Institute of Research and Publication Indonesia
IJIRSE: Indonesian Journal of Informatic Research and Software Engineering
Journal Homepage: https://journal.irpi.or.id/index.php/ijirse
Vol. 5. No.2. September 2025, pp: 98 - 106
E-ISSN: 2775-5754 | P-ISSN: 2797-2712

Information Gain Feature Selection for Temporal Sentiment Analysis of Pedulilindungi Application Review using Naïve Bayes Classifier Algorithm

Siti Syahidatul Helma[1*], Dini Hidayatul Qudsi[2], Ivan Chatisa[3]

[1,2,3]Department of Information Technology, Caltex Riau Polytechnic, Indonesia
E-Mail: [1]helma@pcr.ac.id, [2]dinihq@pcr.ac.id, [3]ivan@pcr.ac.id

Makalah: Received July 20th 2025; Revised July 31th 2025; Accepted August 05 th 2025
Corresponding Author: Siti Syahidatul Helma

Abstract

The Indonesian government requires the public to use the Pedulilindungi application to mitigate the spread of the COVID-19 virus. Users can download and access the Pedulilindungi application through the Google Play Store. There, users can directly assess an application by providing reviews that can describe user responses and satisfaction with the application. These reviews generate large amounts of text data that can be analysed using a text mining approach. Through the text mining process, review data is extracted and analysed to uncover patterns and user sentiments over time. This study applied the Naïve Bayes Classifier (NBC) algorithm to create a time-based temporal sentiment classification model. Prior to classification, a feature selection process with Information Gain is performed. Based on the experimental results, the best evaluation was produced on temporal data dated September 03, 2021, with an accuracy of 91.9% and precision and recall values of 99.9% and 91.9%, respectively.

Keyword: Feature Selection, Information Gain, Naïve Bayes Classifier, Pedulilindungi, Text Mining

1. INTRODUCTION

The COVID-19 pandemic has spread to 220 countries including Indonesia, with a total of 4,129,020 cases as of August 30, 2021, according to data obtained from the official website of the World Health Organization (WHO) [1]. One of the policies implemented by the Indonesian government to break the chain of COVID-19 transmission is contact tracing of individuals confirmed to have COVID-19 or those who have had close contact with confirmed cases using the Pedulilindungi application [2]. Based on the Minister of Communication and Information Technology Decree No. 171 of 2020, the Pedulilindungi app is used for tracing, tracking, and issuing warnings and fencing alerts to app users, in this case, the Indonesian public [3]. According to the Minister of Home Affairs of the Republic of Indonesia's Instruction No. 38 of 2021 regarding the Implementation of Community Activity Restrictions (PPKM), all members of the public are required to use the Pedulilindungi app for screening when in crowded areas, public facilities, or other places or locations, where this policy is effective from August 31, 2021, to September 6, 2021, and is updated periodically [4].

Users can download and use the Pedulilindungi application through the Google Play Store to access the application. On the Google Play Store, users can directly rate the application by giving it a score from 1 to 5 and provide a review reflecting their feedback and satisfaction with the application [5]. Using a sentiment analysis approach, we can use user reviews to evaluate applications and understand user sentiment towards them effectively and efficiently [6]. Sentiment analysis is the process of understanding, interpreting, and identifying the intent contained in textual opinion statements, which is part of the field of text mining [7]. Text mining is the process of mining large amounts of data to obtain new information and discover interesting sentence patterns using machine learning algorithms [8]. One type of sentiment analysis that can be performed is temporal sentiment analysis. Temporal sentiment analysis aims to analyze temporal trends in data with time variables, thereby identifying prominent sentiments in specific periods [8]. Research related to reviewing data from the Google Play Store app market has also been used previously to perform sentiment analysis on several applications, such as the Zoom Meeting app [5], Provider by.U [9], Go-Jek [10], Grab [11], Halodoc [12], and other applications.

One of the machine learning algorithms, the Naïve Bayes Classifier (NBC), can be applied to model data classification [13]. By applying text mining and data mining techniques, relevant and specific information can be obtained by classifying data into three opinion categories: negative, positive, and neutral [14]. The NBC algorithm is a probabilistic classification algorithm based on Bayes' theorem. It is widely used in several cases because it is simple, efficient, performs well on datasets [13], and has a high learning efficiency rate by estimating all probabilities during modeling with training data. [15] conducted previous research on the NBC algorithm to categorize documents into 20 categories, with 1,000 documents for each category. In that study, the researchers also compared the NBC algorithm and Support Vector Machine (SVM) using 10-fold cross-validation, with the NBC algorithm showing a significant improvement over SVM, with an improvement rate of +28.78% compared to SVM's +6.36%. In 2016, Dey, L. *et al.* conducted a sentiment analysis on movie reviews from the website www.imdb.com, comprising 5,000 positive and 5,000 negative review records, using the NBC and K-NN approaches. The NBC algorithm achieved an accuracy rate above 80% and performed better than the K-NN approach [16].

On the other hand, NBC is very sensitive to too many features, which can result in low classification results [17]. One way to address this issue is to apply feature selection techniques using the Information Gain method to select features that influence the dataset. Information Gain is one of the feature selection techniques that can be used to select the best terms in text data classification by measuring the likelihood of a word's occurrence and non-occurrence using entropy values [17] [18]. Maulida *et al.* 2016 used Information Gain to select features in Indonesian thesis abstract documents using various threshold values of 0.02, 0.05, and 0.07. In that study, Information Gain reduced features by up to 89% at the 0.07 threshold [18]. In 2019, Information Gain was used to select features in a tweet text dataset to classify Indonesian-language hate speech tweets using the NBC algorithm. Using Information Gain with a threshold of 80% of the entire dataset improved algorithm performance, achieving accuracy, precision, and recall of 98%, 100%, and 96%, respectively [19].

