SENTIMENT ANALYSIS OF PEDULILINDUNGI APPLICATION REVIEWS USING NAIVE BAYES CLASSIFIER AND SUPPORT VECTOR MACHINE

Irfani Firdausy[1], Suhartono[2], M. Imamudin[3]
[1-3] UIN Maulana Malik Ibrahim, Malang, Indonesia
[*]Email for Correspondence: irfani@gmail.com

Keywords:

analysis sentiment, pedulilindungi, naive bayes classifier, support vector ,machine

ABSTRACT

The spread of the Covid-19 virus since the end of 2019 in Indonesia has resulted in the Indonesian government taking several actions in various sectors. One of the government's efforts to handle and monitor the condition of the Covid-19 pandemic using information technology is by launching the PeduliLindungi application. With this application the government can monitor public data related to vaccination, tracing, telemedicine, and looking for rooms at the nearest hospital. The launch of the application and the obligation to use the PeduliLindungi application has received a response from the public, this can be monitored/seen from reviews on social media, news and also reviews on the Google Play Store regarding the application. Reviews from users on the Google Play Store can be used as parameters for input or feedback. This data is quite a lot and requires a long time to process it, even though the existing reviews could be useful as input for criticism and suggestions in future application development. From user review data, a classification process can be carried out based on sentiment type. Sentiment analysis is a branch of text classification which aims to classify sentiment (opinion) whether the text contains negative opinions, positive or neutral opinions. The aim of this research is to apply sentiment analysis to user review data of the PeduLindungi application into positive and negative classes using the Naive Bayes Classifier and Support Vector Machine classification algorithms. The dataset used after going through data pre-processing was 10,616 records. The results of testing and model evaluation carried out by randomly dividing training data and testing data obtained accuracy values, for the Naive Bayes Classifier method it was 77% and the Support Vector Machine had higher accuracy, namely 81%.

INTRODUCTION

The spread of the Covid-19 virus since the end of 2019 in Indonesia has resulted in the Indonesian government taking several actions in various sectors. One of the government's efforts to handle and monitor the condition of the Covid-19 pandemic using information technology is by launching the PeduliLindungi application. With this application the government can monitor public data related to vaccination, tracing, telemedicine, finding rooms at the nearest hospital (Putra, 2020).

The launch of the application and the obligation to use the PeduliLindungi application has received a response from the public, this can be monitored/seen from reviews on social media, news and also reviews on the Google Play Store regarding the application. On the PeduliLindung application page on the Google Play Store, there are many comments and reviews from application users. When this report was written, there were a total of more than 1 million reviews, and the PeduliLindung application had been downloaded more than 50 million times. Reviews from users on the Google Play Store can be used as parameters for input or feedback (Saputra et al., 2022). This data is quite a lot and takes a long time to process it. In fact, the existing reviews could be useful as input for criticism and suggestions in future application development. From user review data, a classification process can be carried out based on sentiment type. Sentiment analysis itself is a branch of text classification which aims to classify the sentiment (opinion) of a text automatically. Does the text contain negative opinion (negative sentiment), positive opinion (positive sentiment) or neutral (neutral sentiment). In the teachings of the Islamic religion, in terms of expressing opinions, it is recommended to express positive opinions, because this includes doing actions that contain goodness (Saputra et al., 2022).

Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases. The term data mining has the essence of being a scientific discipline whose main goal is to discover, explore, or mine knowledge from the data or information that we have (Kawani, 2019). Data mining, often also referred to as Knowledge Discovery in Databases (KDD). KDD is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets (Deolika et al., 2019).

With the increasing number of electronic documents and information in the form of text from various sources, text mining has become a potential field for study. Text mining is the science of studying how to extract information and look for patterns from a document automatically. The working principle of text mining is the same as data mining, namely mining information from large data sets (Pessôa et al., 2021). It's just that the data used in text mining is in the form of text. There are many applications of text mining such as text classification, information retrieval, topic clustering, topic extraction, document summary and sentiment analysis (Widaningsih & Suheri, 2018).

Important concepts in text classification are features and algorithms. Basically, algorithms in sentiment classification can be defined into two methods, namely rule based methods and statistical based methods. The rule-based method means that the word features used for classification are built manually by human experts, while the statistical-based method means that the word features are selected automatically by a classification algorithm through machine learning (Mustopa et al., 2020).

