Journal of Industrial Engineering and Halal Industries (JIEHIS)
P-ISSN 2722-8150, E-ISSN 2722-8142
Vol. 5 No. 1 June 2024, Pp 36-42
https://doi.org/10.14421/jiehis.4672

Submitted: 2024-06-23
Revised: 2024-08-06
Accepted: 2024-08-08
Online: 2024-08-15

Sentiment Analysis of PeduliLindungi User Using Naïve Bayes Classifier Algorithm and Support Vector Machine

Rizki Rahmatullah[1*], Jundi Nourfateha Elquthb[1], Fanya Nindha Al-Qur'ani[1], Annisa Uswatun Khasanah[1]

[1]Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia
*Corresponding email: 20522091@students.uii.ac.id

Abstract

The Indonesian government is attempting to track the spread of the virus by creating an application named "PeduliLindungi" to deal with the coronavirus's exponential increase in cases across the country. Because it has a feature to disclose the user's location immediately, it is envisaged that this program can reduce the transmission of viruses in monitoring. Indonesians have used the PeduliLindungi, and there are user reviews of both positive and negative experiences. Therefore, to enhance these services, an assessment is required. The text mining method can extract information from users' reviews to collect this data. This method's application additionally uses the Naive Bayes Classifier and Support Vector Machine algorithms, which analyze word associations and do a classification evaluation of the data's accuracy. Based on the two methods' calculations, the NBC algorithm's average classification accuracy was 83.81%, and the SVM algorithm was 93.84%. Following that, discoveries on words that frequently exist or are used by people are obtained through word associations in the sentiment analysis of positive or negative reviews.

Keywords : Classification, Naïve Bayes Classifier, PeduliLindungi, Sentiment Analysis, Support Vector Machine, Word Association.

INTRODUCTION

Corona Disease 2019 (COVID-2019) is a new type of disease caused by the Sars-CoV-2 virus and is zoonotic that can be transmitted by humans and animals. This disease is very easily transmitted from human to human through sneezing and coughing. The negative impact of the emergence of the COVID-19 disease has hit all countries. Indonesia is no exception, which is also affected and included as the country with the highest corona cases. In response to the impact of the increasingly widespread cases of COVID-19, the government designed a program intending to minimize the spread of this virus. Some of these movements and policies are maintaining physical fitness, washing hands, maintaining distance (minimum 1 meter), wearing a mask, minimizing going out of the house, and postponing domestic and foreign trips.

The Indonesian government has designed an application called PeduliLindungi which is a service application to help track the spread of the virus and stop the spread of the COVID-19 virus. This application relies on the user to share location data while traveling to detect whether there is contact with a person with COVID-19. The presence of this application is very effective and efficient for monitoring the user community in carrying out activities and traveling long distances for a long duration. Over time, this application certainly gets reviews regarding its performance. It is known that the PeduliLindungi application, with 803 thousand reviews, received a rating of 4.4 on the Google Play Store and received a rating of 2.6 out of 5 on the AppStore. From the many reviews given by users of the PeduliLindungi application, it is known to be very diverse, ranging from positive to negative reviews. Most users complain that the PeduliLindungi application service is still challenging to understand, frequent login failures occur, and a lack of response to user complaints.

Rating and comments that contain positive and negative reviews from user can be used to evaluate and improve the application. Text mining is a popular method to extract review effectively. In this method, sentiment analysis is commonly used to extract data from a textual document and process it to determine the sentiment. Usually, the data analyzed is from customer opinions about a product or service (Fitri, et al., 2019). In this study,

the classification is applied to conduct the sentiment analysis. This method is included in supervised learning, and for the last ten years, it has progressed in machine learning. Several algorithm methods develop in this machine learning, such as Bayesian Classifier, Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Neural Networks (Ranjan, 2015).

The Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) algorithms are useful to classify data that has already been gathered about user information. The Naïve Bayes Classifier algorithm is a statistical classification algorithm used to estimate probabilities into subgroups and measure high-accuracy performance in a large database (Saleh, 2015). While the Support Vector Machine algorithm was chosen because it can identify hyperplanes that can separate two sets of data from two different classes (Vapnik V. N., 1999) and perform distance determination computation using the Support Vector Process more quickly (Vapni, et al., 1995). These two algorithms are usually used for categorizing text by extracting from certain sources to enrich knowledge (Hassan, et al., 2011)

The selection of the two algorithms is based on comparisons with other algorithms such as Random Forest, KNN, and Decision Tree. Whereas the Random Forest has a complex decision tree and a long prediction process because it requires a lot of decision trees when processing the input as a whole (Abidin, et al., 2020). The K-Nearest Neighbor (KNN) algorithm needs to improve; namely, when classifying data, it is necessary to calculate the distance of k-neighbors, which is quite intensive. If there are irrelevant features, it can decrease the accuracy level (Ray, 2019). The Decision Tree algorithm has disadvantages such as a long computational processing time and weaknesses in unbalanced data in a complex decision tree structure (Hakim, et al., 2017). Then in this research we will compare the accuracy performance values of which algorithm is better in this problem.

Text Mining

Text mining is a process of mining data in the form of text with data sources, usually obtained from documents that aim to find words that can represent the document's contents so that connectivity analysis can be carried out to acquire new useful information (Senellart, et al., 2008). The purpose of text mining itself is to extract useful information from data sources. The data source used in text mining is a collection of documents that have an unstructured format through the identification and exploration of exciting patterns. Specific text mining tasks include text categorization and text clustering (Gupta, et al., 2009).

Sentiment Analysis

Sentiment analysis or extracting opinions is one of the main tasks in Natural Language Processing (NLP) which studies a person's idea of a particular entity. Sentiment analysis itself comes from the perspective of product users (Fang, et al., 2015). Apart from being based on one's opinion, sentiment analysis can be used on a person's behavior or emotions, which can be used as evaluation material (Isnain, et al., 2021).

Sentiment analysis is also helpful in classifying products and services, targeting consumers in the market, and monitoring social media to improve the branding of a product brand. One example is marketing research, where a real-time consumer sentiment analysis will be carried out regarding customer attitudes and opinions toward a market (Ciocodeica, et al., 2022). This method can aim to determine the advantages and disadvantages of the products based on data obtained via the Internet. This is intended as user feedback to be used as evaluation material for product development.

Classification

Classification is a multivariate technique for separating different data from an object (observation) and allocating the new thing (observation) into a predefined group. A suitable classification method will produce negligible misclassification or a slight chance of error [16]. In this research, two frequently used algorithms are used, namely SVM and NBC. Each method or algorithm used has its own performance and advantages. Where Support Vector Machine is a method that can work very well on high-dimensional datasets and for the Naive Bayes Classifier itself it also has advantages in processing discrete and quantitative data efficiently (Kesumawati, et al., 2018).

METHOD

Data Collection

Using the Web Scrapping method, the data was collected through user reviews of the PeduliLindungi application service on the Google Play website database. In compiling the data, researchers used the AppFollow website on Google Chrome. In this study, primary data was collected from PeduliLindungi website during March 30 until May 14, 2022, in 4.3 until 4.4.3.1 version. The collected data contains the variable of the date of making reviews and the contents of user reviews related to the use of the application, there were 125,523 data of customer reviews that will be carried out in sentiment analysis.

Data Processing

Pre-processing is a stage that is carried out after data mining has been carried out. The goal is to perform data cleaning to improve the data performance. The data processing steps are as follows;

1. Data Cleaning
   The data cleaning stage is the process of removing duplicate data or words that have copies of the same sentence in the data that has been obtained.
2. Case Folding
   This stage itself is a normalization stage that is carried out to reduce vocabulary and allow for better generalization.
3. Tokenizing
   Tokenizing is separating words from a sentence with no word relationship. The separated words are known as tokens. The tokens that have been obtained will help in understanding the context or developing an NLP model.
4. Stemming
   In this stage, words that contain inductive and deductive affixes will be changed into essential words. According to Kowsari et al. (2019), Stemming is a way to modify words through different linguistic processes, such as adding affixes.

