Information Technology and Systems
e-ISSN: 3031-1187
Homepage: https://sanscientific.com/journal/index.php/its
3(1) 14-26 (2026)
https://doi.org/10.58777/its.v3i1.500

Performance Analysis of Ensemble Learning in Sentiment Classification of BRImo App Reviews

Novi Puspita Sari [1*]
[1]Program Studi S1 Teknologi Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jakarta

Received: October 10, 2025; Accepted: November 14, 2025

Abstract
The use of mobile banking services in Indonesia continues to increase along with the development of information technology, including the BRImo application owned by Bank Rakyat Indonesia (BRI), which has reached more than 50 million downloads and one million reviews on the Google Play Store. These reviews serve as an important data source for understanding user perceptions and experiences. This study analyzes the performance of the Ensemble Learning method for sentiment classification of BRImo reviews by combining Support Vector Machine (SVM) and Decision Tree. The data was obtained through web scraping techniques, then processed through preprocessing stages including cleaning, case folding, normalization, tokenization, stopword removal, and stemming. Next, a lexicon approach was used for sentiment labeling, while TF-IDF was used for feature extraction. The dataset consists of 8,002 reviews, split with a ratio of 80:20. The study results show that SVM achieved the highest accuracy at 92.63%, due to its strong ability to optimally separate high-dimensional text data. The Ensemble model combining SVM and Decision Tree achieved an accuracy of 89.38%, slightly lower than SVM, but still providing stable predictions. This is because the Ensemble leverages the strength of two algorithms, making it capable of reducing result variance. Meanwhile, the Decision Tree recorded the lowest accuracy at 86.45%, indicating its limitations in handling the complexity of text data. Thus, although the Ensemble does not surpass SVM, the model combination still produces a more balanced and consistent performance. This study has limitations in terms of data coverage and a lexicon approach that is sensitive to context. The findings have implications for the development of the BRImo application based on user perceptions. The novelty of the research lies in the application of the SVM–Decision Tree Ensemble in sentiment analysis of mobile banking applications in Indonesia.

Keywords: Specimen Analysis; BRImo; Ensemble Learning; SVM; Decision Tree

How to cite: Novi Puspita Sari. (2025). Performance Analysis of Ensemble Learning in Sentiment Classification of BRImo App Reviews (ITS) 3(1), 14-26.

*Corresponding author: Novi Puspita Sari (npuspitasari326@gmail.com)

1. Introduction

One of the most popular banking innovations is mobile banking. Undoubtedly, modern advancements affect many aspects of life, including how customers interact with financial services. A smartphone can be used to access these banking services directly, for all kinds of transactions, especially money transfers, balance checks, bill payments, and investments.(Marwi, 2024). In addition, this service also provides convenience, speed, and efficiency. BRImo is the mobile banking service of Bank Rakyat Indonesia (BRI) (Trisnaningrum et al., 2024). As the number of users and the amount of reviews on the BRImo application increase, it becomes a new challenge to continuously improve the quality of services to enhance customer satisfaction, by leveraging user perceptions obtained from reviews on the BRImo application.

To understand the sentiment or perception of these users, sentiment analysis becomes an effective method. Based on previous research, the study conducted by (Lowell et al., 2025) comparing SVM and Naïve Bayes for BRImo reviews and found that SVM excelled in training accuracy, but the model's performance on test data was still moderate (83.11%). Most studies compare single algorithms such as SVM, Naïve Bayes, or transformer models, but the application of an ensemble combining two models is still rarely explicitly researched in the context of BRImo. Comparing the clear performance among classic models, namely SVM, Decision Tree, and Ensemble, tested on the same dataset (8,002 reviews), allows for fair and measurable performance comparison. The dataset size is relatively larger and more representative compared to smaller studies using only 199 reviews, making the results statistically more robust. The implementation of a complete preprocessing pipeline, from cleaning to TF-IDF, ensures better input quality for classification. Ensemble learning is an approach in machine learning to improve prediction accuracy using multiple predictive models; in this study, a hard voting ensemble type is used to combine predictions from two algorithm models, namely SVM and Decision Tree. (Sondakh, S.Kom, M.T, Ph.D et al., 2023).

2. Literature Review

Sentiment Analysis

Sentiment analysis is a technique used to identify and classify public opinions into positive, negative, or neutral categories based on text (L. A. Fudholi et al., 2024). In the context of digital applications, sentiment analysis is widely used to understand user perceptions through reviews on platforms such as the Google Play Store (Nurwahidah et al., 2023). This method helps companies evaluate user satisfaction levels and improve service quality.

Machine Learning for Text Classification

Machine learning has become the main approach in sentiment classification. Two widely used algorithms are:

1) Support Vector Machine (SVM)

   SVM is a classification algorithm that works by finding the optimal hyperplane to separate classes (Rusman et al., 2023). In text analysis, SVM often delivers high performance because it can handle high-dimensional data such as TF-IDF (Aryanti & Suria, 2025).