Sentiment analysis, also called opinion mining, is a field of study that analyzes people's opinions, sentiments, judgments, attitudes, and emotions toward entities and their attributes expressed in written text. Entities can be products, services, organizations, individuals, events, issues, or topics (Wilie, 2023). Many related names and slightly different tasks, for example, sentiment analysis, opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, influence analysis, emotion analysis and review mining, now all come under the umbrella of sentiment analysis (Azizah et al., 2022).

The sentiment classification process using statistical-based methods using the Naive Bayes Classifier (NBC) algorithm and Support Vector Machine (SVM) (ANA FATIMAH FITRIANI, 2019)is widely used in sentiment analysis in previous studies. The data pre-processing process and the use of feature extraction, such as the frequency of occurrence of words in a document (Term Frequency) or the frequency of occurrence of words in a document over the entire document (Term Frequency-Inverse Document Frequency / TF-IDF) are expected to increase the accuracy of classification results. Therefore, the author was moved to conduct further research regarding sentiment analysis of user reviews of the PeduliLindungi application (Aripin, 2018).

METHOD

The data collection method used is the data scraping method using the Python programming language (Sukma, 2021). The library package used is google_play_scraper PeduliLindung application review data which is located at https://play.google.com/store/apps/details?id=com.telkom.tracencare&showAllRe views=true, the review taken is a text review in Indonesian (Devasia, T., Vinushree, T. P., & Hegde, 2016).

There are several processes carried out in this stage, namely Case Folding, Cleaning, Tokenizing, Spelling Normalization, Stemming, Filtering / Stopword (Rozi et al., 2021). There are two approaches that can be used in the feature extraction process, namely the statistical approach and the semantic approach. This process is carried out to form a model that will be used to classify new data. This process is carried out using two methods, namely Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) (Mustopa et al., 2020). Analysis of the classification results was carried out by comparing the results of the Naive Bayes Classifier classification and the results of the Support Vector Machine classification for each combination using a confusion matrix (Accuracy, Recall, Precision, F1-score) (Qadrini et al., 2021).

RESULTS AND DISCUSSION

Classification results from two models using the Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) methods (Deolika et al., 2019), using test data that was run in the previous chapter. To apply it to the Naive Bayes Classifier and Support Vector Machine methods, the PeduliLindung application user review data obtained must be prepared through a series of pre-processing and transformation steps (feature extraction) with TF-IDF. This step must be taken before using the data in the analysis process by these two methods (FUADIN, 2017).

Table 1. Comparison of Accuracy, Precision, Recall, F1-score NBC and SVM

Test Data		NBC				SVM			
Training %	Testing%	Accuracy %	Precision %	Recall %	F1-score%	Accuracy %	Precision %	Recall %	F1-score%
80	20	77	81	73	74	81	81	80	80
70	30	77	81	73	73	80	80	79	79
60	40	77	81	72	73	81	80	79	80
50	50	77	80	72	73	80	79	78	79
Average		77	81	73	73	81	80	79	80

Table 1 shows a comparison of test results, showing that a series of tests were carried out 4 times with varying percentages of training data and testing data. From this observation, it can be seen that the accuracy value is stable, there are no large differences or differences, the accuracy is 77%. In Figure 6.1, we present a graph illustrating the accuracy test of the Naive Bayes Classifier.