Based on previous research and supported by the issues presented, this study conducted sentiment analysis on user reviews of the Pedulilindungi application obtained from the Google Play Store app market using the NBC classification algorithm with the Information Gain technique as feature selection. This study will provide helpful information for stakeholders, particularly application providers and developers. This study is also useful for determining the effect of Information Gain implementation on the performance of the NBC classification model, thereby improving the performance of temporal sentiment analysis modelling.

2. MATERIALS AND METHOD

This research consists of four main stages, namely data collection, data preprocessing, classification, and sentiment analysis classification. The research stage scheme can be seen in Figure 1. The data used is user review data from the Pedulilindungi application, collected from the Google Play store app market via the URL link "*play.google.com/store/apps/details?id=com.telkom.tracencare&hl=id&gl=US&showAllReviews=true*" and the *google_play_scraper* library available in the Python programming language. The data collection process yielded 24,214 review records from August 31 to September 6, 2021, by implementing the obligation to use the Pedulilindungi app as stated in the Indonesian Minister of Home Affairs Instruction No. 38 of 2021.

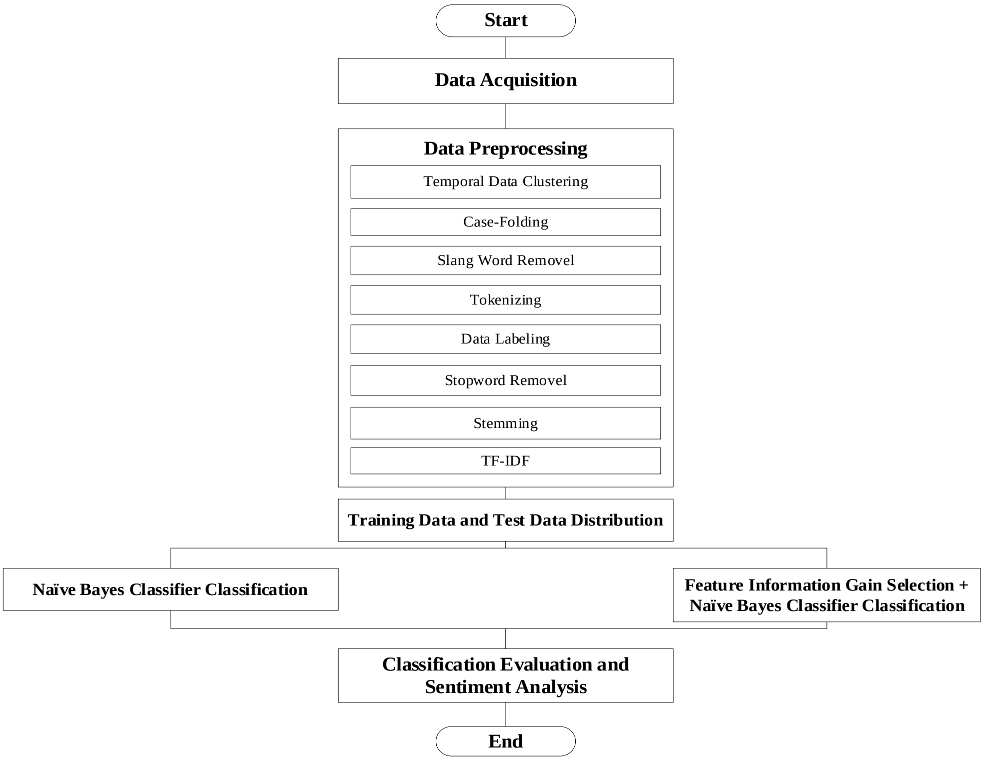

Start
Data Acquisition
Data Preprocessing
Temporal Data Clustering
Case-Folding
Slang Word Removel
Tokenizing
Data Labeling
Stopword Removel
Stemming
TF-IDF
Training Data and Test Data Distribution
Naïve Bayes Classifier Classification
Feature Information Gain Selection + Naïve Bayes Classifier Classification
Classification Evaluation and Sentiment Analysis
End

Figure 1. Research Methodology

2.1 Pedulilindungi

The Pedulilindungi application has been established through the Decree of the Minister of Communication and Information Technology Number 171 of 2020 concerning the Establishment of the Pedulilindungi Application in the Context of Implementing Health Surveillance in Handling Coronavirus Disease 2019 (COVID-19). The Pedulilindungi application, released on March 27, 2020, was built and developed by PT. Telekomunikasi Indonesia Tbk, whose copyright is exclusively licensed to the Ministry of Communication and Information Technology of the Republic of Indonesia. It is used to trace, track, and issue warnings to application users, in this case, the Indonesian public [3]. The Pedulilindungi app must be maintained and developed to fulfill its role and functions in tracing, tracking, warning, and fencing COVID-19 among the Indonesian public [20].

2.2 Text Mining

Text mining, also known as text data mining, is the process of extracting knowledge from textual databases through a semi-automated process to identify patterns in the data [21]. Text mining is discovering information where users interact with a collection of documents over time using analytical tools [22]. According to Berry and Kogan (2010), text mining can address classification, clustering, information extraction, and information retrieval issues. The primary process in this technique is identifying words that represent the content of documents for subsequent analysis of the relationships between documents [23].

