**Jurnal Informatika dan Rekayasa Perangkat Lunak** ISSN 2656-2855
Vol. 8, No. 1, March 2026, Page. 59-70 e-ISSN 2685-5518

# Comparative Performance Analysis of ML, DL, and Transformer Models for Sentiment Classification of Indonesian Mobile Banking User Reviews

**Waeisul Bismi**[1*], **Siti Marlina**[2], **Muhammad Qommarudin**[3]
[1*]Program Studi Informatika, Universitas Bina Sarana Informatika
[2] Program Studi Sistem Informasi, Universitas Bina Sarana Informatika
[3] Program Studi Informatika, Universitas Nusa Mandiri

*Email: Waeisul.wbn@bsi.ac.id

**Abstract**

*The rapid development of digital technology has encouraged the adoption of mobile banking applications in Indonesia, but it has also led to an increase in user complaints and reviews regarding performance and ease of use. This study aims to conduct a comparative analysis of the performance of Machine Learning, Deep Learning, and Transformer (IndoBERT) models in classifying the sentiment of user reviews of Indonesian-language mobile banking applications. Data was collected through web scraping from the Google Play Store on ten leading banking applications in Indonesia with a total of 200,000 reviews. After going through the preprocessing stages of cleaning, normalisation, tokenisation, and stemming, automatic labelling was carried out based on ratings into three sentiment classes: positive, neutral, and negative. Machine learning models (Naïve Bayes, Logistic Regression, Random Forest, and SVM) were built using TF-IDF feature representation, while deep learning models (LSTM, Bi-LSTM, GRU, and CNN) utilised 128-dimensional word embeddings. The Transformer-based IndoBERT model was fine-tuned with a sequence classification configuration. The evaluation used accuracy, precision, recall, and weighted F1-score metrics, accompanied by an analysis of training and testing time efficiency. The results show that the Bi-LSTM model performs best with an accuracy of 83.47% and an F1-score of 80.78%, followed by CNN (83.11%) and SVM (82.85%), while IndoBERT records an accuracy of 81.73% with a precision of 76.96%. In terms of efficiency, Logistic Regression showed an optimal balance between accuracy and training time (27.7 seconds), while deep learning and transformer models required higher computational resources. This study emphasises the importance of model selection based on requirements, between maximum accuracy and computational efficiency, and enriches the literature on Indonesian sentiment analysis in the domain of digital financial services.*

***Keywords****: Sentiment Analysis, Mobile Banking, Machine Learning, Deep Learning, IndoBERT*

## INTRODUCTION

The development of digital technology has changed the way Indonesians conduct financial activities (Purwanto et al., 2022). One of the most prominent innovations in the banking sector is Mobile Banking, which allows users to conduct financial transactions anytime and anywhere via mobile devices (Salman, 2023). Bank Indonesia data shows significant growth in the value of Mobile Banking transactions, reaching trillions of rupiah every month (Anggraeni & Khadafi, 2022). Applications such as BCA Mobile, BRImo, Livin' by Mandiri, and BNI Mobile have become the main services in Indonesia's digital banking ecosystem. This growth indicates high technology adoption, but it is also accompanied by increasing user expectations and complaints regarding the stability, security, and ease of use of the applications.

User reviews on Google Play Store are now a rich source of data for evaluating the quality of digital banking services (Azarya & Budi, 2025). Each user can provide opinions that reflect satisfaction, frustration, or suggestions regarding specific features. The large number of reviews and their free-text nature make manual analysis difficult, requiring an automated approach based on sentiment analysis. Sentiment analysis serves to identify emotions or opinion tendencies (positive, negative, or neutral) from user review texts (Mola et al., 2024). With accurate analysis results, banks can understand public perceptions more quickly and based on

data, as well as make targeted service improvements.

Previous studies have utilised machine learning methods for sentiment analysis in mobile banking applications in Indonesia. Models such as Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Logistic Regression are widely used due to their simple implementation and high accuracy. However, conventional machine learning approaches still have limitations because they rely on statistical feature representations such as bag of words and TF-IDF, which do not capture the semantic context between words. In Indonesian, which has complex morphology and a variety of informal expressions (for example, in user reviews), this approach often loses contextual meaning. Therefore, Deep Learning-based approaches, such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), have emerged, which are capable of automatically learning word sequence patterns and sentence context without having to define manual features. Deep learning models have been proven to improve text classification accuracy in various domains, including sentiment analysis.