2) Decision Tree

   Decision Tree is a tree-based classification method with decision rules. Its advantages lie in its interpretability and its ability to model non-linear relationships (Kurniawati, 2024). However, this model tends to overfit on complex data.

Ensemble Learning

Ensemble Learning is a method of combining several base algorithms (base learners) to obtain more stable and accurate prediction results (Rayadin et al., 2024). Ensemble techniques such as boosting, bagging, and majority voting have been proven to improve model performance in various classification tasks (Rachmatullah, 2025). Majority voting is a simple ensemble method that works by selecting the most frequent prediction from several models.

Lexicon-Based Approach in Sentiment Labeling
The lexicon approach uses a list of words that have been assigned polarity values to determine the sentiment of a text. Lexicon-based methods are often used in research that requires automatic labeling on a large scale because they are efficient and do not require human annotators (Rizkia et al., 2025). Indonesian language lexicons such as INA SentiWord or other polarity dictionaries are commonly used in sentiment analysis research in Indonesia.

Previous Research Related to Sentiment of the BRImo Application
Several previous studies have examined sentiment analysis of the BRImo application, among others, Research (Iffa et al., 2025) using Naïve Bayes, SVM, and Logistic Regression models on 1,500–3,000 BRImo reviews. The results show that SVM has the best performance, but the dataset used is limited and most of the labeling was done manually. Research (D. H. Fudholi, 2022) applying deep learning methods such as LSTM and IndoBERT on small datasets (≤2,000 reviews). Deep learning models provide high accuracy, but require large computational resources and do not incorporate ensemble methods. Research (Hermawan et al., 2025) using a simple lexicon approach to assess BRImo sentiment descriptively without involving a machine learning model, so the results do not focus on algorithm performance.

3. Methods

Research Stages
The research stages conducted in this study are shown in Figure 1 below:

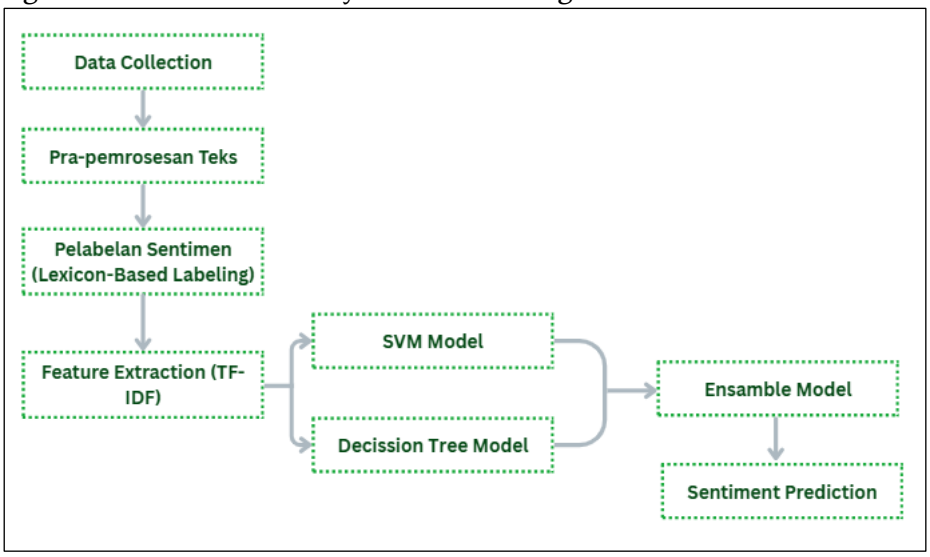

Figure 1. Research Stages

1. **Pengumpulan Data**
   The initial stage of the research is to collect review data of the BRImo application from the Google Play Store using web scraping techniques. This process resulted in 8,002 reviews containing users' opinions regarding their experience using the BRImo application. Reviews can consist of long or short sentences, formal or informal language, a mix of Indonesian and English, and even emoticons. This raw data serves as the foundation for the entire sentiment analysis process that will be carried out in the next stage.

2. **Text Preprocessing**
   The preprocessing stage is carried out to clean and standardize the text so that it can be properly processed by machine learning algorithms. This process includes:
   1) Cleaning
      Removing symbols, excessive punctuation, URLs, emoticons, numbers, and unnecessary characters. Cleaning aims to eliminate noise in the data.
   2) Tokenizing
      Splitting sentences into word pieces (tokens). For example, "BRImo sering error" to ["brimo",

"sering", "error"].

3) Stopwords Removal

Removing common words that do not have significant meaning in sentiment analysis, such as 'that', 'and', 'in', 'to'.

4) Stemming

Change the word to its root form. For example: “mengirimkan” becomes “kirim”.

Preprocessing is very important because the quality of the input greatly affects the model's output.