In text mining, certain stages are required to process textual data into a more structured format. One of the stages in text mining is preprocessing. It is the stage where data is prepared for further processing in the data analysis stage [24]. The preprocessing stage in this study includes tokenizing, which is the stage of separating input strings based on the set of words that compose them [25]. There is the stopword removal stage, which is the stage of removing words contained in the stopword dictionary. Stopwords are very common and frequently occurring words, such as conjunctions or linking words that do not affect sentiment, for example, words like "to," "that," "this," "is," and so on [26]. The next step, the process of mapping and parsing various forms (variants) of words into their base forms (stems) is performed, commonly referred to as the stemming process [27].

2.3 Naïve Bayes Classifier

The Naïve Bayes Classifier (NBC) algorithm is based on probability and statistical methods developed by the English scientist Thomas Bayes. NBC can be used to predict future probabilities based on previous experiences. According to Andini (2013), the advantage of using the NBC algorithm is that it only requires a small amount of training data to estimate the parameters (mean and variance of variables) needed for classification. This is because the independent variables are assumed, and only the variables of each class will have their variance determined, not the entire covariance matrix. NBC belongs to the Bayesian learning algorithm by calculating explicit probabilities to describe the sought hypothesis [28]. Mahmudy and Widodo (2015) state that the NBC algorithm can directly determine the hypothesis without going through a process or performing a search by calculating the frequency of each word's occurrence in the training data. The advantage of applying the NBC algorithm is that it can reduce data noise in large datasets [29]. However, the drawback is that it is susceptible to too many features, resulting in low classification accuracy [30]. NBC is a classification algorithm based on Bayes' theorem with the assumption of independence [7], used to predict data as accurately as possible [5]. In applying the NBC algorithm, the following equation is used:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

Where X is data with an unknown class, H is defined as the hypothesis of data X is a specific class, P(H|X) is the probability of hypothesis H based on condition X, P(H) is the probability of hypothesis H (prior probability), P(X|H) is the probability of X based on that condition, and P(X) is the probability of X.

2.4 Information Gain

Information gain is an approach to feature evaluation based on information theory, which measures how much a feature contributes to reducing entropy or uncertainty in a dataset [31]. Entropy is a value of uncertainty in a class calculated using the probability of occurrence in each feature [32]. The features with a higher Information Gain value are more important for improving the classification process [33].

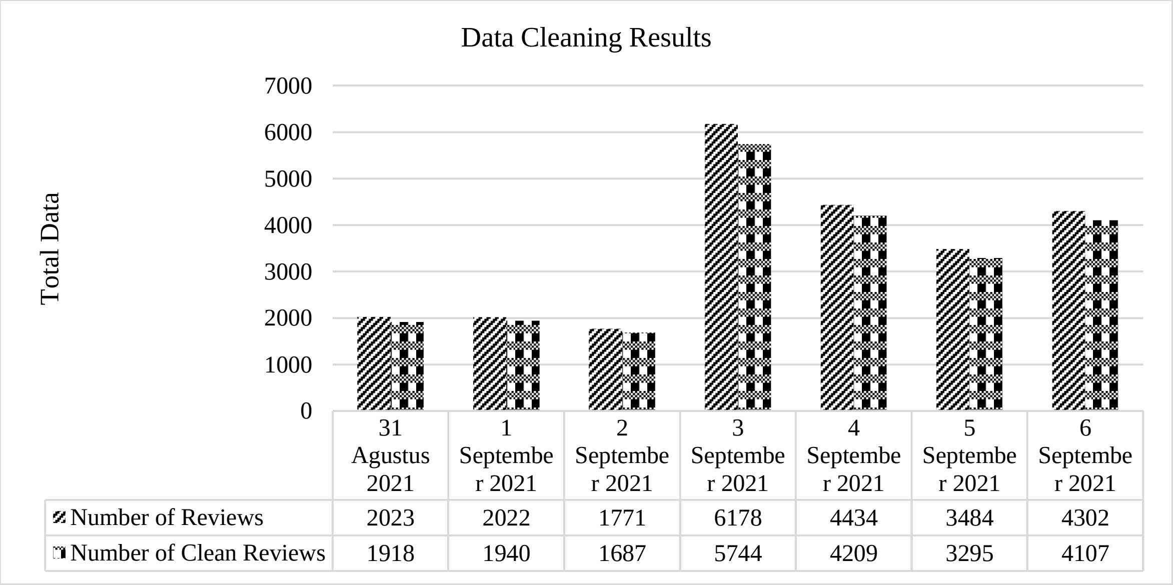
3. RESULTS AND DISCUSSION

Several processes are carried out at this stage, starting with (a) the data preprocessing stage, (b) classification modeling with the NBC algorithm, and (c) sentiment analysis of the classification results.