Furthermore, recent developments in the field of natural language processing (NLP) have been marked by the emergence of Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and IndoBERT, which was developed specifically for the Indonesian language. Transformer models use a self-attention mechanism to understand the bidirectional contextual relationships between words in a sentence, thereby providing a more accurate representation of meaning compared to conventional deep learning models. In the context of sentiment analysis, Transformer models have been proven to improve classification performance, especially for complex natural language texts, such as user reviews of applications.

Although various studies have examined the performance of classical machine learning algorithms in Mobile Banking application sentiment analysis, comprehensive comparisons between Machine Learning, Deep Learning, and Transformer models in the context of user reviews in Indonesia are still very limited. In fact, this cross-paradigm comparative analysis is important for determining the most effective and efficient approach to understanding public perceptions of national digital banking services.

## LITERATUR REVIEW

The development of digital technology and the internet has driven the increased use of Mobile Banking applications in Indonesia, which has also given rise to the need to understand user perceptions and satisfaction with these services. One approach that is widely used to automatically explore public opinion is text-based sentiment analysis of user reviews on the Google Play Store. A number of previous studies have applied various machine learning models to classify user sentiment towards mobile banking applications such as BCA Mobile, BRImo, BSI Mobile, and Mandiri Mobile Banking.

Research by (Sari et al., 2023) used the Naïve Bayes algorithm to analyse BCA Mobile user reviews collected through web scraping from the Google Play Store. With text preprocessing and word weighting using TF-IDF, this model achieved an accuracy of 82% with the majority of positive sentiments related to the ease and security of using the application. Similar results were also shown by (Nadira et al., 2023), who applied Naïve Bayes with the InSet (Indonesia Sentiment Lexicon) dictionary to Victoria Mobile Banking application review data. The model achieved a precision of 90.4%, recall of 100%, and accuracy of 93.1%, proving the effectiveness of Naïve Bayes on Indonesian-language text. Meanwhile, several studies compared the performance of different algorithms, such as the study by (Mifathusalam et al., 2023),which compared Random Forest and Naïve Bayes for sentiment classification on 2,453 BCA Mobile reviews, with the result that Random Forest excelled with an accuracy of 93.93% and an F1-score of 91.43%. Similar results were shown by (Astuti et al., 2022), who compared Support Vector Machine (SVM) and Naïve Bayes on BRImo review data. The study showed that SVM was superior with an accuracy of 97.69%, while Naïve Bayes achieved 96.53%.

Research by (Zelina & Afiyati, 2024), which analysed Motion Banking reviews, also found that SVM with a linear kernel produced an accuracy of 93.7%, which was better than Decision Tree, which only achieved 83%. Additionally, (Rizky Pratama et al., 2023)

combined the Lexicon-Based and Support Vector Machine methods in sentiment analysis of the BRImo and BCA Mobile applications. The results showed an accuracy rate of 94% for BRImo and 95% for BCA Mobile, with sentiment visualisation generated through Power BI. Meanwhile, (Aryanusa et al., 2025) compared Logistic Regression and Multinomial Naïve Bayes on 200,000 reviews from four popular applications (blu by BCA, BRImo, BNI Mobile, and Livin' by Mandiri). This study shows that Logistic Regression has higher accuracy (92–93%) than Naïve Bayes (71–74%), and finds that most reviews are negative, particularly regarding login and verification features. Another study by (Arifiyanti et al., 2023) evaluated four supervised learning algorithms, Multinomial Naïve Bayes, Support Vector Machine, Decision Tree, and K-Nearest Neighbour for sentiment analysis on the BSI Mobile application. The results showed that Naïve Bayes provided the best results with an ROC area of 0.84, followed by SVM (0.82).

