Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 2 (2023) 414 - 421 ISSN Media Electronic: 2580-0760

Sentiment Analysis of Twitter Users to the PeduliLindungi Using Naïve Bayes Algorithm

[1]Lia Ellyanti, [2]Yova Ruldeviyani, [3]Lelianto Eko Pradana, [4]Andro Harjanto
[1,2,3,4]Magister of Technology Information, Faculty of Computer Science, Universitas Indonesia
[1]lia.ellyanti@ui.ac.id, [2]yova@cs.ui.ac.id, [3]lelianto.eko11@ui.ac.id, [4]andro.harjanto@ui.ac.id

Abstract

Covid-19 was declared as a pandemic by World Health Organization (WHO) in March 2020, has a major impact on the lives. Indonesian's government has made several efforts to suppress the spread of the virus by requiring the societies to use PeduliLindungi in every activity. There are many pros and cons from the societies in using PeduliLindungi, many reviews about the performance of this application found through playstore, app store or social media. Twitter is one of social media that allows the societies to express their feeling, idea, opinion, or critics about any topics. This study takes the review of PeduliLindungi from Twitter with period from June up to December 2021, which has the highest cases of covid-19 and tighter movement restriction from the government. The data collected were manually labeling into positive and negative class and processed using sentiment analysis with Naïve Bayes algorithm, give the result 64.69% positive sentiment and 35.5% negative sentiment regarding PeduliLindungi. The model tested using Naïve Bayes algorithm with 10-fold cross validation has the highest performance, the accuracy obtained is 95.86%, with precision 96.99% and recall 94.12%. The positive sentiment indicates the pro expression from society, like the data integration with vaccine certificate, PCR or antigen result, that makes the activities to entry public transport or public space easily. The negative sentiment indicates the cons expression from the societies, related with the performance of the application and the data security. The result of this study expected being reference, give insight, and information for developers and governments to build a better strategy in improving the performance of PeduliLindungi application.

Keywords: pedulilindungi; sentiment analysis; twitter; naïve bayes

1. Introduction

Covid-19 was declared as a pandemic by World Health Organization (WHO) in March 2020, has a major impact on the lives of most of the world's population [1]. Indonesian's government has made several efforts to suppress the spread of the virus by limiting public activities by using a tracing application called PeduliLindungi. PeduliLindungi is an application developed by collecting data from users regarding the spread of COVID-19 in the community, helping to track and confirm cases of covid, give alert if we are at red zone that indicate area with high rate covid status, and vaccination status for each individual that provides color indicators such as black, red, orange/yellow or green [2]. The Indonesian government requires the societies to use PeduliLindungi in every activity in public spaces. The uses of PeduliLindungi applies to all regions in Indonesia, in small cities or big cities. PeduliLindungi provide the vaccination status and healthy status of the user. The user with green color status allow to entry the public space which indicate that the user get full vaccination and has health condition, the red color status not allow to entry the public spaces because the user still got first doses vaccination and the user with black color status also not allow to entry the public space which it shown that the user got covid. Many public spaces require to use PeduliLindungi for check-in application, like shopping mall, concert venue, seminar, meeting venue, office building and etc. With this regulation, it will make it easier for the government to limit the movement of the individual that affected by Covid-19 so they will not transmit it to others and limit the movement of individuals who have not yet received the full vaccine.

There are many pros and cons in using PeduliLindungi application, which this application still under development and had to be improved. Many reviews about the performance of this application find through playstore, app store and social media. It is important for developers to know feedback regarding the performance of this application. The improvement of PeduliLindungi to get the best performance is needed,

Accepted: 17-11-2022 | Received in revised: 25-03-2023 | Published: 28-03-2023

because this application is a support for individual activities during the Covid-19 pandemic.

Research by Ali Mustopa et al, conducted a user review analysis of PeduliLindungi from Google Play with 1364 review data was collected from April to June 2020 using SVM and Naïve Bayes algorithm [3]. The accuracy of Naïve Bayes based PSO has value 69.00%, and the accuracy of SVM based PSO has value 93.0%. Also, research by Zulkifli et al, analyzed the data from Google Play Store review using 1000 data to get the review positive, neutral or negative by using Naïve Bayes with the accuracy about 73% [4].

All previous research using Google Play review as main data to get sentiment analysis of PeduliLindungi application. This study takes the comments or review from Twitter media social with the user of PeduliLindungi on web based, Android or IOS based. The use of social media today has a lot of influence on our actions and interactions in everyday life [5]. Social media enable everyone from anywhere or any background to broadcast messages in any language, related to any subject, with little to no filtering [6]. Twitter is a social media that offers a stream of topics to be discussed, and makes easy to get valuable and timely information [7]. Twitter allows the societies to express their feeling, idea, opinion, or critics about any topics with a very high user adoption and a quick increase in conversation volume.