3.1 Data Preprocessing

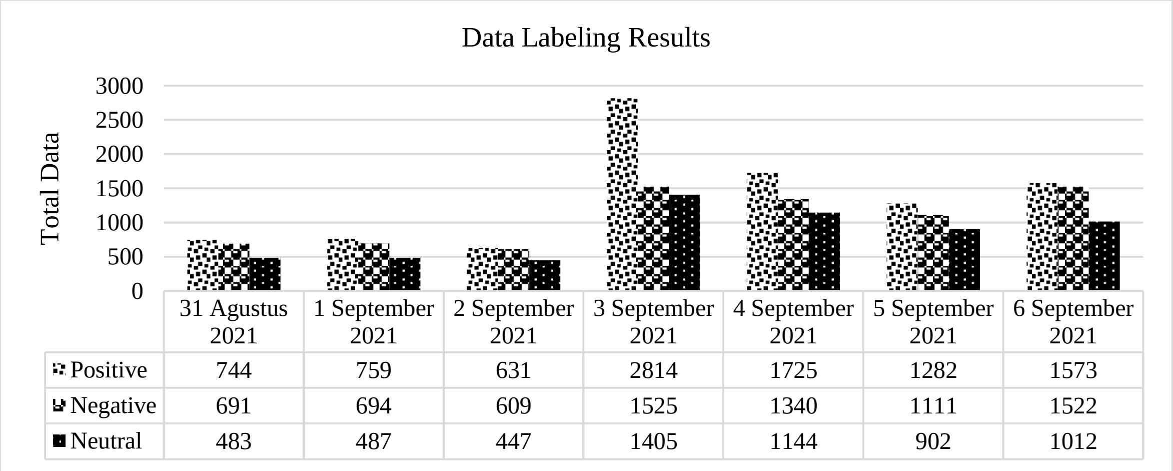
Data preprocessing includes case folding, tokenizing, filtering, slang word removal, stopword removal, data labeling, stemming, and data weighting. The case folding stage involves converting text data to lowercase letters (case folding). Then, the input strings dataset is trimmed during the tokenizing stage based on each composed word, where words are cut to be separated for further processing. The data filtering stage is the process of removing characters other than the alphabet; removing symbols, emoticons, and punctuation marks; removing whitespace (spaces, tabs, newlines); removing URLs or links from each review; and changing slang words or colloquial/slang terms. Then, the data labeling process is carried out to group the data based on a dictionary of negative, positive, and neutral sentiments.

The next stage is the removal of stopwords from the data, where terms or words in each sentence that are unrelated despite having a high frequency of occurrence need to be removed because they do not change the meaning of the review. Stopwords are words that are very common and frequently appear, such as conjunctions or linking words. Stopword removal is done to reduce features in the data and reduce computational load. Some examples of words contained in stopwords are "for," "and," "or," and so on. Then, words that have gone through the cleaning and labeling process will go through the steaming stage, which is finding the root words of each previous data result. The cleaning process resulted in 22,900 clean data records. The results of the cleaning and labeling processes for the review data are shown in Figures 2 and 3, respectively.



	31 Agustus 2021	1 Septembe r 2021	2 Septembe r 2021	3 Septembe r 2021	4 Septembe r 2021	5 Septembe r 2021	6 Septembe r 2021
▨ Number of Reviews	2023	2022	1771	6178	4434	3484	4302
▦ Number of Clean Reviews	1918	1940	1687	5744	4209	3295	4107

Figure 2. Data Cleaning Results



	31 Agustus 2021	1 September 2021	2 September 2021	3 September 2021	4 September 2021	5 September 2021	6 September 2021
⁘ Positive	744	759	631	2814	1725	1282	1573
▩ Negative	691	694	609	1525	1340	1111	1522
■ Neutral	483	487	447	1405	1144	902	1012

Figure 3. Data Labeling Results

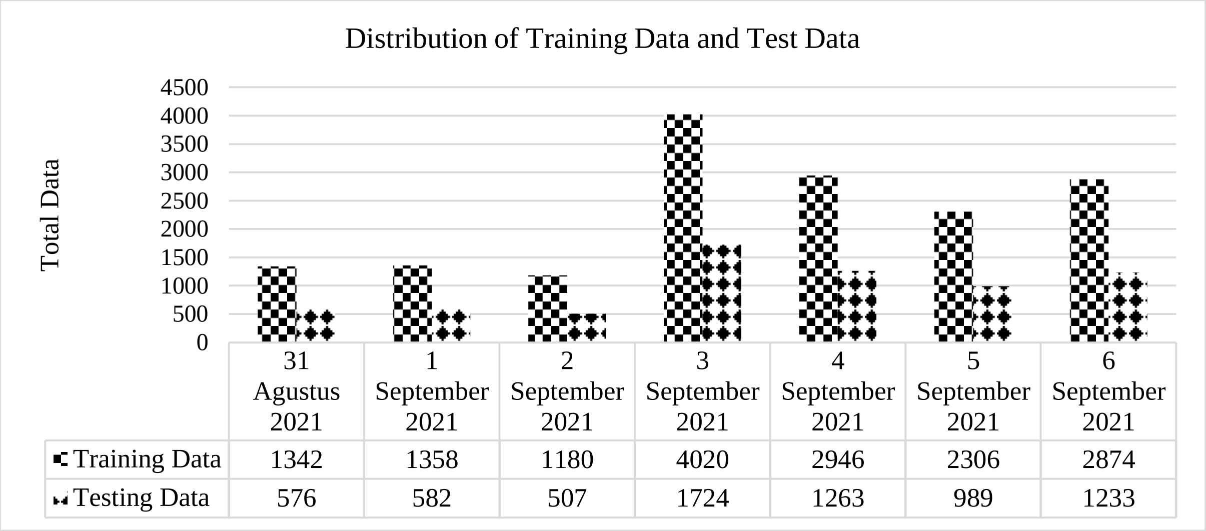
After that, a weighting process is carried out where each word is given a weight using Term Frequency and Inverse Document Frequency (TF-IDF), and the results of the weighting process produce features to be used in the classification process, as shown in Table 1.