These results indicate that classic machine learning algorithms such as Naïve Bayes, SVM, and Random Forest are still the primary choice in Indonesian text sentiment analysis due to their ease of implementation and relatively high interpretability of results. However, all of the above studies still focus on conventional machine learning approaches. Studies that integrate Deep Learning methods such as CNN, LSTM, or Bi-LSTM, as well as Transformer models such as BERT and IndoBERT, are still rare in the context of Mobile Banking applications in Indonesia. In fact, contextually-based models have better capabilities in capturing the semantic meaning and context of sentences than classic models based on statistical features (TF-IDF).

Therefore, this research is important to compare the performance of Machine Learning, Deep Learning, and Transformer models in analysing the sentiment of Mobile Banking application user reviews in Indonesia, thereby providing empirical contributions to the development of a more accurate and adaptive Indonesian language opinion analysis system.

## METHOD

The stages of the research methodology used in this study can be seen in Figure 1 below.

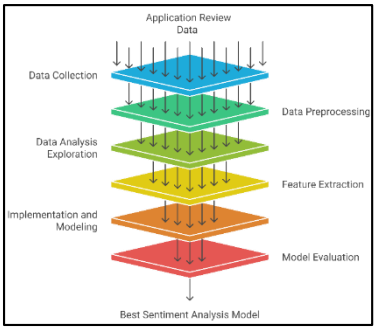



Figure 1. Research Methodology Flow

### Data Collection

Data is an important requirement, especially for researchers in analysing a phenomenon or searching for information informasi (Dwicahyo & Indah Ratnasari, 2023). The data collection process in this study was carried out using the web scraping technique, which is a technique for extracting data from websites (Flores et al., 2020) or, in this study, from user reviews available on Google Play Store. The main data sources were obtained from ten leading mobile banking applications in Indonesia.

The scraping method was carried out using the python-based google-play-scraper library in Google Colaboratory *Colaboratory* (Reyan & Purwaningtyas, 2025) with the parameters of Indonesian language (lang='id') and Indonesia country (country='id') to ensure the relevance of the data collected. Each application was targeted to obtain approximately 1,000 reviews, which were taken sequentially based on the latest reviews (sort by newest) to obtain a representation of the current condition of the application.

The data variables successfully collected include review text content, numerical ratings (1-5 stars), and timestamps recording the time the reviews were written. This data collection process produced a comprehensive dataset capable of representing users' actual perceptions and experiences of Mobile Banking services in Indonesia.

### Data Preprocessing

The data pre-processing stage is a critical step in this research, in order to prepare the review text data before sentiment analysis is carried out. This process is carried out through three main stages, namely text cleaning, normalisation, and data labelling.

1) The data cleaning process aims to improve data quality by detecting and correcting errors (Hosseinzadeh et al., 2023) nd eliminating noise (Romli, 2021) and characters that are irrelevant to the analysis, including the removal of URLs, user mentions (@username), hashtags, numbers, punctuation marks, and emojis that often appear in Mobile Banking application reviews. In addition, whitespace normalisation is performed to remove excess spaces that can interfere with the tokenisation process.
2) The normalisation stage aims to make several variables have the same value range (Kusnaidi et al., 2022), starting with case folding, which converts all text to lowercase for consistency in analysis. Next, tokenisation is performed using the NLTK library, which is a ready-to-use open source program package specifically for linguistic languages (Rifano et al., 2020) to break down text into smaller word units. The stopword removal process is applied using a list of Indonesian stopwords to remove irrelevant words in a sentence (Rinandyaswara et al., 2022). The final stage of normalisation is stemming using the Sastrawi algorithm, which is capable of processing words so that they can be converted into their basic form (Pardede & Darmawan, 2025), thereby reducing word variation and improving model accuracy.
3) The sentiment labelling process is carried out into three sentiment classes, namely negative, neutral, and positive sentiments (Rachmadana Ismail et al., 2023), which are done automatically based on the ratings given by users in their reviews. The label "negative" is given to reviews with a rating of 1-2 stars, which represents user dissatisfaction, the label "neutral" is given to reviews with a rating of 3 stars, which indicates a normal or neutral response, and the label "positive" is given to reviews with a rating of 4-5 stars, which reflects user satisfaction. This labelling approach is based on the assumption that there is a strong correlation between numerical ratings and the sentiment expressed in the text content of reviews.