Review Analysis

In this study, several methods of analysis were implemented as follows:

1. The descriptive analysis provides an overview of the reviews in the PeduliLindungi application.
2. Sentiment Analysis analyzes and categorizes user reviews into positive or negative labels.
3. Machine Learning is used to classify predetermined labels using the Naïve Bayes Classifier and Support Vector Machine algorithms.
4. The word cloud is used to visualize reviews or words that users often use in providing reviews of the PeduliLindungi application. Text mining is a process of mining data in the form of text with data sources, usually obtained from documents that aim to find words that can represent the document's contents so that connectivity analysis can be carried out to acquire new useful information [11]. The purpose of text mining itself is to extract useful information from data sources. The data source used in text mining is a collection of documents that have an unstructured format through the identification and exploration of exciting patterns. Specific text mining tasks include text categorization and text clustering [12].

RESULT AND DISCUSSION

Pre-Processing Data

1. Data Cleaning

Five thousand seven review data from scrapping results will be used for sentiment analysis. Then, after cleaning process as many as 3683 reviews are remaining. In this study, sample data were processed using Python to extract reviews. In addition, RStudio software is also used to visualize words in reviews using the word cloud.

Table 1. Case Folding Process

Change the Word to Lowercase	
Input	**Output**
Udah tidak ada lagi covid ini, hapus aja dari PlayStore ini aplikasi	*udah tidak ada lagi covid ini hapus aja dari playstore ini aplikasi*
Deleting Numbers	
Input	**Output**
mengubah tahun dan bulan aja masih harus geser jauh ditambah ngeblank lagi. Bayangin kalau lahir tahun 60-70an berapa kali harus tekan2?	*Mengubah tahun dan bulan aja masih harus geser jauh ditambah ngeblank lagi. Bayangin kalau lahir tahun an berapa kali harus tekan?*
Removing Punctuation	
Input	**Output**
Tidak bisa berkomentar. SERBA LEMOT, mempersulit, memperhambat. R I B E T - L A M B A T	*Tidak bisa berkomentar SERBA LEMOT mempersulit memperhambat R I B E T L A M B A T*

2. Data Folding

Several processes are conducted at this stage, such as changing words to lowercase, deleting numbers, and removing punctuation.

3. Tokenizing

The tokenization process helps describe the meaning of words by analyzing word order and aims to make it easier to calculate the frequency of occurrence in documents.

Table 2. Tokenizing Process

Input	Output
update aplikasi untuk memperbaiki bug saat scan qr yang gagal	[*'update', 'aplikasi', 'baik', 'bug', 'scan', 'qr', 'gagal'*]

4. Stemming

Stemming is a filtering process performed on data.

Table 3. Stemming Process

Input	Output
"Update"	"Update"
"Aplikasi"	*"Aplikasi"*
"Memperbaiki"	*"Baik"*
"Dibuka"	*"Buka"*
"Penyimpanan"	*"Simpan"*

Sentiment Analysis

The sentiment analysis process automatically extracts and processes data using an algorithm to get a weighting score to continue labeling related to user reviews.

1. Sentiment Score Calculation

Sentiment score calculations Obtain to define positive and negative review.

Table 4. Sentiment Score

No	Review	Score
1	"[*'terima', 'kasih', 'aplikasi'*]"	1
2	"[*'update', 'aplikasi', 'baik', 'bug', 'scan', 'qr', 'gagal'*]"	-2
3	"[*'aplikasi', 'bagus'*]"	0

2. Sentiment Score Labeling

Sentiment score labeling is done based on three different class categories: the positive sentiment category class, the neutral sentiment category class, and the negative sentiment category class. Labeling criteria can be seen based on the conditions.