3. Pelabelan

In this study, sentiment labeling was carried out using an Indonesian lexicon approach, which is a method that relies on a list of words that have been assigned polarity values (positive, negative, or neutral). Each word in the lexicon has a certain score representing its emotional tendency. The labeling process is done by summing the polarities of the words in each review and then determining the sentiment based on the total value obtained:

1) positive value → positive sentiment
2) negative value → negative sentiment
3) value approaching 0 → neutral sentiment

4. Feature Extraction (TF-IDF)

After the text data is cleaned, the text is converted into a numerical representation using TF-IDF (Term Frequency–Inverse Document Frequency). TF-IDF gives higher weights to words that frequently appear in a particular document but rarely appear in other documents. This method helps the model distinguish important words from common ones. Examples of high-value words are error, failed, loading, great, easy, fast.

5. Machine Learning Model Training

At this stage, the feature extraction results are used to train several classification models.

1) SVM (Support Vector Machine) Model

SVM builds the best separating line (hyperplane) to differentiate between positive, negative, and neutral sentiment classes. SVM is known for excelling at handling high-dimensional textual data such as TF-IDF.

2) Decision Tree Model

Decision Trees map data into a tree structure based on decision rules. This model is easy to understand and can capture non-linear relationships, but it is less stable on large datasets and prone to overfitting.

These models were trained using an 80% training data and 20% testing data ratio.

6. Ensemble Learning (Majority Voting)

This stage combines predictions from the two previous models using the majority voting technique. The way this technique works starts with the SVM giving a positive prediction, the Decision Tree giving a negative prediction, and then the Ensemble chooses the prediction based on the most votes. If the voting result is a tie, the system will follow the model with the best accuracy, which is the SVM model. The Ensemble Model aims to increase prediction stability, reduce errors from a single model, and achieve a more balanced final prediction. The accuracy of the Ensemble model is lower than that of the SVM model, but the Ensemble model can provide more consistent predictions across data variations.

7. Sentiment Prediction

The sentiment prediction stage produces sentiment outputs for each review, namely positive, negative, and neutral. This prediction is influenced by the majority voting results from the SVM and Decision Tree models. The sentiment output can then be further analyzed to understand users' perceptions of the BRImo application.

Data Collection Techniques

Data collection in this study was carried out through the process of collecting user review data of the

BRImo application by scraping data using a Python library, namely Google Play Scraper, by connecting the AppId on the Google Play Store for the BRImo application with a total of 10,256 datasets in CSV format. Data was collected from March 12, 2019, to June 9, 2025. In addition, data collection was carried out through a literature review of journal articles, scholarly works, and research reports on sentiment analysis, ensemble learning, and BRImo mobile banking users, adapted to support the data collection system. The scientific basis and arguments in the analysis and discussion of the study results were reinforced by this literature review.

4. Results

1. Data Selection

Review data of the BRImo application successfully collected using scraping techniques with the Google Play Scraper library integrated in Google Colab amounted to 10,256 datasets, consisting of 5,000 most relevant reviews and 5,000 latest reviews.

	userName	content	score	at
2	Dodi Setiawan	susah banget bikin username nya	1	6/9/2025 8:03
3	Renuard Rajagukguk	aplikasi bank yg terpercaya mudahh dan g	1	6/9/2025 8:01
4	Agung Prasetyo	Sulit sekali buka rekening melalui brimo,	1	6/9/2025 7:59
5	raden agan	Apk tolol bikin username aja susah	1	6/9/2025 7:49
6	Andi Safaat	ni aplikasi gmana sih,semua verifikasi dal	1	6/9/2025 7:46
7	Agam Abdillah	Nice	5	6/9/2025 7:41
8	Elena Skin3	cepat transaksi nya mantap....	5	6/9/2025 7:40
9	Melody Afrilliani	Saya kecewa dengan BRIMO karena lupa p	3	6/9/2025 7:39
10	Adi Satria Wibowo	bikin username selalu tidak bisa, sudah b	1	6/9/2025 7:38
11	Rindra Nuari Wijayanto	Mohon dibantu ini kenapa pembuatan us	3	6/9/2025 7:33
12	Jho Hari	apk apa ini cuma buat username saja di p	1	6/9/2025 7:27
13	Kawi Metalcore	mantap, gesit, sangat bisa diandalkan ðŸ‘	5	6/9/2025 7:25
14	Miranda Akib	jadi lebih mudah	5	6/9/2025 7:22
15	Dyah Ratna Sari	sering banget terjadi gangguan disaat but	3	6/9/2025 7:16

Figure 2. Results of Review Data Scraping

In Figure 2, the results of data scraping are selected into 4 attributes, namely UserName, content (review), score (rating), and at (date/time). Duplicate data was found, so the data was deleted, resulting in 8,054 data ready to be processed in the next stage.

2. Text Preprocessing

This preprocessing stage is the stage where review data is cleaned before the classification process. This stage consists of Cleaning, Case Folding, Normalization, Tokenizing, Stopword Removal, and Stemming.