After going through all stages of pre-processing, the clean and structured data can be viewed in data exploration and is ready for feature extraction and input into the sentiment analysis model building process.

**Data Analysis Exploration**

The analytical data exploration (EDA) stage is a process of examining or understanding data from Mobile Banking application reviews and exploring insights or key characteristics of the data (Siambaton & Husein, 2022). Analysis of the distribution of ratings and sentiments reveals the composition of user preferences through chart visualisation, while textual analysis using word clouds shows the words that frequently appear in positive and negative reviews, providing an intuitive picture of the aspects of the application that are most often praised or criticised. Quantitative analysis of word frequency in negative reviews identifies the most frequently appearing words that represent systematic patterns of user complaints.

**Feature Extraction**

At the feature extraction stage, the technique used is TF-IDF Vectorizer (Term Frequency-Inverse Document Frequency) to evaluate how important a word (term) is in a document in the context of a larger document collection (Septiani & Isabela, 2022) and is a word weighting technique that calculates the Term Frequency value and counts the occurrence of a word in the entire text document collection (Hafizh Mahendra et al., 2023), which is able to capture the importance of a word in a document relative to the entire corpus. The selection of TF-IDF is based on its ability to give higher weight to words that are significant in a particular document, while reducing the weight of words that commonly appear throughout the document.

Then, the data is split using stratified sampling by dividing the data into two parts, namely training data and testing data. The data is divided into 80% for training data and 20% for testing data (Dennis et al., 2022). This stratified division ensures that the distribution of sentiment classes (positive, neutral, negative) in both subsets remains proportional to the distribution of the original dataset, thus avoiding bias in model evaluation. The feature extraction results produce a TF-IDF matrix with dimensions that are optimal for training various models, where the training data is used for model

learning, while the testing data serves as a final evaluation of the model's generalisation ability on previously unseen data.

**Implementation And Modeling**

This study applies three comprehensive modelling approaches for sentiment analysis of Mobile Banking application reviews. The first approach uses traditional machine learning algorithms that enable sentiment classification tasks. The models implemented are:

1) Naive Bayes, because it has superior performance for testing categorical data types and attributes in data that are independent of each other (Nadira et al., 2023).
2) Logistic Regression, due to its ability to identify linear relationships between input and output variables (Adrian & Verawati, 2025) and provide easily interpretable results, making it ideal for initial analysis (Wahid & Utomo, 2024).
3) Support Vector Machine (SVM), because it has a stronger and mathematically defined concept and aims to find the optimal hyperplane by maximising the distance between data classes (Aldren Marpaung et al., 2024).
4) Random Forest, because it can produce relatively low errors, has good performance, and is suitable for large amounts of data (Mifathusalam et al., 2023) and combines several decision trees to provide more stable results and resistance to overfitting (Wahid & Utomo, 2024).

The second approach utilises Deep Learning techniques by implementing four different neural network architecture models, with inputs processed through an embedding layer with a dimension of 128, followed by sequence padding up to a length of 100 tokens. The hyperparameters applied include a dropout rate of 0.5 for regularisation, a learning rate of 0.001, and a batch size of 32 with early stopping to prevent overfitting and 20 epochs. The architectural models used include:

1) LSTM (Long Short-Term Memory), because this type of artificial neural network architecture can learn and process data sequentially. LSTM is suitable for analysing sequential data such as text and can recognise patterns in data to make predictions about sentiment (Rolangon et al., 2023).
2) Bi-LSTM (Bidirectional LSTM) processes information bidirectionally, as it captures contextual information at a higher level and has the ability to handle long-term dependencies in complex data (Rolangon et al., 2023).
3) GRU (Gated Recurrent Unit) has a structure similar to LSTM, but is simpler and more efficient because it only uses two gates, namely the update gate and the reset gate. The update gate is used to control how much new information will be stored in the cell memory, while the reset gate is used to control how much old information will be stored in the cell memory (Rolangon et al., 2023).
4) CNN (Convolutional Neural Network) for local feature extraction is faster in training and able to avoid overfitting, while still retaining important information for accurate sentiment classification (Widaad & Anggraini, 2024).