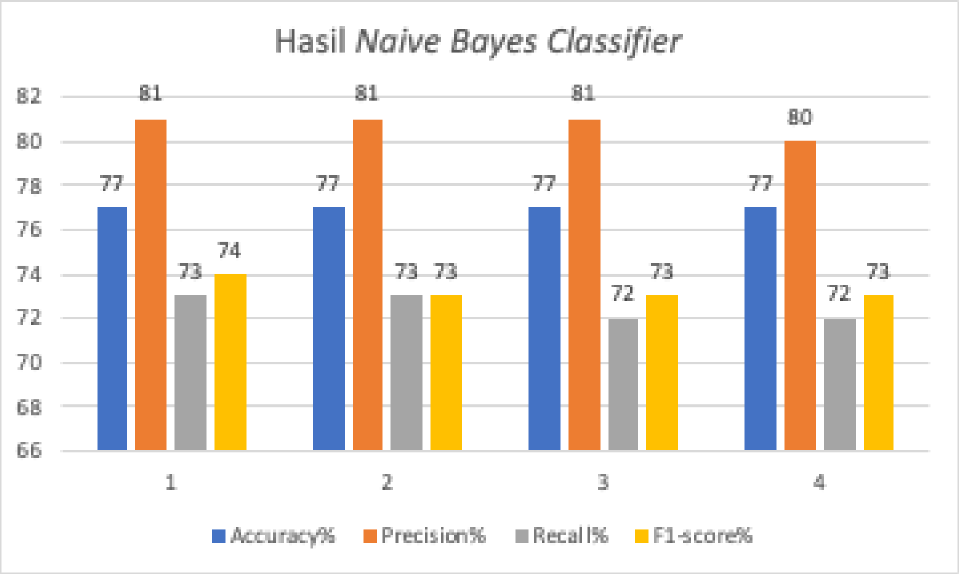


Figure 1. Naive Bayes Classifier results

Figure 1 shows that with increasing training data, there is a significant increase in accuracy values. The highest point in accuracy value appears at a ratio of 80% training data and 20% testing data (trial 1), reaching 77%. There was a steady increase despite the imbalance in the data and the random selection process of the data. To see the accuracy testing of the Support Vector Machine, the details can be seen in Figure 2

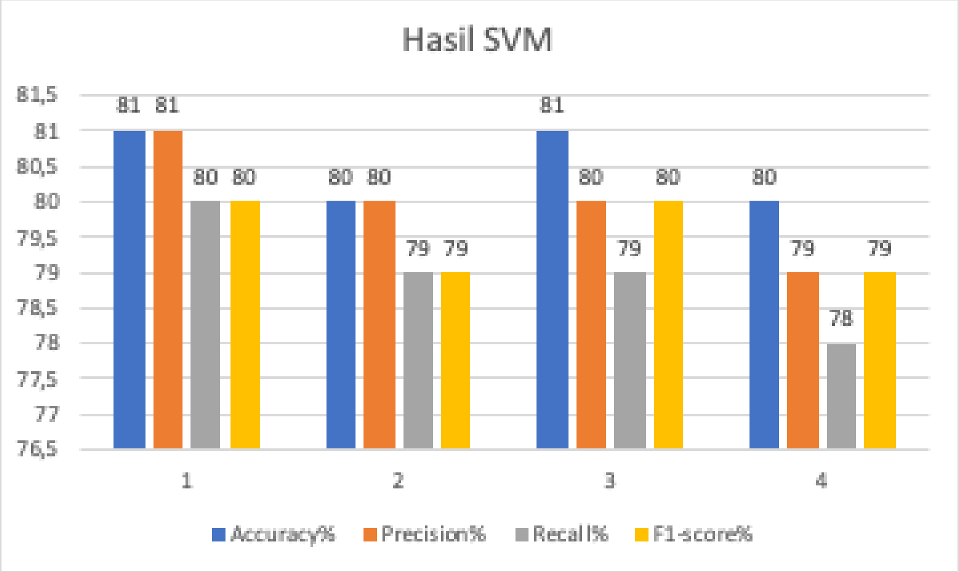


Figure 2. Support Vector Machine results

Figure 2 shows a similar pattern, with an increase in the proportion of training data, there is an increase in accuracy values. The highest point of accuracy value was recorded at a ratio of 80% training data and 20% testing data, which reached 81%. There was a steady increase despite imbalances in the data and random processes in data selection. To compare the two models, please see Figure 3.