Table 1. The Features Generated at Each Temporal

No.	Date	Feature
1	31 Agustus 2021	1185
2	1 September 2021	1271
3	2 September 2021	1105
4	3 September 2021	1636
5	4 September 2021	1550
6	5 September 2021	1426
7	6 September 2021	1630

3.2 Naïve Bayes Classifier Model

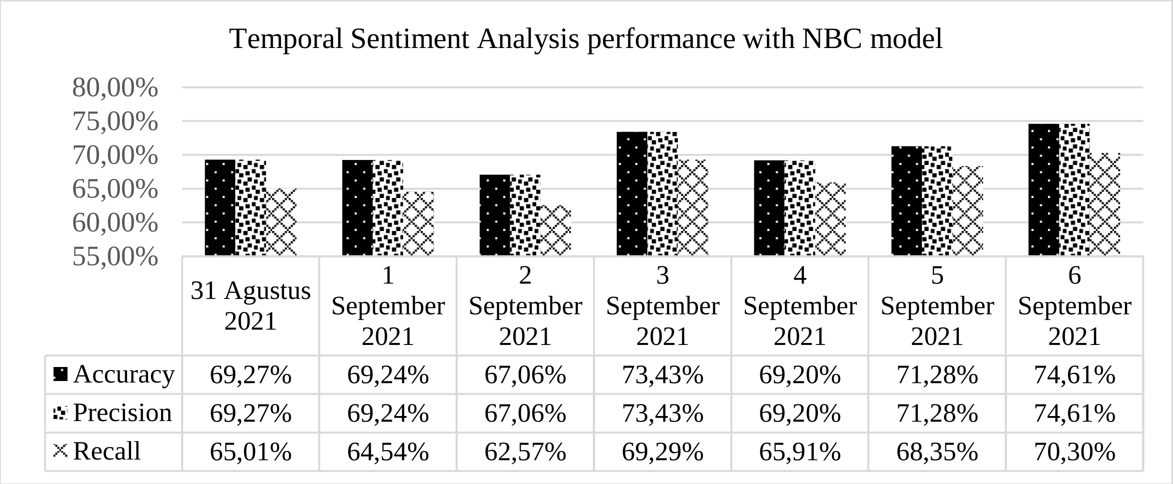
The weighted data is divided into two parts: Training Data and Test Data. This study used the Hold-Out data distribution technique, dividing the data into 70% training and 30% test data, resulting in the following data distribution, as shown in Figure 4.



	31 Agustus 2021	1 September 2021	2 September 2021	3 September 2021	4 September 2021	5 September 2021	6 September 2021
▩ Training Data	1342	1358	1180	4020	2946	2306	2874
◆ Testing Data	576	582	507	1724	1263	989	1233

Figure 4. Distribution of Training Data and Test Data

The data training process used the training data to create a model with the NBC algorithm, where data modeling was performed based on temporal data. The modeling results with NBC were then used to perform testing with the test data, resulting in a confusion matrix evaluation in Figure 5.

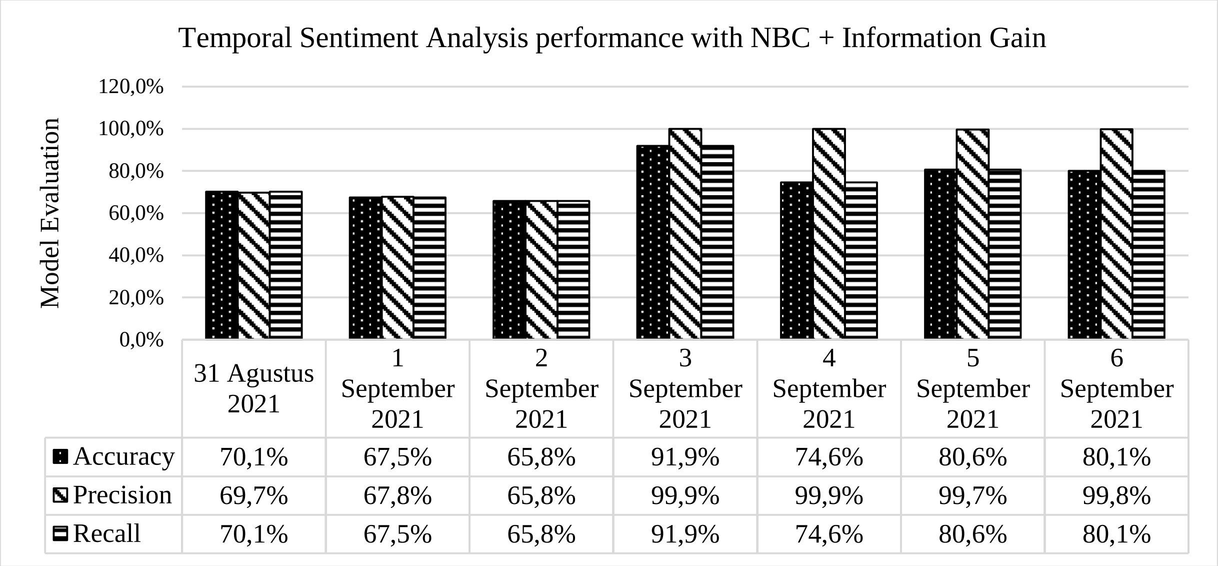


	31 Agustus 2021	1 September 2021	2 September 2021	3 September 2021	4 September 2021	5 September 2021	6 September 2021
■ Accuracy	69,27%	69,24%	67,06%	73,43%	69,20%	71,28%	74,61%
⁘ Precision	69,27%	69,24%	67,06%	73,43%	69,20%	71,28%	74,61%
× Recall	65,01%	64,54%	62,57%	69,29%	65,91%	68,35%	70,30%

Figure 5. Evaluation of Temporal Sentiment Analysis with the NBC Model

Based on the results of sentiment classification testing from user review data of the Pedulilindungi application from August 30, 2021, to September 6, 2021, the accuracy was less than optimal, with the best evaluation results obtained in the experiment using temporal data from September 6, 2021, with an accuracy of 74.61%, with precision and recall values of 74.61% and 70.30%, respectively. The lowest evaluation results were obtained from the temporal data on September 2, 2021, with accuracy, precision, and recall values of 67.06%, 67.06%, and 67.06%, respectively.