Positive sentiment: score > 0

Neutral sentiment: score = 0

Negative sentiment: score < 0

Table 5. Sentiment Labeling

No	Review	Score	Sentiment
1	"[*'terima', 'kasih', 'aplikasi'*]"	1	Positive
2	"[*'update', 'aplikasi', 'baik', 'bug', 'scan', 'qr', 'gagal'*]"	-2	Negative
3	"[*'aplikasi', 'bagus'*]"	0	Neutral

After pre-processing data and sentiment analysis on user review data for the PeduliLindungi application, it was found data was identified the label and ready to be used as shown below.

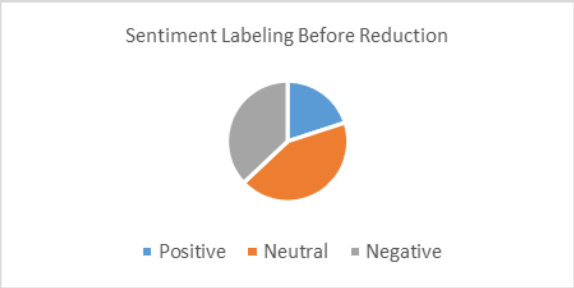

Figure 1. Sentiment Labeling Before Reduction

After doing sentiment analysis, it is known that this study only used two classes of positive and negative sentiments. Based on this, the data obtained is used to proceed to the following process: as many as 2,105 review data.

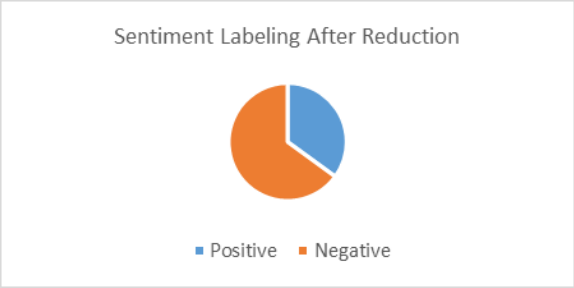

Figure 2. Sentiment Labeling After Reduction

Classification

Based on the sentiment data that has been obtained, the classification data will be processed three times for each algorithm, both Naïve Bayes Classifier and Support Vector Machine. The amount of data used at this classification stage is 2,105 review data. There are three proportions of the distribution of training data and testing data, namely 90:10, 80:20, and 70:30. The following results from the classification accuracy level that has been carried out using sentiment data.

Ratio	N	NBC	SVM
90:10	1	82.46%	91.94%
	2	87.68%	91.94%
	3	81.52%	91.94%
Average		**83.89%**	**91.94%**
80:20	1	85.75%	94.29%
	2	80.76%	94.77%
	3	85.04%	95.01%
Average		**83.85%**	**94.69%**
70:30	1	85.28%	94.77%
	2	81.65%	94.93%
	3	94.18%	94.93%
Average		**83.70%**	**94.88%**
Total Average		**83.81%**	**93.84%**

Based on the results of processing the classification data above, the pattern of accuracy in the NBC algorithm is that the greater the training data, the higher the accuracy results. Whereas in the SVM algorithm, the level of accuracy will be higher if the distribution of testing data is bigger. Then based on the two algorithms it was found that the accuracy value of the SVM algorithm is better than the NBC algorithm, with an average accuracy rate of 93.84%. This result is supported because the SVM algorithm has good generalization capabilities when using data with small samples and has separators between classes resulting in better performance in solving complex problems.

Word Visualization

Visualization is carried out to display review words that users often discuss. Extraction results have been obtained from sentiment analysis based on the categories of positive and negative reviews visualized using a word cloud. On positive sentiment, it was obtained from 707 reviews that the results of the words that users often discussed were the word "application" with a frequency of 296 times, the word "help" 220 times, and the word "vaccine" 192 times. In comparison, the negative sentiment has been processed as much as 1,358 review data and obtained several words that users often use. It is known that the word "application" appears the most, with a frequency of 680 times, the word "vaccine" 671 times, and the word "certificate" 489 times—the following results from visualizing sentiment analysis and word associations using the word cloud.