The results of preprocessing are shown in the following Table 1:

Table 1. Text Preprocessing Results

Review Data	very good **ðŸ‘□ðŸ‘□ðŸ‘□ðŸ‘□ðŸ‘**
Cleaning Results	very good
Data Ulasan	GiLa makes creating a username really difficult, the username is always unavailable
Case Folding Result	It's crazy how hard it is to make a username, the username is always taken.
Review Data	I created a Brimo account but the username is not available... please help me with a solution
Normalize Result	I created a Brimo account but the username is not available... please give me a solution
Review Data	The transaction was fast, excellent....
Tokenizing Result	'fast', 'transaction', 'yes', 'great....'
Review Data	very fast and reliable
Stopword Result	'fast', 'reliable'
Review Data	What kind of app is this, just for a username and it's made so complicated?
Stemming Result	difficult username app

3. Pelabelan

After the Cleansing stage, the next step in this sentiment analysis is data labeling. Data labeling is done based on reviews from BRImo application users, identified using an Indonesian lexicon dictionary divided into three classes: positive, negative, and neutral.

Table 2. Labeling Results

score	review	sentiment
1	It's really hard to make a username, isn't it?	Negative
1	Trusted bank app, easy and simple	Positive
3	Disappointed that Brimo forgot the account password, banned, returning the account is a complicated process, simple bank is easy, needs Gmail number verification, valid data, old bank account recovery updates to keep up with the times.	Neutral

The score table represents review ratings, the review table contains review results that have gone through the preprocessing stage, and the sentiment table is a sentiment classification table based on positive, negative, and neutral reviews.

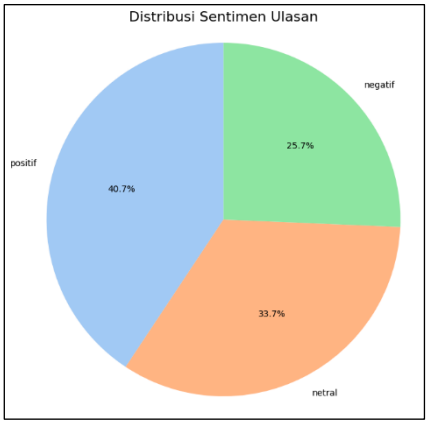

Figure 3. Pie Chart of Labeling Results

Figure 3 shows data of 10,256 obtained from scraping user reviews of the BRImo application, which became 8,002 data points after the cleansing process. There are 40.7% of reviews with positive sentiment, 25.7% of reviews with negative sentiment, and 33.7% of reviews with neutral sentiment.

4. Transformation

At this transformation stage, testing of the model used for sentiment analysis of the BRImo application is carried out. The data is divided into two types: training data and test data with an 80:20 ratio as shown in Table 3. The review data must be converted into a numerical form so that it can be processed in the modeling stage.

Table 3. Results of Training Data and Testing Data Distribution

Comparison Ratio	Number of Training Data	Number of Test Data
80:20	6401	1601

5. Data Mining

At the data mining stage, after the data is divided into training and testing data and converted into numerical form, the next step is the model algorithm classification stage, which is divided into three classification processes as follows:

1) SVM *(Support Vector Machine)*

The SVM classification model used to find the best hyperplane to separate classes in this sentiment analysis, namely positive, negative, and neutral classes, uses a linear kernel. Based on the training results in Figure 5, the use of training data and testing with test data is as follows:

```
SVM Accuracy: 0.9262960649594004

SVM Accuracy: 92.63%

SVM Classification Report:
              precision    recall  f1-score   support

     negatif       0.94      0.90      0.92       389
      netral       0.87      0.93      0.90       565
     positif       0.97      0.94      0.96       647

    accuracy                           0.93      1601
   macro avg       0.93      0.92      0.92      1601
weighted avg       0.93      0.93      0.93      1601
```

Figure 5. SVM Model Calculation Results

The findings indicate that the Support Vector Machine (SVM) model has an accuracy rate of 92.63%.

Table 4. SVM Classification Report

Class	Precision	Recall	F1-Score
Negative	0.94	0.90	0.92
Neutral	0.87	0.93	0.90
Positive	0.97	0.94	0.96

Table 4 shows that the positive class has higher precision, recall, and F1-score values compared to the other classes (Negative and Neutral).

2) Decision Tree

Decision Tree classification is processed using the Python library scikit-learn.tree for DecisionTreeClassifier in Google Colab. Based on the process in Figure 6, the results obtained are as follows:

```
Decision Tree Accuracy: 0.8644597126795752

Decision Tree Accuracy: 86.45%

Decision Tree Classification Report:
              precision    recall  f1-score   support

     negatif       0.83      0.85      0.84       389
      netral       0.86      0.84      0.85       565
     positif       0.89      0.89      0.89       647

    accuracy                           0.86      1601
   macro avg       0.86      0.86      0.86      1601
weighted avg       0.86      0.86      0.86      1601
```

Figure 6. Decision Tree Model Calculation Results

The results show that the Decision Tree model has an accuracy rate of **86.45%.**

Table 5. Decision Tree Classification Report

Class	Precision	Recall	F1-Score

Negative	0.83	0.85	0.84
Neutral	0.86	0.84	0.85
Positive	0.89	0.89	0.89

From the Classification report shown in Table 5, it can be seen that the highest precision, recall, and F1-score values are in the positive class, just like the SVM classification results. The positive class is more dominant compared to the negative and neutral classes.