Less influential features in the data may influence the less-than-optimal accuracy value, so performing feature selection using the Information Gain method with a threshold of 0.0005 to select influential features on the dataset is necessary. The results of classifying user review data from the Pedulilindungi application using the NBC algorithm with Information Gain as the feature selection technique are as follows, as shown in Figure 6.



	31 Agustus 2021	1 September 2021	2 September 2021	3 September 2021	4 September 2021	5 September 2021	6 September 2021
■ Accuracy	70,1%	67,5%	65,8%	91,9%	74,6%	80,6%	80,1%
▧ Precision	69,7%	67,8%	65,8%	99,9%	99,9%	99,7%	99,8%
▤ Recall	70,1%	67,5%	65,8%	91,9%	74,6%	80,6%	80,1%

Figure 6. Evaluation of Temporal Sentiment Analysis with the NBC and Information Gain

Based on the results of the experiment using feature selection with the Information Gain technique in sentiment classification from user review data of the Pedulilindungi application from August 30, 2021, to September 6, 2021, the best evaluation was obtained in the experiment using temporal data from September 3, 2021, with an accuracy of 91.9%, with precision and recall values of 99.9% and 91.9%, respectively. The lowest evaluation results were obtained from the temporal data on September 2, 2021, with accuracy, precision, and recall values of 65.8%, 65.8%, and 65.8%, respectively.

3.3 Sentiment Analysis

Sentiment analysis based on time in review data shows various words appearing at each time point. In the review dated September 3, 2021, which is temporal data with the highest testing accuracy using the Information Gain technique for feature selection, positive reviews are dominated by words such as oke; bagus; mantap; bantu; top; keren; manfaat; lumayan; mudah; aplikasi; sertifikat; and others.

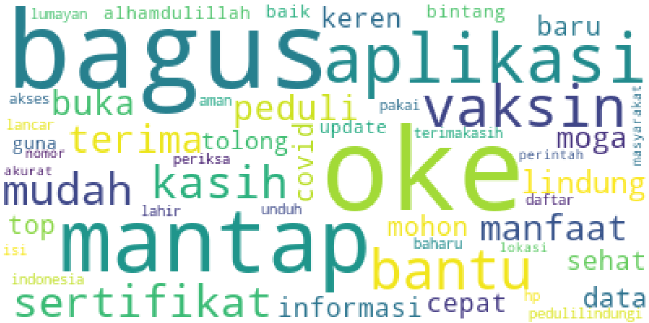


Figure 7. Wordcloud of Positive Sentiment in Reviews on September 3, 2021

Meanwhile, the words that dominated positive reviews on September 2, 2021, temporal data with the lowest testing accuracy, were oke; bagus; mantap; bantu; terima kasih; aplikasi bagus; moga manfaat; mudah; akses; lumayan; and others.

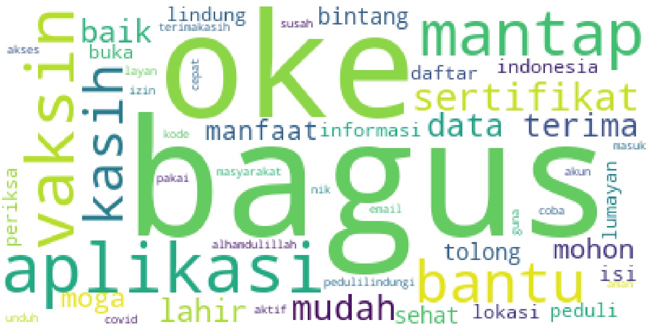


Figure 8. Wordcloud of Positive Sentiment Reviews on September 2, 2021

Based on Figure 9, the words that dominated the negative sentiment for reviews on September 3, 2021, are aplikasi; sertifikat; vaksin; error; ribet; lambat; susah; buka; payah; jelek; and others. Conversely, in Figure 10, the September 2, 2021 reviews are dominated by the words aplikasi; vaksin; sertifikat; ribet; buruk; error; otp; jelek; lambat; bug; and so on.

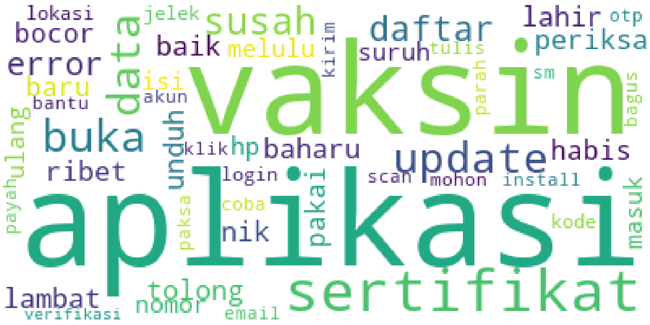


Figure 9. Wordcloud of Negative Sentiment in Reviews on September 3, 2021

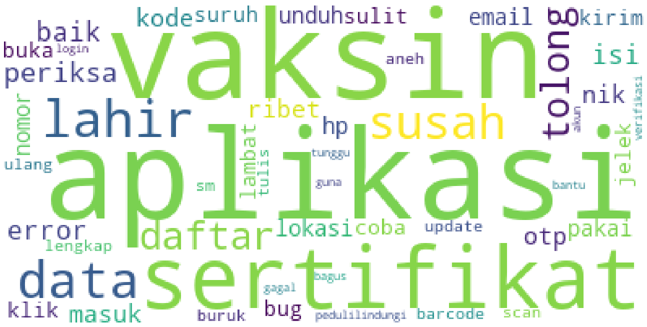


Figure 10. Wordcloud of Negative Sentiment in Reviews on September 2, 2021

Based on Figure 11, the words that dominate the neutral sentiment in reviews on September 3, 2021, are aplikasi; vaksin; good; bantu; sertifikat; update; mantul; buka; baru; unduh; and others.