The comments collected from twitter that contain keyword PeduliLindungi will be the sources for this study. The data collected will be processed using sentiment analysis as the subfiled of text mining using Naïve Bayes algorithm. Text mining is a process for extracting information, knowledge or patterns that are not yet known [8]. Text mining is done based on the results of information search, data collection, statistical analysis, machine learning and computational [9]. Text classification using Naïve Bayes is often found in categorizing text easily and quickly to implement. Sentiment analysis is an activity to extract and analyze information, opinions, and sentiments on a different situation, topic, event, product or service [10]. The use of sentiment analysis is applied to three categories of detail, namely documents, sentences, and aspects or entities [11]. Naïve Bayes is the most common and simple method for classifying texts based on Bayes' theorem in the form of assumptions with independent features [12]. Naïve Bayes is a linear classifier that provides opportunities for improvement of classification which is supported by text adaptation with better distribution [13]. Many well-known documents with text classification and sentiment analysis use the Naïve Bayes algorithm, with the advantage are create models and predict models quickly with small training data. Naïve Bayes provides predictions on text classes with negative, neutral and positive classifications, and the next algorithm is calculated using several work evaluations [14].

This study aims to find out how the societies sentiment, especially Twitter social media, is towards the use of the PeduliLindungi application. The results of this study expected being reference, give insight, and information for developers and governments to improve their application by presenting positive or negative sentiments from using PeduliLindungi.

2. Research Methods

The detail of research method shown at figure 1. First step is data collection from Twitter media social and will continue to processing and classification the data collected by using Naïve Bayes algorithm. The data that have been classified will be evaluated to know the performance of machine learning used at this research.

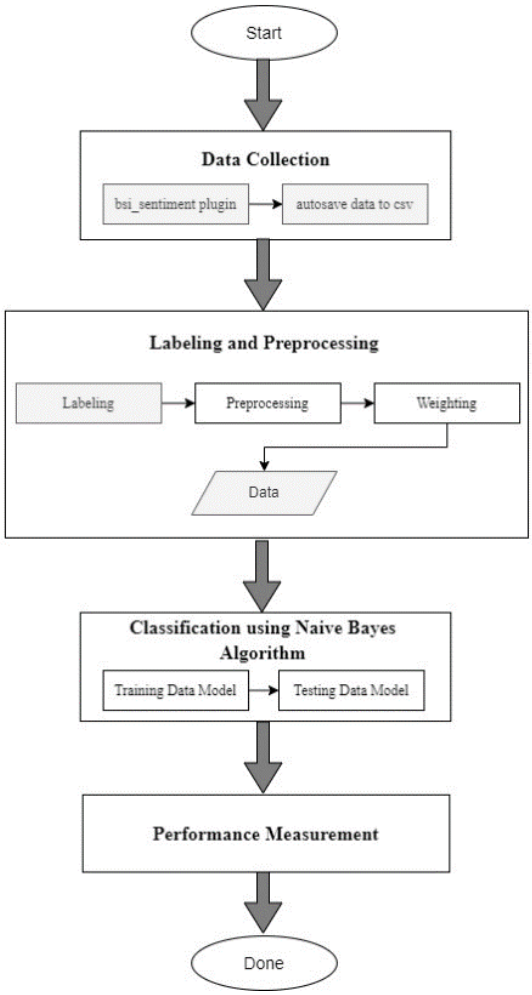

Figure 1. Research Method

2.1 Data Collection

The data use for this research collected from Twitter social media using the Twitter API by filtering the words containing "PeduliLindungi". Data was

DOI: https://doi.org/10.29207/resti.v7i2.4684

collected from 13 June 2021 up to 11 December 2021, where it has the highest cases of covid-19 and tighter movement restriction from the government that make uses of PeduliLindungi become higher. The data collection begins by crawling process using python plugins named bsi_sentiment by utilizing SNScraper which is used as a twitter data scraper and then be automatically saved into a csv format file.