The third approach implements a Transformer-based model using IndoBERT (Indonesian Bidirectional Encoder Representations from Transformers), which is a Transformer-based model optimised for the Indonesian language with its ability to process text bidirectionally. IndoBERT is able to capture the context of words based on the information before and after them (Yoga Pratama et al., 2025). The indobenchmark/IndoBERT-base-p1 model was fine-tuned on the Mobile Banking review dataset with a training strategy optimised for computational efficiency.

Text preprocessing uses a special IndoBERT tokeniser with max_length=128, truncation, and padding configurations to standardise input length. The model architecture is initialised as sequence classification with three output labels, then transferred to the GPU if available to speed up computation. The training configuration includes 2 epochs with a batch size of 8 for training and 16 for evaluation, equipped with 50 warmup steps, 0.01 weight decay, and mixed precision (FP16) for memory efficiency.

The evaluation strategy is carried out every 50 steps with the best model stored based on accuracy, while the evaluation metrics include accuracy, precision, recall, and F1-score

calculated on a weighted basis. Thus, it is expected to be able to capture the nuances of sentiment in Mobile Banking application reviews more accurately than conventional models.

**Model Evaluation**

A comprehensive evaluation was conducted to measure the performance of all models implemented using a multi-aspect approach. The main evaluation metrics used included *accuracy*, *precision*, *recall, and F1-score*, which were calculated on *a weighted basis* to address class imbalance. *Accuracy* measures the proportion of correct predictions overall, while *precision* represents the consistency of positive predictions, and *recall* measures the model's ability to identify actual instances. *F1-score* is a key metric as *the harmonic mean* of *precision* and *recall*, providing a balanced view of model performance. *A confusion matrix* was used for detailed analysis of classification error patterns per sentiment class, identifying whether the model tended to experience *false positives* or *false negatives* in certain categories.

Computational efficiency is evaluated by measuring execution time, which includes *training time* and *testing time*. *Training time* represents the time required to learn patterns from training data, which is influenced by model complexity and dataset size. Testing time measures the speed of predictions on large-scale test data. Time measurements are performed using a high-precision time module, ensuring accuracy in recording performance differences between models.

Comparative analysis was conducted by considering the trade-off between classification performance and computational complexity. Models with high accuracy but requiring long training times may not be practical for production implementation, while fast but less accurate models are ineffective for classification tasks. The efficiency score was calculated as the ratio of accuracy to total execution time, providing a quantitative indicator of each model's computational effectiveness.

Visual analysis through an accuracy versus training time scatter plot helps identify models that achieve an optimal balance between performance and efficiency, enabling model selection based on specific application requirements, whether prioritising maximum accuracy or response speed.

**RESULT AND DISCUSSION**

Based on the research method applied, this study collected data through web scraping techniques on 10 mobile banking applications on Google Play Store, obtaining a total of 200,000 data points, the details of which can be seen in Table 1 below.

Tabel 1. Data Collection Results

| No. | Bank Name | Number of data |
|---|---|---|
| 1 | Bank Mandiri | 20,000 |
| 2 | Bank BRI | 20,000 |
| 3 | Bank BCA | 20,000 |
| 4 | BNI Bank | 20,000 |
| 5 | BTN Bank | 20,000 |
| 6 | BSI Bank | 20,000 |
| 7 | Neo Bank | 20,000 |
| 8 | Jago Bank | 20,000 |
| 9 | Amar Bank | 20,000 |
| 10 | SeaBank | 20,000 |

The collected data was then processed. The results of data processing in this study, which involved text cleaning, normalisation and labelling techniques, yielded 146,623 data points. Samples of the processed data can be seen in Figure 2 below.