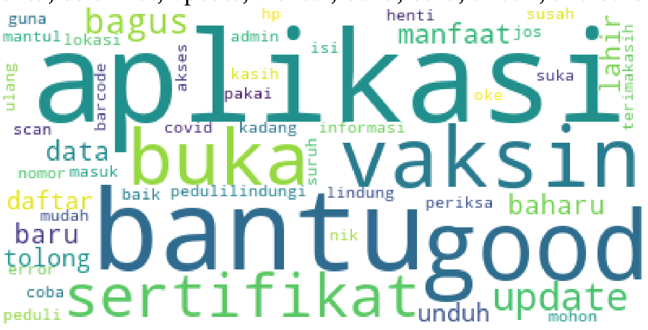


Figure 11. Wordcloud of Neutral Sentiment Reviews on September 3, 2021

Meanwhile, in Figure 12, reviews on September 2, 2021, are dominated by aplikasi; buka; bantu; vaksin; good; bagus; sertifikat; manfaat; error; lahir; and others.

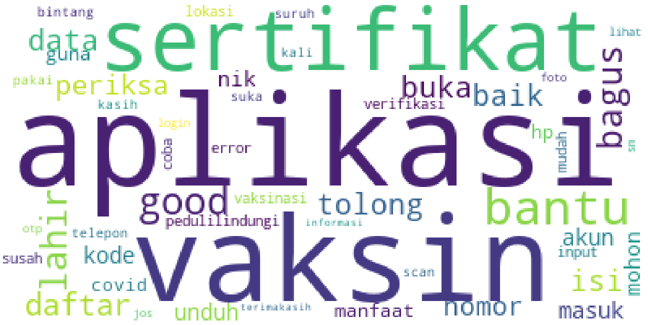


Figure 12. Wordcloud of Neutral Sentiment Reviews on September 2, 2021

4. CONCLUSION

Based on the results of this study, feature selection using the Information Gain technique can improve the accuracy of sentiment classification in user review data for the Pedulilindungi application using the Naïve Bayes Classifier (NBC) algorithm, where the best evaluation was obtained on temporal data for September 3, 2021, which applied the Information Gain technique for feature selection at a threshold of 0.0005, yielding an accuracy of 91.9%, with precision and recall values of 99.9% and 91.9%, respectively. The dominant words appearing in positive sentiment on that date were oke; bagus; mantap; bantu; top; keren; manfaat; lumayan; mudah; aplikasi; sertifikat; and others. The words that dominated the negative sentiment for reviews on September 3, 2021, were aplikasi; sertifikat; vaksin; error; ribet; lambat; susah; buka; payah; jelek; and others. Meanwhile, the words that dominated the neutral sentiment were aplikasi; vaksin; good; bantu; sertifikat; update; mantul; buka; baru; unduh; and others.

REFERENCES

[1] World Health Organization WHO, “WHO Coronavirus (COVID-19) Dashboard,” WHO (World Health Organization). [Online]. Available: https://covid19.who.int/region/searo/country/id

[2] F. N. Afiana, I. R. Yunita, L. D. Oktaviana, and U. Hasanah, “Pelatihan Teknis Penggunaan Aplikasi PeduliLindungi Guna Melacak Penyebaran COVID-19,” *JPMM (Jurnal Pengabdi. Mitra Masyarakat)*, vol. 2, no. 2, pp. 98–106, 2020.

[3] KOMINFO, “Keputusan Menteri Komunikasi dan Informatika Nomor 171 Tahun 2020 tentang Penetapan Aplikasi Pedulilindungi Dalam Rangka Pelaksanaan Surveilans Kesehatan Penanganan Corona Virus Disease 2019 (Covid-19),” 2020. [Online]. Available: https://jdih.kominfo.go.id/produk_hukum/view/id/735/t/keputusan+menteri+komunikasi+dan+informatika+nomor+171+tahun+2020

[4] Menteri Dalam Negeri Republlik Indonesia, “Instruksi Menteri Dalam Negeri Republik Indonesia Nomor 38 Tahun 2021 Tentang Pemberlakukan Pembatasan Kegiatan Masyarakat (PPKM) Level 4, Level 3, dan Level 2 Corona Virus Disease 2019 di Wilayah Jawa dan Bali,” Jakarta, 2021.

[5] N. Herlinawati, Y. Yuliani, S. Faizah, W. Gata, and S. Samudi, “Analisis Sentimen Zoom Cloud Meetings di Play Store Menggunakan Naïve Bayes dan Support Vector Machine,” *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 2, p. 293, 2020, doi: 10.24114/cess.v5i2.18186.

[6] Suwanda Aditya Saputra, D. Rosiyadi, W. Gata, and S. M. Husain, “Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization Suwanda,” *Resti*, vol. 3, no. 3, pp. 377–382, 2019.

