifikasi Opini Pengguna Smartphone Pada Twitter Di Indonesia," *jifosi.upnjatim.ac.id*, vol. 1, no. 1, Accessed: Jun. 01, 2022. [Online]. Available: http://jifosi.upnjatim.ac.id/index.php/jifosi/article/view/32

[15] A. Kaur and N. Gumber, "Sentimental Analysis on Application Reviews on Educational Apps," *academia.edu*, Accessed: Jun. 01, 2022. [Online]. Available: https://www.academia.edu/download/35954453/IJETR022706.pdf

[16] A. Khan, B. B.-2011 N. P. Conference, and undefined 2011, "Sentiment classification using sentence-level semantic orientation of opinion terms from blogs," *ieeexplore.ieee.org*, Accessed: Jun. 01, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6136319/