By using the code program as per shown in figure 2, will be obtained a csv formatted file contains all tweets with labels and polarities. *q* is the keyword that will be used for collecting the data from Twitter. *Since* and *until* used to limit the date of retrieved data, *since* is the initial date for the data taken and *until* is the end date for the data retrieved. *Lang* is the language of the taken tweet, in this case *id* is used to retrieve Indonesian-language tweets. *Max_tweet* is the maximum number of tweet data that will be retrieved from twitter.

```
from bsi_sentiment.twitter import search_tweets_sn

tweets = search_tweets_sn(
    q="pedulilindungi",
    since="2021-12-10",
    until="2021-12-12",
    lang="id",
    max_tweets= 500
)

tweets.get_sentiment(method="textblob-nb")
tweets.to_csv('./result_dec1011.csv')
```

```
Downloading tweets:  40%|████      | 202/500 [00:03<00:04, 60.09it/s]
Analyzing tweets   : 100%|██████████| 202/202 [14:31<00:00,  4.31s/it]
Writing tweets     : 100%|██████████| 202/202 [00:00<00:00, 41523.69it/s]
```

Figure 2. Data Collection Process

The complete data are containing tweets with several columns in it, namely id, permalink (link to the website of the tweet in question), username, text, date, classification, p_pos (positive probability), p_neg (negative probability). After getting the data, will be done a manual labeling review to confirm the results of the labeling by the bsi_sentiment plugin, that will be final data processed on a machine learning used at this study.

2.2 Data Labeling

After data collection process, the next step is labeling manually to review and confirm the labelling results provided by the bsi_sentiment plugin. The manual labeling done by the writer and two other people to determine the tweet as positive or negative class. The final data with class labeling will be processed on a machine learning used at this study.

2.3 Data Preprocessing

There are several steps are used for processing the data has been collected from twitter. Data preprocessing is done by eliminating inappropriate data or converting data into a form that is easier to process by the system. First step is case folding, to process of cleaning the username, hashtag, url, punctuation, and symbol characteristics. Tokenizing for dividing each word in a sentence into individual word units, normalization for changing irregular words into regular words, filtering used to eliminate the word that frequently appear with no meaning and stemming for converting suffixed words into basic words. Figure 3 shows the step of data preprocessing used for this research.

Figure 3. Data Preprocessing Step

2.4 Weighting System

The next stage is the data transformation stage with weighting using TF-IDF (Term Frequency - Inverse Document Frequency) to determine how frequently a word appears in document and improve the accuracy of the classification results. The weight of a sentence is determined by adding the weights of its terms, which can be words, phrases, or other artificial forms [15].

2.5 Naïve Bayes

Naïve Bayes is one of the techniques for developing a classification that combines probability and statistical methods with "naive" to manage conditions between attributes independently [16]. Naïve Bayes shown high accuracy and easy to train compared to the other machine learning algorithms with low error rate for large dataset [17]. Naïve Bayes algorithm is used to predict the likelihood of a specific word belonging to a specific class. [18]. The benefits of this classification include the assumption of independence, which is required to obtain the calculation quickly, and the presence of a probabilistic hypothesis [19].

Data sets that have gone through the labeling and preprocessing will be classified using the Naïve Bayes Algorithm, which will generate training data and testing data. The trained model being used to determine positive or negative sentiments at the testing stage. Formula 1 is a formula for using the Naïve Bayes Classifier [18].

$$P\left(\frac{Y}{X}\right) = \frac{P\left(\frac{X}{Y}\right)P(Y)}{P(X)} \quad (1)$$

Y is the the hypothesis of X with specific class, X is the the probability of the true hypothesis, P(Y|X) is the probability of Y to X, P(Y) is the probability of Y independent of the tuple value and P(X): represents the probability of X with a specific value.

Next step is to measure the performance of the program. To analyze and validate the performance level of the program based on the processed data by using K-fold cross validation. At this stage, the percentage of the performance level is obtained. K-Fold cross validation is an approach in sharing data to measure the performance or quality of a model with the aim of getting the right training data [16]. K-Fold

DOI: https://doi.org/10.29207/resti.v7i2.4684

Cross Validation divides data into folds or partitions of the same class size (K1, K2,..Kn). Training and submission is repeated *n* times. During iterations, K1 becomes test data, K2 becomes training data, K3 becomes test data, and so on. Data divided into k parts allows each piece of data to stop predicting data sooner than not first [20]. The model's accuracy will be tested using test data in each fold, and this process will be repeated until the model is finished. The accuracy will be totaled and divided by the number of k.