| score | cleaned_text | processed_text | sentiment |
|---|---|---|---|
| 1 | setelah di update malah lemot bangetttt | update malah lemot bangetttt | negatif |
| 5 | Awesome | awesome | positif |
| 5 | bagus sekali mempermudah transaksi tp kadang was juga kalo sering kepending karna kesalahan sistem | bagus sekali mudah transaksi kadang was kalo sering pending karna salah sistem | positif |
| 5 | jaringan licin sering lelet dan gangguan | jaring licin sering lelet ganggu | positif |
| 5 | bagus banget | bagus banget | positif |
| 5 | pengalaman memakai aplikasi livin sangat baik | pengalaman pakai aplikasi livin sangat baik | positif |
| 5 | sangat baik | sangat baik | positif |
| 1 | Payah Gak ada KSM Mungkin udah waktunya ganti ke bank lain | payah gak ksm mungkin udah waktu ganti bank | negatif |
| 1 | susah login aplikasi sedang dalam perbaikan teruss | susah login aplikasi teruss | negatif |
| 2 | Verifikasi wajah sangat sulit sudah di lakukan sesuai arahannya tetap aja gagal verifikasi wajah | verifikasi wajah sangat sulit padahal sesuai arah tetap aja gagal verifikasi wajah | negatif |
| 5 | kenapa KSM di Livin saya gak muncul ya | ksm livin gak muncul | positif |
| 5 | sangat cepat | sangat cepat | positif |
| 4 | aplikasi ini mempermudah transaksi tapi kenapa ya dari pagi aplikasi nya TDK bisa di buka mau transaksi jadi ga bisa padahal biasa nya gpp | aplikasi mudah transaksi pagi aplikasi nya tdk buka mau transaksi jadi padahal biasa nya gpp | positif |
| 1 | no di rekomendd | rekomendd | negatif |
| 5 | bagus bgt | bagus bgt | positif |
| 1 | bukan oknum | bukan oknum | negatif |
| 5 | mempermudah transaksi ya pake livin | mudah transaksi pake livin | positif |
| 2 | susah di buka | susah buka | negatif |
| 1 | kenapa sekarang Livin JD LEMOT untuk login | sekarang livin lemot login | negatif |
| 1 | ini aplikasi lelet banget koneksinya ke internet walaupun jaringan sedang bagus tampilan login gak usah pakai ikut ikutan momentum peringatan peringatan negara mending perbaikin tuh ping internetnya biar login gak jadi lelet | aplikasi lelet banget koneksi internet walaupun jaring sedang bagus tampil login gak usah pakai ikut ikut momentum ingat ingat negara mending perbaikin tuh ping | negatif |

Figure 2. Data Preprocessing Results

After the data processing stage, the results of the data exploration reveal the characteristics and patterns of the dataset, which show an uneven distribution of sentiment in the Mobile Banking application reviews, with the majority of users tending to give positive evaluations, as

seen in the bar chart visualisation in Figure 3 and the ratings in the pie chart in Figure 4 below.

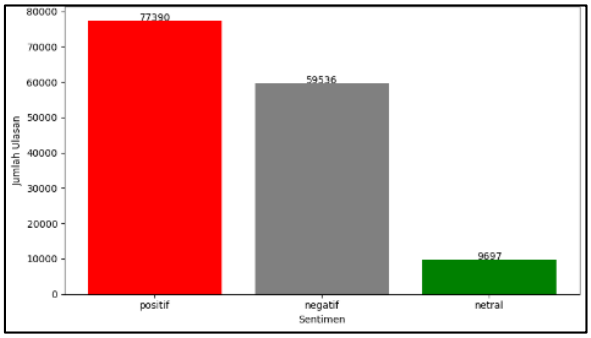



Figure 3. Sentimen Distribution

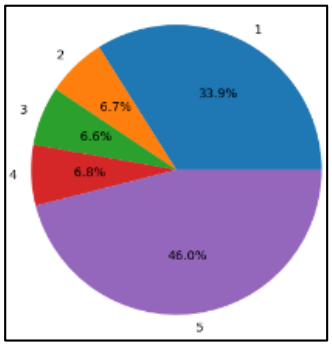



Figure 4. Rating Distribution

Textual analysis through word clouds successfully identified significant keyword patterns that differed between positive and negative reviews. Positive reviews are shown in Figure 5, which is dominated by terms such as "easy", "good", and "fast", while negative reviews are shown in Figure 6, which contains many words such as "error", "slow", and "difficult" that reflect the main pain points of users.