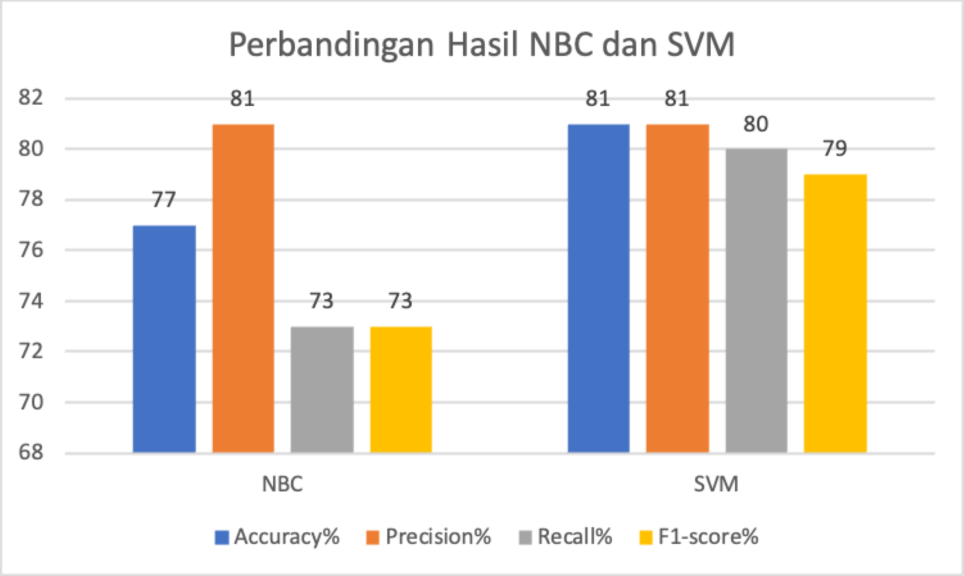
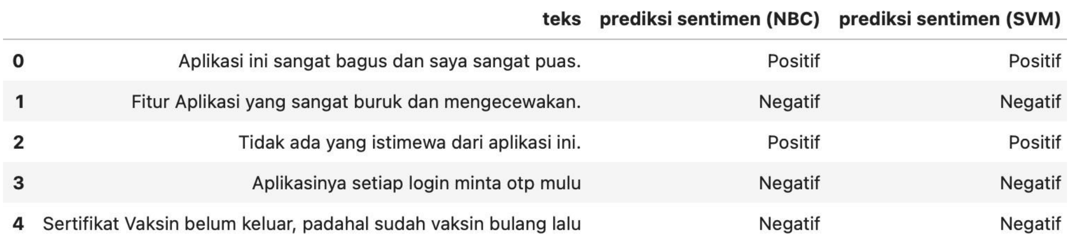


Figure 3. Comparison graph of model results

Figure 3 presents the average Accuracy values of two different methods: Naive Bayes Classifier has 77% Accuracy, 81% Precision, 73% Recall and 73% F1-score, while Support Vector Machine shows 81% Accuracy, 81% Precision, 80% Recall and F1-score 79%. Differences in performance can be seen in each method used.

a. The Naive Bayes Classifier model achieved an average accuracy of 77%. This means that of all the predictions made by the Naive Bayes Classifier model, around 77% of them are correct predictions.
b. The SVM model has an average accuracy of 81%. This indicates that the SVM model performs better than the NBC model, with a higher level of accuracy.

The results of each method are stored in a model file, namely naive_model.pkl and svm_model.pkl. To test the accuracy value, the researcher tried to create 5 sentences and these would be predicted by the two models, the results can be seen in Figure 4.



	teks	prediksi sentimen (NBC)	prediksi sentimen (SVM)
0	Aplikasi ini sangat bagus dan saya sangat puas.	Positif	Positif
1	Fitur Aplikasi yang sangat buruk dan mengecewakan.	Negatif	Negatif
2	Tidak ada yang istimewa dari aplikasi ini.	Positif	Positif
3	Aplikasinya setiap login minta otp mulu	Negatif	Negatif
4	Sertifikat Vaksin belum keluar, padahal sudah vaksin bulang lalu	Negatif	Negatif

Figure 4. Prediction results of sentiment analysis from several sentences using NBC and SVM

In Figure 4 of 5 sentences, both methods can predict correctly whether a sentence falls into positive sentiment and negative sentiment.

Researcher's knowledge information tries to display the information and knowledge obtained from this research:

a. Distribution graph of user review ratings for the PeduliLindung application

After taking data from Google Play, using the Python programming language and the google_play_scrapper library of 20,000 records, the data is saved in a CSV file, and information on the distribution of ratings (scores) from users is obtained as in Figure 5.