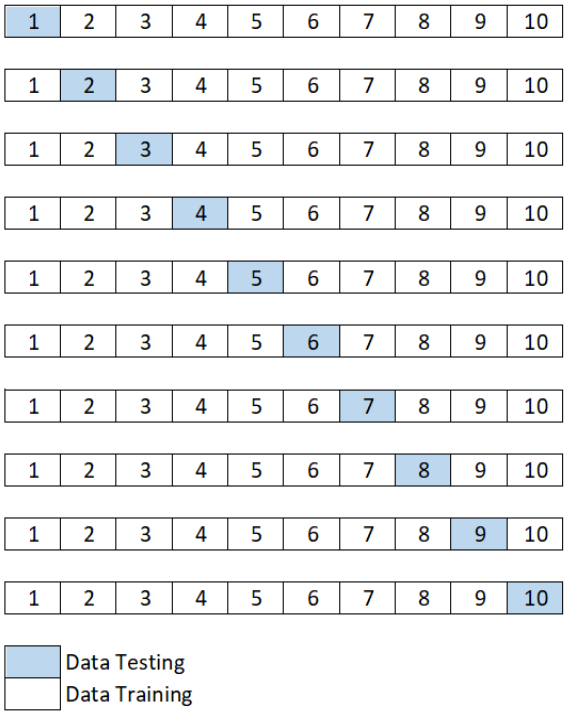

Figure 4. The 10-Fold Cross Validation

This research using 10-fold cross validation as shown ini figure 4, which this k-fold cross-validation is advised to generate estimated results with less bias. [21]. To prevent biased evaluation for real-world data sets, 10-fold cross-validation is recommended [22]. Data subsets utilized to evaluate the model should also reflect this class ratio. 10-fold cross validation repeat ten times for fitting procedure, with each fit performed on a training set consisting of 90% and the remaining 10% being used as a validation holdout set. For this 10-fold cross validation, the data is divided into ten equal-sized folds, providing ten data sets for evaluating the effectiveness of the model or algorithm for each of the ten data subsets, cross validation will employ 9-folds for training and 1-fold for testing.

2.6 Performance Evaluation

The evaluation of the models was done using a variety of indices, including the accuracy, precision, recall, and F1 score [23]. The accuracy of actual and predicted class is represented by the confusion matrix by comparing the predicted class and actual class. The accuracy shows the percentage of actual events in all examined data. Precision shows the comparation between true event and total true event predicted by classifier. Recall shows the comparation between true predictive event and total number positive event. F1 score shows the weighted average of precision and recall for that class. Formula 2, 3 , 4 and 5 are for the evaluation of the models.

$$Accuracy = \frac{true\ positive + true\ negative}{total\ number\ of\ predictions} \quad (2)$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (3)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (4)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

3. Results and Discussions

3.1 Dataset Result

The data from twitter that collected using keyword "**PeduliLindungi**" shown 51740 twitter comments. We retrieved the data by dividing it into chunks one month at the time to make the process quicker rather than retrieving it all at once. There was a spike in tweets in August and September based on the data we obtained. This happens because in both months, all public service activities must be accessed using the PeduliLindungi application.

Table 1. Positive and Negative Sentiment Classification

Classification	Total
Positive sentiment	33513
Negative sentiment	18287
Total	51740

The results of data collected shown in table 1. The tweets have been classified using bsi_sentiment plugin extension to assign positive and negative sentiment based on the obtained polarity, and review manually to confirm the labeling. There are 33513 tweets that are classified as positive sentiment and 18287 tweets that are classified as negative sentiment. From figure 5 shows the comparison between positive and negative sentiment from the data collected, which there are 64.69% positive sentiment shown with blue color, this percentage is obtained from the comparison between the total of positive sentiment and the total of overall tweet comments result, and 35.3% negative sentiment shown with orange color. The positive sentiment represents good response or positive feedback from the societies for PeduliLindungi uses. The negative sentiment represents negative feedback or bad response from the society for PeduliLindungi uses. Since it was first launched, the PeduliLindungi application has provided significant improvements, but need more improvement to get the best application for the societies.

DOI: https://doi.org/10.29207/resti.v7i2.4684

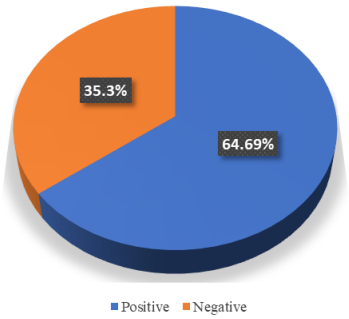

Figure 5. Comparison of Positive and Negative Sentiments about PeduliLindungi from Twitter

Table 2 shows the example of data collected from Twitter, *class pos* show the positive classification and *class neg* show negative classification, *p_pos* means as probability positive and *p_neg* means probability negative. Sentiment is scored based on the words contained in a tweet. Tweets will be labeled as positive classification if the *p_pos* scores greater than *p_neg* scores and tweets will be labeled as negative classification if the *p_neg* scored greater than *p_pos* scored. The final result of review manual labeling will be data that processed in machine learning using Naïve Bayes algorithm.