[7] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” *Integer J.*, vol. 2, no. 1, pp. 32–41, 2017, [Online]. Available: https://t.co/jrvaMsgBdH

[8] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” 2017, [Online]. Available: http://arxiv.org/abs/1707.02919

[9] S. Fransiska, R. Rianto, and A. I. Gufroni, “Sentiment Analysis Provider By.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method,” *Sci. J. Informatics*, vol. 7, no. 2, pp. 203–212, 2020, [Online]. Available: https://journal.unnes.ac.id/nju/sji/article/view/25596

[10] S. Wahyu Handani, D. Intan Surya Saputra, Hasirun, R. Mega Arino, and G. Fiza Asyrofi Ramadhan, “Sentiment analysis for go-jek on google play store,” *J. Phys. Conf. Ser.*, vol. 1196, no. 1, 2019, doi: 10.1088/1742-6596/1196/1/012032.

[11] R. Wahyudi and G. Kusumawardana, “Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine,” *J. Inform.*, vol. 8, no. 2, pp. 200–207, 2021, doi: 10.31294/ji.v8i2.9681.

[12] A. Hendra and F. Fitriyani, “Analisis Sentimen Review Halodoc Menggunakan Nai ̈ve Bayes Classifier,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 2, pp. 78–89, 2021, doi: 10.14421/jiska.2021.6.2.78-89.

[13] Mustakim *et al.*, “Data Sharing Technique Modeling for Naive Bayes Classifier for Eligibility Classification of Recipient Students in the Smart Indonesia Program,” *J. Phys. Conf. Ser.*, vol. 1424, no. 1, 2019, doi: 10.1088/1742-6596/1424/1/012009.

[14] S. G. Setyorini and Mustakim, “Application of the nearest neighbor algorithm for classification of online taxibike sentiments in indonesia in the google playstore application,” *J. Phys. Conf. Ser.*, vol. 2049, no. 1, 2021, doi: 10.1088/1742-6596/2049/1/012026.

[15] S. Hassan, M. Rafi, and M. S. Shaikh, “Comparing SVM and Naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment,” *Proc. 14th IEEE Int. Multitopic Conf. 2011, INMIC 2011*, pp. 31–34, 2011, doi: 10.1109/INMIC.2011.6151495.

[16] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016, doi: 10.5815/ijieeb.2016.04.07.

[17] Dinda Ayu Muthia, “ANALISIS SENTIMEN PADA REVIEW BUKU MENGGUNAKAN ALGORITMA NAÏVE BAYES,” *J. Paradig.*, vol. XVI, no. 1, p. 12, 2014.

[18] I. Maulida, A. Suyatno, and H. R. Hatta, “Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain,” *J. SIFO Mikroskil*, vol. 17, no. 2, pp. 249–258, 2016, doi: 10.55601/jsm.v17i2.379.

[19] Ivan, Y. A. Sari, and P. P. Adikara, “Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 5, pp. 4914–4922, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[20] C. E. Putri and R. E. Hamzah, “Aplikasi Pedulilindungi Mitigasi Bencana Covid-19 Di Indonesia,” *J. Pustaka Komun.*, vol. 4, no. 1, pp. 66–78, 2021, doi: 10.32509/pustakom.v4i1.1321.

[21] J. Ma, W. Xu, Y. Sun, E. Turban, S. Wang, and O. Liu, “An ontology-based text-mining method to cluster proposals for research project selection,” *IEEE Trans. Syst. man, Cybern. a Syst. humans*, vol. 42, no. 3, pp. 784–790, 2012.

[22] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[23] M. W. Berry and J. Kogan, *Text mining: applications and theory*. John Wiley \& Sons, 2010.

[24] N. M. S. Hadna, I. S. Paulus, and W. Winarno, “Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter,” *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. March, pp. 1–8, 2016.

[25] W. Gata, “Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS,” vol. 6, pp. 1–13, 2017.

[26] C. D. M. P. Raghavan and H. Schütze, “An Introduction to Information Retrieval,” *Cambridge Univ. Press*, pp. 1–519, 2009, doi: 10.1210/endo-38-3-156.

[27] F. Tala, “A study of stemming effects on information retrieval in Bahasa Indonesia,” 2003.

[28] S. Andini, “Klasifikasi Dokument Teks Menggunakan Algoritma Naive Bayes dengan Bahasa Pemograman Java,” *J. Teknol. Inf. \& Pendidik.*, vol. 6, no. 2, pp. 140–147, 2013.

[29] Firdaus Mahmudy and Wahyu Widodo, “Klasifikasi Artikel Berita Secara Otomatis Menggunakan,” *Tekno*.

[30] F. Ariani, Amir, N. Alam, and K. Rizal, “Klasifikasi Penetapan Status Karyawan Dengan MenggunakanMetode Naïve Bayes,” *Paradigma*, vol. XX, no. 2, pp. 33–38, 2018, doi: 10.31294/p.v.

[31] M. A. Thanoon, M. J. M. Zedan, and A. N. Hameed, “Feature selection based on wrapper and information gain,” in *2019 1st AL-Noor International Conference for Science and Technology (NICST)*, 2019, pp. 32–37.

[32] N. A. Shaltout, M. El-Hefnawi, A. Rafea, and A. Moustafa, “Information gain as a feature selection method for the efficient classification of influenza based on viral hosts,” *Lect. Notes Eng. Comput. Sci.*, vol. 1, pp. 625–631, 2014.

[33] X. Wang, M. Zuo, and L. Song, “A feature selection method based on information gain and BP neural network,” in *Chinese Intelligent Systems Conference*, 2017, pp. 23–30.