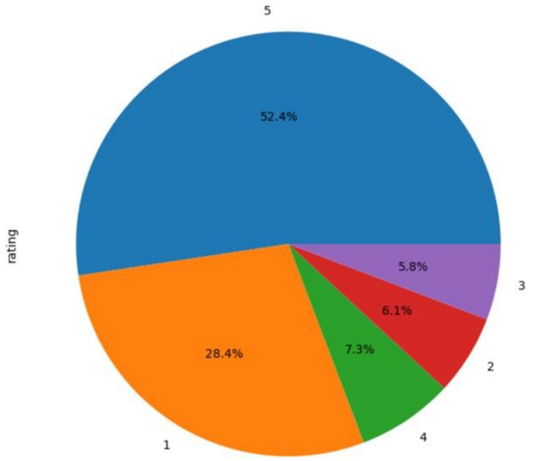


Figure 5. Distribution graph of user review ratings for the PeduliLindungi application

In Figure 5, users give a rating (score) from 1 to 5 for the PeduliLindung application, the higher the score means giving a good assessment, conversely if giving a low value means giving a less good assessment, for score 1 data there are 5681 (28.4%) , score 2 was 1218 (6.1%), score 3 was 1156 (5.8%), score 4 was 1470 (7.3%), and score 5 was 10475 (52.4%).

b. Rating distribution graph (score)

Categories become positive and negative labels. From the rating data (scores) they are then grouped into 2 categories, where if the score is > 3 it is labeled Positive and the score <= 3 is labeled Negative . From the grouping results with the column name rating_category, a graph is created like Figure 6.

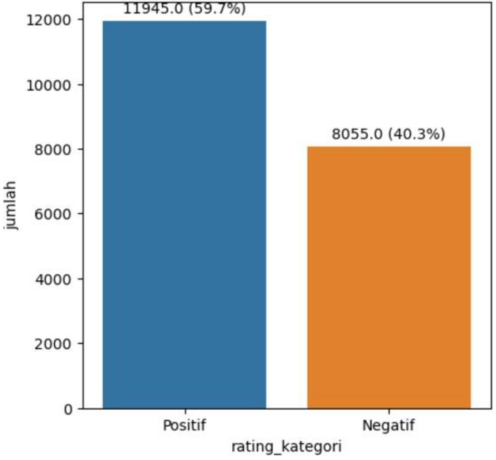


Figure 6. Graph of Rating Score Division into Positive and Negative Labels

In Figure 6, 11,945 (59.7%) positive category (label) ratings were obtained and 8,055 (40.3%) positive category ratings were obtained from 20,000 records.

c. Graph of the number of positive and negative labels after data pre-processing

After pre-processing the data, the amount of data was reduced, from 20000 to 10616, the reduction in data was due to the deletion of empty text (no values) and the deletion of duplicate records, because the same or duplicate text data could affect the accuracy results.

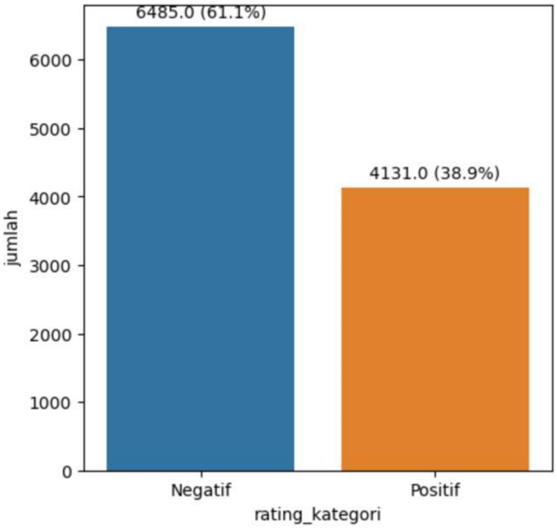


Figure 7. Graph of the Number of Positive and Negative Labels After Data Pre-processing

In Figure 7, after going through data pre-processing, we obtained a positive category (label) rating of 4132 (38.9%) and a negative label of 6485 (61.1%).

d. Data on positive sentiment and negative sentiment words that often appear as a result of the Naive Bayes Classifier and Support Vector Machine methods

To display positive and negative sentiment words, the author uses a wordcloud, which is a visualization of a collection of words that are often mentioned in a particular media (Mustopa et al., 2020). For example, on social media, wordcloud will collect many trending words. Words that appear frequently will be the largest and most prominent, while other small words will be seen surrounding the largest words, as seen in Figure 8 and Figure 9.