Table 2. Example of Data Collected

Tweet	Class	p_pos	p_neg
Di pokasi wisata, penggunaan barcode aplikasi PeduliLindungi dioptimalkan. (*At recreation place, the use of PeduliLindungi is optimized*)	Pos	0.75	0.25
Pelayanan PeduliLindungi ga niat banget kalo ada komplain kesalahan data, udah hampir seminggu sampe 2 kali bikin laporan masih belum ada perbaikan sampe sekarang. (*The service of PeduliLindungi not really help, if there are complaints of data errors, it's been almost a week or two times to make a report and there's still no improvement until now*)	Neg	0.38	0.62
Beragam manfaatnya aplikasi PeduliLindungi yang harus kamu gunakan dalam keseharianmu. (*The various benefits of the PeduliLindungi application that you must use in your daily life*)	Pos	0.7	0.3
Sepertinya ada bug di aplikasi Peduli Lindungi di iPhone. *(Seems there is a bug at PeduliLindungi app using on iPhone)*	Neg	0.57	0.43
Wow keeewl!!! Baru tau gw klo antigen juga di record di PeduliLindungi, kirain PCR doang. (*Wow, great! I just found that the antigen also recorded at PeduliLindungi, I thought only PCR recorder*)	Pos	0.64	0.36
Amazed ternyata hasil tes masuk PeduliLindungi beneran kampung banget gue (*Amazing! that the result of antigen or PCR test recorded at PeduliLindungi)*	Pos	0.56	0.44
Mantap juga nih integrasi data sistem vaksinasi kemenkes. Dari PeduliLindungi sampe JAKI in-sync semua. (*The data integration of the Ministry of Health's vaccination system is very great, that synchronized to PeduliLindungi and JAKI app*)	Pos	0.62	0.38
Ntar lagi akan ada issue, data PeduliLindungi bocor ke Arab. (*Later, there will be an issue, PeduliLindungi's data will leaks to Arab*)	Neg	0.41	0.59
PeduliLindungi ngabisin baterai. (*PeduliLindungi darining the battery of mobile phone).*	Neg	0.17	0.83

3.2 Test Result

Naïve Bayes algorithm generate the training data and testing data to determine positive or negative sentiments at the testing stage. Table 3 shows the data that split into training and testing data. This study tests the performance using 2 scenarios. For first testing, the comparison data used is 80:20 shown in table 4, for second testing the comparison data used is 90:10 shown in table 5.

Table 3. Scenario Performance System Test

Scenario	Training Data (%)	Testing Data (%)
First	80	20
Second	90	10

Table 4. Total of Training and Testing Data for First Scenario

Data	Total
Training Data	41392
Testing Data	10348

Table 5. Total of Training and Testing Data for Second Scenario

Data	Total
Training Data	46566
Testing Data	5274

The result of classifier predicted for training and testing data shown in table 6 and 7. True Negative (TN) indicate the number of negative data that is correctly detected, False Positive (FP) indicate negative data but is detected as positive data. True Positive (TP) indicate positive data that is correctly detected, and False Negative (FN) indicate the incorrectly predicts the negative class.

Table 6. Classifier Predicted for First Scenario

	Training Data	Testing Data
True Positive (TP)	13090	3325
True Negative (TN)	26810	6701
False Positive (FP)	1492	321
False Negative (FN)	0	0

DOI: https://doi.org/10.29207/resti.v7i2.4684

Table 7. Classifier Predicted for Second Scenario

	Training Data	Testing Data
True Positive (TP)	14554	1625
True Negative (TN)	30161	3351
False Positive (FP)	1851	198
False Negative (FN)	0	0

Table 8 shows some examples of predicted results using the Naïve Bayes Algorithm which generate training data and testing data, the model did not have overfitting that occurs when the accuracy of the training data is better than the accuracy of the testing data. In this study, overfitting was avoided by conducting 10-fold cross validation.