3) Ensemble Learning (SVM + Decision Tree)

Model integration was carried out using the Hard Voting Classifier technique. This process was conducted in Google Colab using Python libraries from the scikit-learn.ensemble library. From this ensemble process, the results obtained are as shown in Figure 7 below:

```
Accuracy Ensemble (Hard Voting): 89.38%

Classification Report:
              precision    recall  f1-score   support

     negatif       0.82      0.98      0.89       389
      netral       0.87      0.88      0.88       565
     positif       0.99      0.85      0.91       647

    accuracy                           0.89      1601
   macro avg       0.89      0.90      0.89      1601
weighted avg       0.90      0.89      0.89      1601
```

Figure 7. Results of Ensemble Learning Model Calculations

Based on the calculation results, the Ensemble model's accuracy level is 89.38%.

Table 6. Ensemble Learning Classification Report

Class	Precision	Recall	F1-Score
Negative	0.82	0.98	0.89
Nuetral	0.87	0.88	0.88
Positive	0.99	0.86	0.91

It can be seen that the highest precision and F1-score values are in the positive class, while recall is in the negative class. This still indicates that positive reviews are more dominant compared to the other classes.

6. Evaluation

1) Confusion Matrix SVM

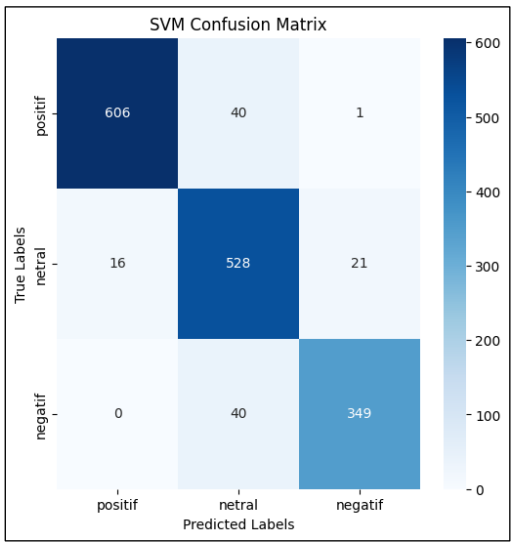

Figure 8. SVM Confusion Matrix

Positive Class = there are 606 positive data correctly predicted as belonging to the positive

sentiment class True Positive (TP), and 16 data predicted as neutral sentiment and 0 predicted as negative sentiment False Positive (FP), as well as 40 data predicted as neutral sentiment and 1 as negative, so there are 41 data that are actually positive but misclassified False Negative (FN).
Negative Class = there are 349 negative data that are correctly predicted as True Negative (TN). Furthermore, there is 1 data predicted as negative but actually positive, and 21 actually neutral but predicted as negative. Therefore, the False Positive (FP) amounts to 22 data. Then, in the neutral prediction column, there are 40 data predicted as neutral but actually negative, which are False Negatives (FN).
Neutral Class = there are 528 neutral data correctly predicted as True Positive (TP) Neutral. And 40 data were classified as positive but predicted as neutral, 40 data were classified as negative but predicted as neutral, so the False Positive (FP) totals 80 data. Meanwhile, 16 data were predicted as positive but should be neutral, and 21 data were predicted as negative but should be neutral, making the False Negative (FN) a total of 37 data.