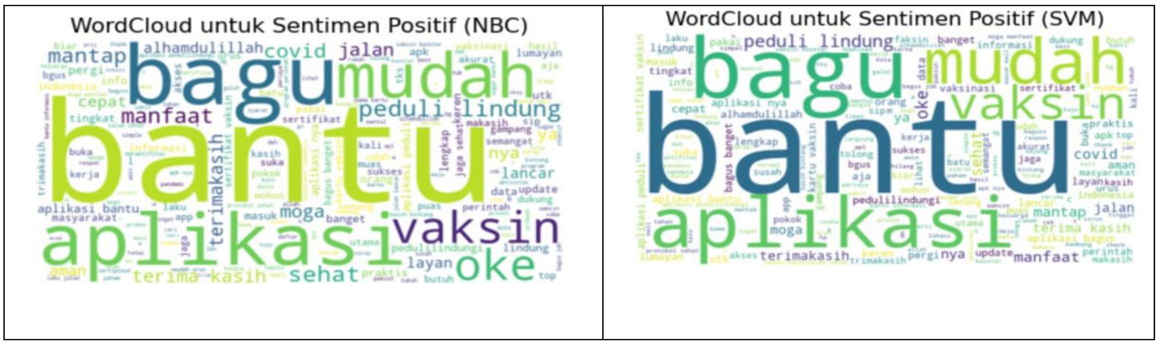


Figure 8. Wordcloud of Positive Sentiment Words from the Naive Bayes Classifier Method (left) and Support Vector Machine (right)

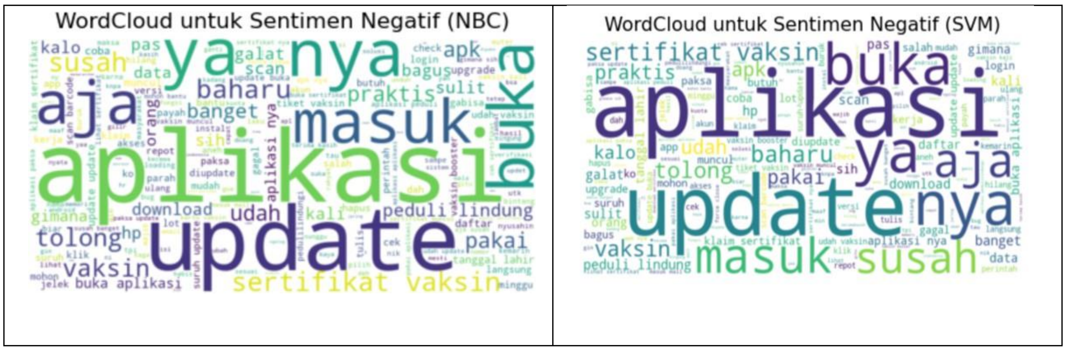


Figure 9. Wordcloud of Negative Sentiment Words from the Naive Bayes Classifier Method (left) and Support Vector Machine (right)

The number of words labeled with positive sentiment that have the highest frequency of appearance is shown in Table 2, while the number of words labeled with negative sentiment is shown in Table 2.

Table 2. Comparison of 3 Positive Sentiment Words with the Most Frequency of Appearance

No	Positive Words	Frequency of Occurrence (NBC)	Frequency of Occurrence (SVM)
1	help	479	528
2	Good	441	517
3	application	433	499

In Table 2, the word "help" is ranked first with the highest frequency of appearance for positive sentiment, with the Naive Bayes Classifier method it has a frequency of appearance of 479 times, while for Support Vector Machine the frequency of appearance is 528 times.

Table 3. Comparison of 3 Negative Sentiment Words with the Most Frequency of Appearance

No	Negative Words	Frequency of Occurrence (NBC)	Frequency of Occurrence (SVM)
1	application	2452	2394
2	updates	2136	2115
3	vaccine	1632	1564

In Table 3 the word " application " occupies ranking First with frequency appeared the most For sentiment negative , with method *Naive Bayes Classifier* has frequency emergence 2452 times, meanwhile for *Support Vector Machine* frequency its appearance 2394 times.

CONCLUSION

From the results of the research conducted, there are several conclusions from the comparison of two Data Mining-based classification methods, namely Naive Bayes Classifier (NBC) and Support Vector Machine (SVM), regarding sentiment analysis of PeduliLindungi application reviews as follows: Study This succeed build a classification model For predict analysis sentiment into two classes that is class positive and negative . Data retrieved from user reviews from Indonesia and Indonesian on Google Play which was taken on October 20 2022, and evaluated use Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) methods . Comparison and evaluation of two methods show that a Support Vector Machine (SVM) has accuracy more tall of 81%, compared to the Naive Bayes Classfier (NBC) which has 77% accuracy Sentiment said positive with frequency emergence the most is the word “ help ”, to Naive Bayes Classifier method emerged 479 times, and Support Vector Machine 528 times . For the word sentiment negative with frequency emergence the most is the word “ application ”, with the Naive Bayes Classfier (NBC) method emerged 2452 times , while Support Vector Machine (SVM) appeared 2394 times.