Table 8. Example of Predicted Result

Tweet	Actual	Predicted
PeduliLindungi eror terus. (PeduliLindungi shown error continuously)	Neg	Neg
Kartu vaksin itu terintegrasi dengan aplikasi PeduliLindungi. *(*The vaccine card is integrated with the PeduliLindungi application*)*	Pos	Pos
Apps pedulilindungi keren, tracing nya kaya apps luar negeri. (PeduliLindungi apps are cool, the process tracing like foreign apps)	Pos	Pos
Verifikasi hasil tes di app Peduli Lindungi tidak bekerja sejak semalam. (Verification of test results in the Peduli Lindungi app has not worked since last night)	Neg	Neg

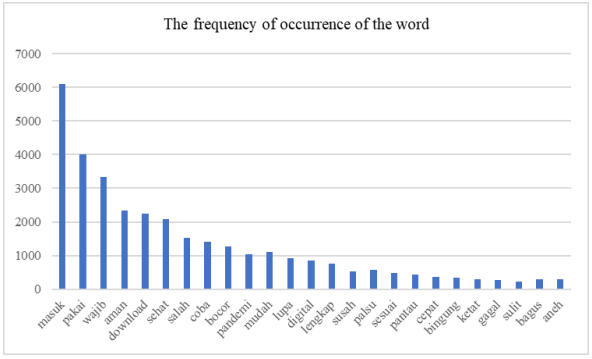

Figure 6. The frequency of occurrence of the word

Figure 6 shows the frequency of occurrence of the word after processing from the data collected, with exception PeduliLindungi word as the keyword for collecting the data. Shown that word like "*masuk"* (checkin), "*pakai"* (use), "*wajib"* (required), "*aman"* (safe), "*unduh"* (download), "*sehat"* (healthy) as the top five word occurred at data collected. This indicate that since the collecting of the data, PeduliLindungi being application that most downloaded and used by the societies since covid being pandemic situation in Indonesia. As the application that support at pandemic situation, the feedback or sentiment of the society about PeduliLindungi very diverse. The word like "*mudah*" (easy), "*lengkap*" (complete), "*sesuai*" (suitable), "*cepat*" (fast), "*bagus*" (good) indicated the positive feedback from the Twitter's user. The word like "*bocor"* (leak), "*susah"* (difficult), *"palsu"* (feak), *"bingung"* (confused) *"gagal"* (fail), *"aneh"* (odd) indicated the negative feedback from the Twitter's user.

Wordcloud at figure 7, 8, 9 and 10 indicate the variation of word and most frequent word captured from Twitter to express how the feel of the societies when they used the PeduliLindungi application in their daily life for accessing the public space. Figure 7 and 8 show the wordcloud which contain positive word in Indonesian and translate to English. Figure 9 and 10 show the wordcloud which contain positive word in Indonesian and translate to English.

Figure 7. Wordcloud in Indonesian for Positive Sentiment of PeduliLindungi

Figure 8. Wordcloud in English for Positive Sentiment of PeduliLindungi

Figure 9. Wordcloud in Indonesian for Negative Sentiment of PeduliLindungi

DOI: https://doi.org/10.29207/resti.v7i2.4684

Figure 10. Wordcloud in English for Negative Sentiment of PeduliLindungi

For positive sentiment, shown word like "*lengkap*" (complete), "*mudah*" (easy), "*sesuai*" (suitable), "integrasi" (integration), "cepat" (fast), "efektif" (effective), "valid" (*valid*), "setuju" (agree), "*akses"* (access), "*membantu*" (help), "*keuntungan*" (benefit), "*melindungi*" (protect) indicates that PeduliLindungi accepted by the societies and give the advantage for them at this pandemic situation. The societies expressed the uses of PeduliLindungi that integrated with vaccine data, PCR or antigen result, makes the activities to entry public transport or public space easily, without bringing the hardcopy of vaccine certificate or antigen result. Also, they expressed that PeduliLindungi effective help the government for monitoring the movement of society, such as the people who allow to entry the public space like mall should have green status at their PeduliLindungi, the people that have black color status could not entry to any public space so they will keep distance and do self-isolation at their house, so covid case can be suppressed.

For negative sentiment, shown word like "*eror*" (error*)*, "*salah*" (wrong*)*, "*palsu*" (fake), "*bingung*" (confused), "*ribet*" (complicated*)*, "*repot*" (troublesome*)*, "*kesal*" (annoyed*)*, "*lelah' (*tired*)*, "*susah*" (difficult), "*bocor*" (leaking), "*salah*" (false) indicates that PeduliLindungi uses make the societies activities burdened. They expressed that PeduliLindungi sometimes shown error status, could not register to PeduliLindungi and could not use when in crowded place, sometimes could not loginto the app and suddenly logout so they could not check in easily at those places and make a long queue. They also complain about vaccine certificate that not shown at PeduliLindungi., although they already get vaccine test. They also complain about fake vaccine certificates or fake antigen result that can be integrated with PeduliLindungi, this must be prevented by the Government and developers to tight monitoring of data uploaded to PeduliLindungi. The societies still worried about the security of their data that registered to PeduliLindungi, related with the data leak that has occurred.