2) Confusion Matrix Decision Tree

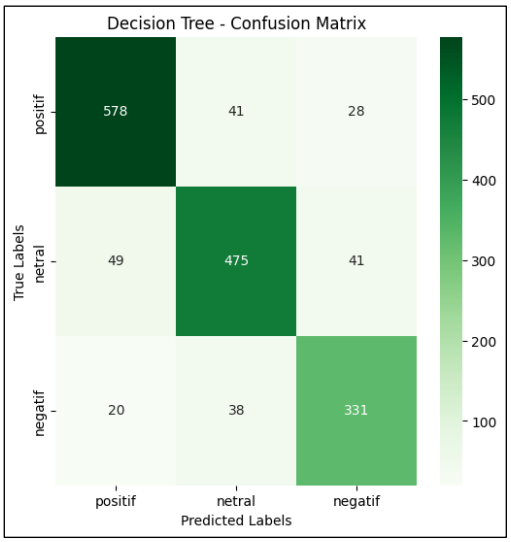

Figure 9. *Decision Tree Confusion Matrix*

Positive Class = there are 578 data correctly predicted as positive True Positive (TP), 49 data actually neutral predicted as positive, and 20 negative data predicted as positive False Positive (FP), totaling 69. Meanwhile, 41 data predicted as neutral that are actually positive and 28 predicted as negative that are actually positive, so the False Negative (FN) totals 69.
Negative Class = there are 331 data correctly predicted as negative review data True Positive (TP) Negative. As many as 28 data actually positive were predicted as negative and 41 data actually neutral were predicted as negative, so False Positive (FP) amounts to 69 data. Meanwhile, 20 data were predicted as positive but actually negative and 38 data actually negative were predicted as neutral, resulting in False Negative (FN) of 58.
Kelas Netral = terdapat 475 data benar diprediksi netral True Positif(TP) Netral. Terdapat 41 data sebenarnya positif terpediksi netral dan 38 data sebenarnya negatif terprediksi netral, maka False Positif (FP) berjumlah 79 data. Selanjutnya terdapat 49 terprediksi positif yang sebenarnya netral dan 41 data terprediksi negatif sebenarnya netral, maka jumlah False Negatif(FN) 90 data.

3) Confusion Matrix Ensemble Learning

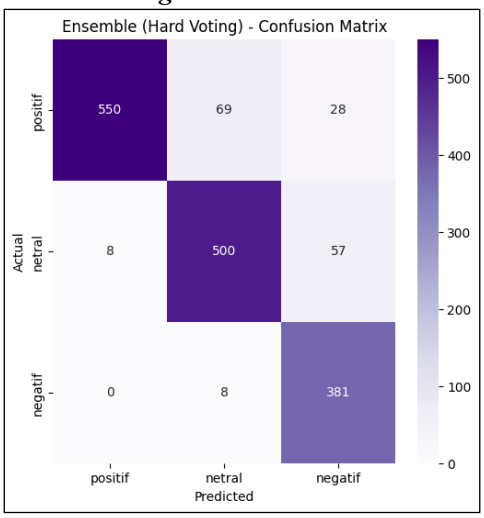

Figure 10. *Ensemble Confusion Matrix*

Kelas Positif = terdapat 550 data benar yang diprediksi sebagai ulasan positif True Positif(TP), sebanyak 8 data yang sebenarnya sebagai ulasan netral diprediksi positif False Positif(FP). Dan terdapat 69 data diprediksi sebagai data ulasan netral dan 28 data diprediksi negatif yang sebenarnya adalah data ulasan positif False Negatif(FN) sebanyak 97 data.
Negative Class = there are 381 data points correctly predicted as negative reviews True Positive (TP). A total of 28 data points are predicted as negative but are actually positive reviews, and 57 data points are predicted as negative but are actually neutral reviews, so the False Negative (FN) amounts to 85. Furthermore, 8 data points are predicted as neutral reviews but are actually negative.
Neutral Class = there are 500 review data predicted correctly as neutral reviews, True Positives (TP) neutral. Then, 69 data are predicted as neutral but actually positive, and 8 data are also predicted as neutral reviews but actually negative, so the False Positives (FP) amount to 77 data. Additionally, 8 review data are predicted as positive but actually neutral, and 57 data are predicted as negative but actually neutral, making the total False Negatives (FN) 65 data.

Table 7. Recapitulation of Analysis Results

Model	Accuracy	Precision	Recall	F1-Score
SVM (Support Vector Machine)	92.63%	93%	92.27%	92.27%
Decision Tree	86.45%	86%	86%	86%
Ensemble (SVM+Decision Tree)	89.38%	89%	90%	89.4%

Based on evaluation Table 7, SVM showed the highest performance with an accuracy of 92.63%, followed by Ensemble at 89.38% and Decision Tree at 86.45%. The precision, recall, and F1-score values also confirm the dominance of SVM in consistently classifying sentiment. Although the ensemble did not surpass SVM, this model was able to produce more stable predictions compared to the Decision Tree through a majority voting mechanism. Compared to previous studies that used smaller datasets and no ensemble approach, the results of this study show significant improvements both in terms of accuracy and prediction stability. By utilizing a large dataset and lexicon-based automatic labeling, this research provides a new contribution to a more comprehensive and robust sentiment analysis of BRImo app reviews.

5. Discussion

This discussion elaborates on the interpretation of research results, their relation to theory and previous findings, as well as an analysis of the strengths and limitations of the models tested. The main focus is to evaluate the performance of SVM, Decision Tree, and Ensemble Learning in classifying sentiment in BRImo app reviews based on labeled data using a lexicon-based approach.