REFERENCES

Ana Fatimah Fitriani. (2019). Analisis Kemampuan Technological Pedagogical Content Knowledge (Tpck) Calon Guru Biologi Universitas Islam Negeri Raden Intan Lampung. Jurnal Kajian Pendidikan Ekonomi Dan Ilmu Ekonomi, 1–71.

Aripin, A. A. (2018). Potensi Pemanfaatan Teknologi Blockchain Terhadap Ketepatan Waktu, Efisiensi Dan Keamanan Proses Operasi Pada Subsektor Perbankan. Skripsi Universitas Katolik Parahyangan.

Azizah, L. M., Ajipratama, D. B., Putri, N. A. R., & Damarjati, C. (2022). Analisa Sentimen Masyarakat Terhadap Kebijakan Vaksinasi Covid-19 Di Indonesia Pada Twitter Menggunakan Algoritma Lstm La. Jurnal Iptekkom Jurnal Ilmu Pengetahuan & Teknologi Informasi, 24(2), 161–172. Https://Doi.Org/10.17933/Iptekkom.24.2.2022.161-172

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. Jurnal Teknologi Informasi, 3(2), 179. Https://Doi.Org/10.36294/Jurti.V3i2.1077

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction Of Students Performance Using Educational Data Mining. International Conference On Data Mining And Advanced Computing (Sapience), 91–95.

Fuadin, D. N. (2017). Deteksi Botnet Menggunakan Naïve Bayes Classifier Dengan Smote Dan Metode Bfs. Telematika-Cio, Bidang Keahlian Teknik.

Kawani, G. P. (2019). Implementasi Naive Bayes. Journal Of Informatics, Information System, Software Engineering And Applications (Inista), 1(2), 73–81. Https://Doi.Org/10.20895/Inista.V1i2.73

Mustopa, A., Hermanto, Anna, Pratama, E. B., Hendini, A., & Risdiansyah, D. (2020). Analysis Of User Reviews For The Pedulilindungi Application On Google Play Using The Support Vector Machine

And Naive Bayes Algorithm Based On Particle Swarm Optimization. 2020 5th International Conference On Informatics And Computing, Icic 2020, 2. Https://Doi.Org/10.1109/Icic50835.2020.9288655

Pessôa, L. C., Deamici, K. M., Pontes, L. A. M., Druzian, J. I., & Assis, D. De J. (2021). Technological Prospection Of Microalgae-Based Biorefinery Approach For Effluent Treatment. Algal Research, 60(May). Https://Doi.Org/10.1016/J.Algal.2021.102504

Putra, S. (2020). Analisis Tows (Threats , Opportunity , Weakness , Strenghts) Terhadap Strategi Pemasaran Pada Cv .

Qadrini, L., Sepperwali, A., & Aina, A. (2021). Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial. Jurnal Inovasi Penelitian, 2(7), 1959–1966.

Rozi, F., Sukmana, F., & Adani, M. N. (2021). Pengelompokkan Judul Buku Dengan Menggunakan Algoritma K-Nearest Neighbor (K-Nn) Dan Term Frequency – Inverse Document Frequency (Tf-Idf). Jimp: Jurnal Informatika Merdeka Pasuruan, 6(3), 1–5.

Saputra, I., Djatna, T., Siregar, R. R. A., Kristiyanti, D. A., Yani, H. R., & Riyadi, A. A. (2022). Text Mining Of Pedulilindungi Application Reviews On Google Play Store. Faktor Exacta, 15(2), 101–108. Https://Doi.Org/10.30998/Faktorexacta.V15i2.10629

Sukma, H. (2021). Clustering Data Siswa Smpn-6 Palangka Raya Untuk Menentukan Kelayakan Bantuan. 1–59.

Widaningsih, S., & Suheri, A. (2018). Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian Web Of Science Dengan Menggunakan Text Mining. Seminar Nasional Teknologi Informasi Dan Komunikasi (Sentika), 2018(March), 23–24.