The performance of the analysis that has been made with the help of a python plugin called *sklearn Naïve Bayes* is validated using K-fold cross validation, table 9 below are the comparison result of k-fold, between 5-fold and 10-fold, and the highest performance obtain using 10-fold cross validation.

From table 9 and figure 11, the performance system testing using 10-fold has higher performance compared to 5-fold. The accuracy, precision, recall and F1 score of 10-fold has higher value than 5-fold.

Table 9. k-fold Cross Validation Comparison Result

Fold	Accuracy	Precision	Recall	F1 Score
5	0.95767379	0.96933031	0.93992758	0.95222408
10	0.95864015	0.96998850	0.94129910	0.95334918

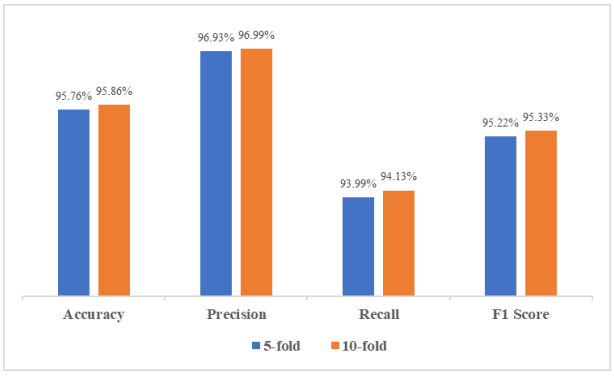

Figure 11. Comparison Performance System for 5-fold and 10-fold

Table 10. Experimental Result using Naïve Bayes with 10-fold

Index	Score
Accuracy	0.9586401499777739
Precision score	0.9699885002664572
Recall Score	0.9412991002852754
F1 Score	0.9533491810713182

The system that has been designed by using Naïve Bayes algorithm and validated using 10-fold has result as per shown at table 10, for the accuracy is 0.9586 or 95.86%, the precision score 96.99% that indicate the model is good enough in predicting positive dataset, recall score 94.12% that indicates the model is good at predicting negative datasets and F1 score 95.33% that indicate the classification model has good precision and recall score.

4. Conclusion

In this study, we conducted a sentiment analysis using Naïve Bayes Classifier to determine the response of twitter users to the keyword "**PeduliLindungi**". The data were 51740 tweets, collected from June until December 2021 which at that period was the highest cases of covid-19 period with tighter movement restriction from the government and it was required to use the PeduliLindungi for all activities. The study shown there are 64.69% positive sentiment and 35.5% negative sentiment regarding PeduliLindungi. The model tested using Naïve Bayes algorithm with 10-fold cross validation has higher performance compare with 5-fold, the accuracy obtained is 95.86%, with precision 96.99% and recall 94.12%. The positive sentiment indicates the pro expression from society related with the integration data. The negative sentiment indicates the cons expression from the societies, related with the performance of the application and the data security.

The result of this study expected being reference for developers to understand their application better,

DOI: https://doi.org/10.29207/resti.v7i2.4684

which the review from Twitter users who use the apps is absolutely needed to give insight or information for developers if there is a bug or problem, related to the performance and the data security. Also, helps the developers know how to get and classify every review from Twitter Users for PeduliLindungi application and give insight for the government to build good strategy in improving the performance of PeduliLindungi application.

This study limited to Twitter media social as the data processed with non-formal language that needed to paraphrase the word to get the good result for the modelling of machine learning. And the data only take during June until December 2021, the data collection can be expanded again by taking in the following period.

For further research, the additional data can be gathered from other social media like Instagram, TikTok, Facebook, or Webpage for collecting and analyzing reviews. It is important to know more about sentiment regarding the uses of PeduliLindungi application so it will help developers to understand their application better, which the review user at android or IOS sometimes contain biases.

References

[1] S. Ekström *et al.*, "General Stress Among Young Adults with Asthma During the COVID-19 Pandemic," *J. Allergy Clin. Immunol. Pract.*, vol. 10, no. 1, pp. 108–115, 2022, doi: 10.1016/j.jaip.2021.10.069.

[2] R. Chatterjee, S. Bajwa, D. Dwivedi, R. Kanji, M. Ahammed, and R. Shaw, "COVID-19 Risk Assessment Tool: Dual application of risk communication and risk governance," *Prog. Disaster Sci.*, vol. 7, p. 100109, 2020, doi: 10.1016/j.pdisas.2020.100109.

[3] E. Bilgic and Y. Duan, "E-commerce and Business Analytics: A Literature Review BT - Digital Economy. Emerging Technologies and Business Innovation," 2019, pp. 173–182.