Performance of Machine Learning Models on the BRImo Review Dataset
The research results show that SVM achieved the highest accuracy of 92.63%. This is consistent with the characteristics of SVM, which is known to be effective in handling high-dimensional data such as TF-IDF representations. In theory, SVM seeks the optimal hyperplane that maximizes the margin between classes, thus providing strong generalization even when the data contains noise or class imbalance. The stable precision, recall, and F1-score results reinforce that SVM is the most suitable model for the context of text-based sentiment analysis on large datasets.
The Decision Tree, on the other hand, shows the lowest performance with an accuracy of 86.45%. Decision tree models tend to be sensitive to small variations in training data and are at risk of overfitting, especially when dealing with informal text commonly found in Google Play Store reviews. Nevertheless, the Decision Tree still makes an important contribution as one of the base learners in an ensemble, particularly in capturing non-linear patterns that may not be detected by SVM.
The Ensemble Model (SVM + Decision Tree) achieved an accuracy of 89.38%, falling between the two individual models. Although it does not surpass the performance of SVM, the ensemble is able to provide more stable predictions. This demonstrates that combining two algorithms with different characteristics can reduce model variance and minimize certain classification errors. The presence of a voting mechanism makes the prediction results more balanced, especially when an individual model is biased toward a particular class.

Impact of Lexicon-Based Labeling Approach
Labeling using an Indonesian lexicon offers advantages in terms of efficiency and objectivity. With a total of 8,002 reviews, manual labeling is inefficient and prone to annotator subjectivity. The lexicon-based approach allows for consistent labeling, but it has limitations in handling context, irony, or informal language styles.
This limitation can be observed from several classification errors, especially in the Neutral class which tends to be more ambiguous. Nevertheless, the consistently high performance of SVM indicates that the quality of the generated labels is adequate for training a machine learning model.

Comparison with Previous Research
In general, the results of this study show a significant improvement compared to previous research on BRImo sentiment analysis. Ramadan et al. (2025) reporting an accuracy of 83.11% for SVM on a much smaller dataset (±1,500 reviews). The increase in accuracy to 92.63% in this study indicates that a larger dataset and stricter preprocessing can significantly improve the quality of the model. Insan et al. (2023) as well as other studies using single algorithms such as Naïve Bayes or Logistic Regression generally achieving 80–88% accuracy. The performance of SVM in this study is superior to all classical models in previous studies. Research using deep learning models like LSTM or IndoBERT can indeed reach 90–94% accuracy, but require high computational resources and manually labeled datasets. With a lexicon-based approach, this study achieved accuracy similar to deep learning SVM, but with much lower computational cost. Almost no previous research has applied Ensemble Learning (SVM + Decision Tree) to BRImo reviews, making this approach one of the important novelties.
Thus, the position of this research provides a new contribution both in terms of methodology (lexicon + ensemble learning) and the larger, more representative quality of the dataset.

Classification Error Analysis
The confusion matrix for each model shows that misclassification most frequently occurs in the Neutral class. This can be understood because this class often contains ambiguous reviews or a mix of positive and negative sentiments. Ensemble Learning slightly improves the balance between classes but still cannot fully overcome the bias in the Neutral class.
Meanwhile, SVM demonstrates the best performance in avoiding False Positives and False Negatives in the Positive class, reflecting the model's stability on reviews containing clearly toned words such as good, fast, or excellent.

Practical Implications for the Development of BRImo
From the sentiment distribution results, 40.7% of user reviews are positive. This indicates that the majority of users are satisfied with BRImo's services. However, 25.7% of negative reviews indicate areas

of the service that need improvement, such as login difficulties, issues with unavailable usernames, system errors or disruptions, and the account verification process.
These findings can be used by BRImo developers to prioritize system improvements and enhance the user experience. In addition, the periodic implementation of sentiment analysis can become part of the service quality monitoring system.
Overall, this study successfully demonstrated that SVM is the most effective model for sentiment analysis of BRImo reviews, while Ensemble Learning provides predictive stability not found in single models. The lexicon-based approach proved efficient for large datasets and is feasible for use in similar studies. Compared to previous research, this study makes a significant contribution in terms of dataset size, ensemble methods, and classification result quality.

6. Conclusion

This study aims to analyze the performance of machine learning models in classifying the sentiment of BRImo app reviews by comparing SVM, Decision Tree, and Ensemble Learning algorithms based on majority voting. A total of 8,002 user reviews were collected from the Google Play Store and labeled automatically using an Indonesian lexicon approach. After undergoing preprocessing and converting the data into TF-IDF features, the three models were tested to assess sentiment classification performance.

The research results show that SVM provides the best performance with an accuracy of 92.63%, along with high and stable precision, recall, and F1-score. The Decision Tree model produced the lowest performance with an accuracy of 86.45%, yet still contributed to capturing non-linear patterns. Meanwhile, the Ensemble model (SVM + Decision Tree) achieved an accuracy of 89.38%, better than the Decision Tree but still unable to surpass SVM. Nevertheless, the ensemble approach has proven to produce more stable and balanced predictions through the majority voting mechanism.

In addition to demonstrating the model's performance, this study also found that user reviews of the BRImo application generally tend to be positive. Methodologically, this research provides a new contribution by integrating a lexicon-based labeling approach with Ensemble Learning, as well as using a larger and more representative dataset compared to previous studies.