Figure 5. Positive Sentimen WordCloud


Figure 6. Negative Sentimen WordCloud

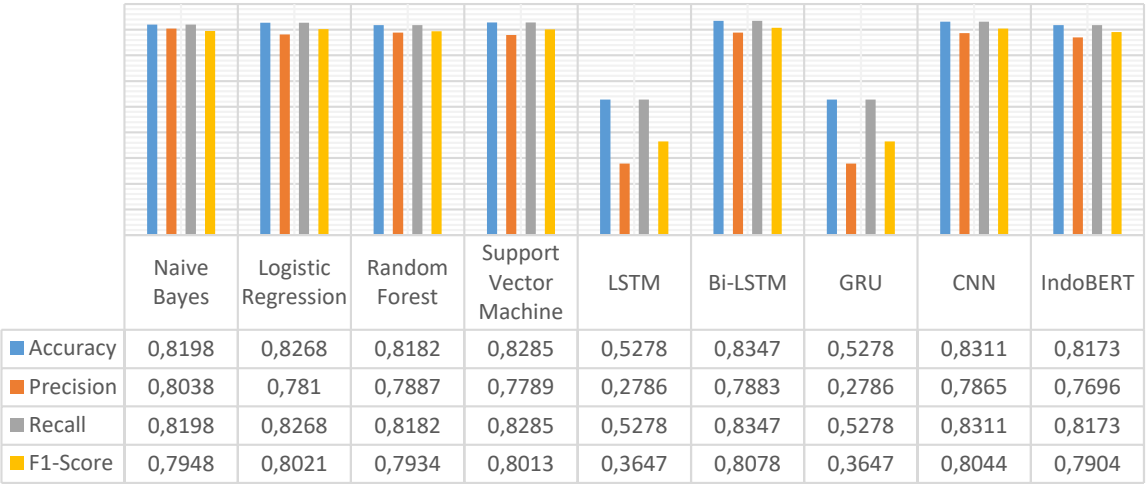
In terms of frequency analysis, there are 20 words that frequently appear in negative reviews, as shown in Table 2. This identification reveals common complaints such as technical problems (error, crash), performance issues (slow), and usability concerns (difficult), which not only form the foundation for developing a sentiment classification model but also provide valuable insights for application developers regarding critical areas that need improvement based on direct user feedback.

Table 2. Frequency Analysis Results

| No. | Word | Number of Words |
|---|---|---|
| 1 | Aplikasi | 17,796 |
| 2. | Guna | 12,436 |
| 3. | Mau | 11,730 |
| 4. | Gak | 10,658 |
| 5. | Masuk | 8,721 |
| 6. | Bank | 8,348 |
| 7. | Buka | 7,977 |
| 8. | Aja | 7,667 |
| 9. | Malah | 7,024 |
| 10. | Udah | 6,760 |
| 11. | Terus | 6,750 |
| 12. | Padahal | 6,744 |
| 13. | Transaksi | 6,594 |
| 14. | Login | 6,351 |
| 15. | Saldo | 6,304 |
| 16. | Susah | 5,762 |
| 17. | Baik | 5,759 |
| 18. | Update | 5,554 |
| 19. | Gagal | 5,044 |
| 20. | Baru | 5. 044 |

The implementation of feature extraction using the TF-IDF Vectorizer successfully converted the processed review text into an optimal numerical representation for modelling. The parameter configuration applied includes max_features=5000, which limits the number of features to the 5000 most significant words, and ngram_range=(1,2), which allows

the extraction of not only single words (unigrams) but also word pairs (bigrams), thus successfully identifying not only single words but also characteristic phrase patterns in Mobile Banking application reviews such as "difficult to log in", "smooth transfer", and "failed update" which have specific meanings in the context of user evaluation. And in the dataset division with a ratio of 80:20 using stratified sampling, a balanced class distribution between training and testing data can be seen in Table 3. This division ensures that the model is evaluated on data that truly represents the variation of the original population, so that the evaluation metrics obtained can be relied upon to measure the model's generalisation ability.