[4] Z. Rais, F. T. T. Hakiki, and R. Aprianti, "Sentiment Analysis of Peduli Lindungi Application Using the Naive Bayes Method," *SAINSMAT J. Appl. Sci. Math. Its Educ.*, vol. 11, no. 1, pp. 23–29, 2022, doi: 10.35877/sainsmat794.

[5] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *Int. J. Inf. Manage.*, vol. 33, no. 3, pp. 464–472, 2013, doi: 10.1016/j.ijinfomgt.2013.01.001.

[6] A. Sarker *et al.*, "Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 10, pp. 1274–1283, 2018, doi: 10.1093/jamia/ocy114.

[7] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Secur. Inform.*, vol. 4, no. 1, 2015, doi: 10.1186/s13388-015-0024-x.

[8] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, pp. 414–419, 2016, doi: 10.14569/ijacsa.2016.071153.

[9] J. E. C. Saire and A. Pineda-Briseno, "Analysis of Covid-19 Impact in Mexico City using Text Mining and Twitter," *Proc. - 2020 Int. Conf. Digit. Transform. Innov. Technol. INCODTRIN 2020*, pp. 33–37, 2020, doi: 10.1109/Incodtrin51881.2020.00018.

[10] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021, doi: 10.1016/j.knosys.2021.107134.

[11] S. U. Hassan *et al.*, "Predicting literature's early impact with sentiment analysis in Twitter," *Knowledge-Based Syst.*, vol. 192, p. 105383, 2020, doi: 10.1016/j.knosys.2019.105383.

[12] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," *Proc. 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2016*, pp. 416–419, 2017, doi: 10.1109/ICATCCT.2016.7912034.

[13] M. Abbas, K. Ali, A. Jamali, K. Ali Memon, and A. Aleem Jamali, "Multinomial Naive Bayes Classification Model for Sentiment Analysis Overview of China View project Classification for Sentiment Analysis View project Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 3, p. 62, 2019, doi: 10.13140/RG.2.2.30021.40169.

[14] B. Seref and E. Bostanci, "Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework," *ISMSIT 2018 - 2nd Int. Symp. Multidiscip. Stud. Innov. Technol. Proc.*, 2018, doi: 10.1109/ISMSIT.2018.8567243.

[15] F. S. Fitri, M. N. S. Si, and C. Setianingsih, "Sentiment analysis on the level of customer satisfaction to data cellular services using the naive bayes classifier algorithm," *Proc. - 2018 IEEE Int. Conf. Internet Things Intell. Syst. IOTAIS 2018*, pp. 201–206, 2019, doi: 10.1109/IOTAIS.2018.8600870.

[16] R. Riyanto, "Application of the Vector Machine Support Method in Twitter Social Media Sentiment Analysis Regarding the Covid-19 Vaccine Issue in Indonesia," *J. Appl. Data Sci.*, vol. 2, no. 3, pp. 102–108, 2021, doi: 10.47738/jads.v2i3.40.

[17] M. Fahmi, Y. Yuningsih, and A. Puspita, "View of Sentiment Analysis Of Online Gojek Transportation Services On Twitter Using The Naïve Bayes Method.pdf," *J. Ilmu Pengetah. dan Teknol. Komput.*, vol. Vol.8 No.2, no. February 2023, 2023, doi: https://doi.org/10.33480/jitk.v8i2.4004.

[18] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "Sentimental Analysis of Tweets Using Naive Bayes Algorithm," *World Appl. Sci. J.*, vol. 35, no. 1, pp. 54–59, 2017, doi: 10.5829/idosi.wasj.2017.54.59.

[19] H. Krishnan, M. S. Elayidom, and T. Santhanakrishnan, "Emotion Detection of Tweets using Naïve Bayes Classifier," *Int. J. Eng. Technol. Sci. Res.*, vol. 4, no. 11, pp. 457–462, 2017.

[20] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.

[21] D. E. Cahyani and A. W. Putra, "Relevance Classification of Trending Topic and Twitter Content Using Support Vector Machine," *Proc. - 2021 Int. Semin. Appl. Technol. Inf. Commun. IT Oppor. Creat. Digit. Innov. Commun. within Glob. Pandemic, iSemantic 2021*, pp. 87–90, 2021, doi: 10.1109/iSemantic52711.2021.9573243.

[22] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[23] R. K. Saroj, P. K. Yadav, R. Singh, and O. N. Chilyabanyama, "Machine Learning Algorithms for understanding the determinants of under-five Mortality," *BioData Min.*, vol. 15, no. 1, pp. 1–23, 2022, doi: 10.1186/s13040-022-00308-8.

DOI: https://doi.org/10.29207/resti.v7i2.4684