Thus, this study can serve as a reference for BRImo developers to understand user perceptions based on review data, as well as a foundation for further studies related to improving mobile banking service quality using machine learning and NLP approaches..

References

Aryanti, N. N. A., & Suria, O. (2025). Analisis Sentimen Terhadap Pemutusan Hubungan Kerja di Indonesia: Komparasi IndoBERT dengan SVM, Random Forest, dan Decision Tree dengan Optimasi TF-IDF. *Rabit: Jurnal Teknologi Dan Sistem Informasi Univrab*, *10*(2), 1158–1176. https://doi.org/10.36341/rabit.v10i2.6364

Fudholi, D. H. (2022). *Klasifikasi Emosi Pada Teks Menggunakan Metode Deep Learning*. https://dspace.uii.ac.id/handle/123456789/40586

Fudholi, L. A., Rahaningsih, N., & Dana, R. D. (2024). Sentimen analisis perilaku penggemar Coldplay di media sosial Twitter menggunakan metode Naive Bayes. *Jurnal Mahasiswa Teknik Informatika*, *8*(3), 4150–4159. https://doi.org/10.36040/jati.v8i3.9827

Hermawan, M. A., Faqih, A., & Dwilestari, G. (2025). Implementasi Akurasi Model Naive Bayes Menggunakan Smote Dalam Analisis Sentimen Pengguna Aplikasi BRIMO. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1). https://doi.org/10.23960/jitet.v13i1.5748

Iffa, M. R., Agustian, S., Safaat, N., & Irsyad, M. (2025). Peningkatan kinerja Support Vector Machine menggunakan model bahasa BERT untuk klasifikasi sentimen dengan dataset terbatas. *ZONAsi: Jurnal Sistem Informasi*, *7*(2), 422–432. https://doi.org/10.31849/zn.v7i2.26847

Kurniawati, K. (2024). *Klasifikasi data mahasiswa lampau menggunakan metode decision tree dan support vector machine*. Universitas Islam Negeri Maulana Malik Ibrahim.

Lowell, A., Lowell, A., Candra, K., & Indra, E. (2025). Perbandingan Metode Support Vector

Machine (SVM) Dan Naive Bayes Pada Analisis Sentimen Ulasan Aplikasi OVO. *Jurnal Media Informatika*, *6*(2), 896–905. https://doi.org/10.55338/jumin.v6i2.5134

Marwi, H. C. (2024). Adaptasi Mobile Banking dalam Transaksi On-Line. *E-Jurnal JUSITI (Jurnal Sistem Informasi Dan Teknologi Informasi)*, *13*(1), 102–115. https://doi.org/10.36774/jusiti.v13i1.1559

Nurwahidah, D., Dwilestari, G., Nuris, N. D., & Narasati, R. (2023). Analisis Sentimen Data Ulasan Pengguna Aplikasi Google Kelas Pada Google Play Store Menggunakan Algoritma Naïve Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3673–3678. https://doi.org/10.36040/jati.v7i6.8829

Rachmatullah, M. I. C. (2025). Ensemble Learning untuk Klasifikasi: Tinjauan Komprehensif Metode, Aplikasi, dan Perkembangan Terkini. *Improve*, *17*(1), 29–31. https://ejurnal.ulbi.ac.id/index.php/improve/article/view/4659

Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki. *BIOS: Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *5*(2), 111–119. https://doi.org/10.37148/bios.v5i2.128

Rizkia, A. S., Wufron, W., & Roji, F. F. (2025). Analisis Sentimen Coretax: Perbandingan Pelabelan Data Manual, Transformers-Based, dan Lexicon-Based pada Performa IndoBERT: Sentiment Analysis of Coretax: A Comparison of Manual, Transformers-Based, and Lexicon-Based Data Labeling on IndoBERT Performance. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(3), 1037–1048. https://doi.org/10.57152/malcom.v5i3.2151

Rusman, J., Haryati, B. Z., & Michael, A. (2023). Optimisasi hiperparameter tuning pada metode Support Vector Machine untuk klasifikasi tingkat kematangan buah kopi. *J-Icon: Jurnal Komputer Dan Informatika*, *11*(2), 195–202. https://doi.org/10.35508/jicon.v11i2.12571

Sondakh, S.Kom, M.T, Ph.D, D. E., Taju, S. W., Tene, M. G., & Pangaila, A. E. T. (2023). Sistem Analisis Sentimen Ulasan Aplikasi Belanja Online Menggunakan Metode Ensemble Learning. *CogITo Smart Journal*, *9*(2), 280–291. https://doi.org/10.31154/cogito.v9i2.525.280-291

Trisnaningrum, N. A. W., Sishadiyati, S., & Priana, E. W. (2024). Mendorong Transformasi Digital Melalui Penggunaan Aplikasi BRImo Pada Bank Rakyat Indonesia Kantor Cabang Krian. *Jurnal Pengabdian Kepada Masyarakat Nusantara*, *5*(3), 3467–3474. https://doi.org/10.55338/jpkmn.v5i3.3672