Wilie, D. P. (2023). Analisis Sentimen Opini Publik Terhadap Chatgpt Di Twitter Menggunakan Metode Naive Bayes. Jurnal Nasional Ilmu Komputer, 4(4), 2746–1343.

Ana Fatimah Fitriani. (2019). Analisis Kemampuan Technological Pedagogical Content Knowledge (Tpck) Calon Guru Biologi Universitas Islam Negeri Raden Intan Lampung. Jurnal Kajian Pendidikan Ekonomi Dan Ilmu Ekonomi, 1–71.

Aripin, A. A. (2018). Potensi Pemanfaatan Teknologi Blockchain Terhadap Ketepatan Waktu, Efisiensi Dan Keamanan Proses Operasi Pada Subsektor Perbankan. Skripsi Universitas Katolik Parahyangan.

Azizah, L. M., Ajipratama, D. B., Putri, N. A. R., & Damarjati, C. (2022). Analisa Sentimen Masyarakat Terhadap Kebijakan Vaksinasi Covid-19 Di Indonesia Pada Twitter Menggunakan Algoritma Lstm La. Jurnal Iptekkom Jurnal Ilmu Pengetahuan & Teknologi Informasi, 24(2), 161–172. Https://Doi.Org/10.17933/Iptekkom.24.2.2022.161-172

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. Jurnal Teknologi Informasi, 3(2), 179. Https://Doi.Org/10.36294/Jurti.V3i2.1077

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction Of Students Performance Using Educational Data Mining. International Conference On Data Mining And Advanced Computing (Sapience), 91–95.

Fuadin, D. N. (2017). Deteksi Botnet Menggunakan Naïve Bayes Classifier Dengan Smote Dan Metode Bfs. Telematika-Cio, Bidang Keahlian Teknik.

Kawani, G. P. (2019). Implementasi Naive Bayes. Journal Of Informatics, Information System, Software Engineering And Applications (Inista), 1(2), 73–81. Https://Doi.Org/10.20895/Inista.V1i2.73

Mustopa, A., Hermanto, Anna, Pratama, E. B., Hendini, A., & Risdiansyah, D. (2020). Analysis Of User Reviews For The Pedulilindungi Application On Google Play Using The Support Vector Machine And Naive Bayes Algorithm Based On Particle Swarm Optimization. 2020 5th International Conference On Informatics And Computing, Icic 2020, 2. Https://Doi.Org/10.1109/Icic50835.2020.9288655

Pessôa, L. C., Deamici, K. M., Pontes, L. A. M., Druzian, J. I., & Assis, D. De J. (2021). Technological Prospection Of Microalgae-Based Biorefinery Approach For Effluent Treatment. Algal Research, 60(May). Https://Doi.Org/10.1016/J.Algal.2021.102504

Putra, S. (2020). Analisis Tows (Threats , Opportunity , Weakness , Strenghts) Terhadap Strategi Pemasaran Pada Cv .

Qadrini, L., Sepperwali, A., & Aina, A. (2021). Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial. Jurnal Inovasi Penelitian, 2(7), 1959–1966.

Rozi, F., Sukmana, F., & Adani, M. N. (2021). Pengelompokkan Judul Buku Dengan Menggunakan Algoritma K-Nearest Neighbor (K-Nn) Dan Term Frequency – Inverse Document Frequency (Tf-Idf). Jimp: Jurnal Informatika Merdeka Pasuruan, 6(3), 1–5.

Saputra, I., Djatna, T., Siregar, R. R. A., Kristiyanti, D. A., Yani, H. R., & Riyadi, A. A. (2022). Text Mining Of Pedulilindungi Application Reviews On Google Play Store. Faktor Exacta, 15(2), 101–108. Https://Doi.Org/10.30998/Faktorexacta.V15i2.10629

Sukma, H. (2021). Clustering Data Siswa Smpn-6 Palangka Raya Untuk Menentukan Kelayakan Bantuan. 1–

59.

Widaningsih, S., & Suheri, A. (2018). Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian Web Of Science Dengan Menggunakan Text Mining. Seminar Nasional Teknologi Informasi Dan Komunikasi (Sentika), 2018(March), 23–24.

Wilie, D. P. (2023). Analisis Sentimen Opini Publik Terhadap Chatgpt Di Twitter Menggunakan Metode Naive Bayes. Jurnal Nasional Ilmu Komputer, 4(4), 2746–1343.