Figure 3. Word cloud of Positive Sentiment

Figure 4. Word cloud of Negative Sentiment

CONCLUSION

Based on the research that has been conducted, the topics reviewed in the positive sentiment class that are often discussed are applications, assistance, vaccines, certificates, love, acceptance, new, good, entry, and benefits. As for the review, the negative sentiment class topics are often discussed: applications, vaccines, certificates, difficult, enter, help, open, data, update, date, birth, and appear. Therefore, the review topics often discussed in positive and negative sentiment class reviews are similar because some users can use the application, and some users have difficulty using the application in this version. Data processing has been done by dividing the training data and the data testing in the experiment as much as three times as randomization for each comparison. The accuracy performance results show that the Support Vector Machine (SVM) algorithm has a higher accuracy of 93.84% compared to using the Naïve Bayes Classifier algorithm (NBC), which is equal to 83.81%. The other than

that occur 8 problems identified in the negative reviews, that problem such as bad service, constraint to save vaccines certificate from apps, last update adding new problem, always failed to save vaccine evidence, always failed to login, nothing notification from apps, open the application need a long time, and hard to scan.

REFERENCES

Abidin, N. Z., Remli, M. A., Ali, N. M., Phon, D. E., Yusoff, N., Adli, H. K., & Busalim, A. H. (2020). Improving intelligent personality prediction using Myers-Briggs type indicator and random forest classifier. *International Journal of Advanced Computer Science and Applications, vol. 11, no. 11*.

Ciocodeica, D. F., Chivu, R. G., Popa, I. C., Mihalcescu, H., Orzan, G., & Bajan, A. M. (2022). The Degree of Adoption of Business Intelligence in Romanian Companies—The Case of Sentiment Analysis as a Marketing Analytical Tool. *Sustainability, vol. 14, no. 12*, 7518.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data, vol. 2, no. 1*, 1-14.

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment Analysis of Social Media Twitter with case of Anti-LGBT Campaign in Indonesia using Naive Bayes, Decision Tree, and Random Forest Algorithm. *Procedia Computer Science, Vol. 161*, 765-772.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence, vol. 1, no. 1*, 60-76.

Hakim, L., Sartono, B., & Saefuddin, A. (2017). *Bagging-based ensemble classification method on imbalance datasets.* Repositories-Dept. of Statistics: IPB University.

Hassan, S., Rafi, M., & Shaikh, S. (2011). Comparing SVM and Naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment. *IEEE 14th International Multitopic Conference*, 31-34.

Indonesia, M. o. (2022).

Isnain, A. R., Marga, S. N., & Alita, D. (2021). Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 1*, 55-64.

Kesumawati, A., & Utari, D. T. (2018). Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine. *AIP Conference Proceedings, vol. 2021, no. 1*.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, vol. 10. no. 4*, 150.

Ranjan, N. (2015). A Survey On text Mining Analytics and Classification Techniques For text Mining. *Int. J. of Dev. Research, vol. 5*, 5952-5955.

Ray, S. (2019). A quick review of machine learning algorithms. *In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 35-39.

Saleh, A. (2015). Implementation of Naïve Bayes Classifier Method for Predict Amount of Household Electricity Use. *Citec Journal*, 209-210.

Senellart, P., & Blondel, V. D. (2008). Automatic discovery of similar words. *Survey of text mining II: clustering, classification, and retrieval*, 25-44.

Vapnik, V. N. (1999). An Overview of Statistical Learning Theory. *IEEE transactions on neural networks vol. 10, no. 5*, 988-999.

Vapnik, V., & Cortes, C. (1995). Support Vector Machine. *Machine Learning, vol. 29*, 273-297.