| | Naive Bayes | Logistic Regression | Random Forest | Support Vector Machine | LSTM | Bi-LSTM | GRU | CNN | IndoBERT |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 0,8198 | 0,8268 | 0,8182 | 0,8285 | 0,5278 | 0,8347 | 0,5278 | 0,8311 | 0,8173 |
| Precision | 0,8038 | 0,781 | 0,7887 | 0,7789 | 0,2786 | 0,7883 | 0,2786 | 0,7865 | 0,7696 |
| Recall | 0,8198 | 0,8268 | 0,8182 | 0,8285 | 0,5278 | 0,8347 | 0,5278 | 0,8311 | 0,8173 |
| F1-Score | 0,7948 | 0,8021 | 0,7934 | 0,8013 | 0,3647 | 0,8078 | 0,3647 | 0,8044 | 0,7904 |

Figure 7. Results of Machine Learning, Deep Learning, and Tranformer-IndoBERT Models

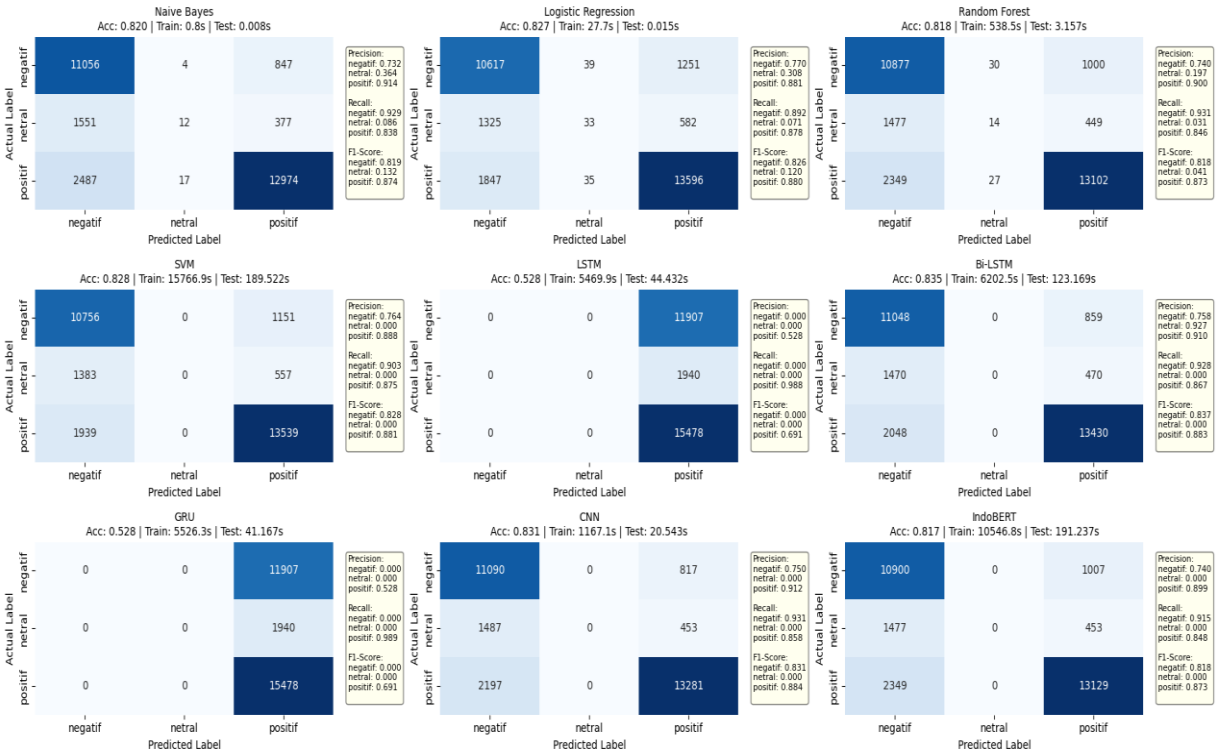



Figure 8. Confusion Matrix of Machine Learning, Deep Learning, and Transformer-IndoBERT

Table 3. Dataset Divison

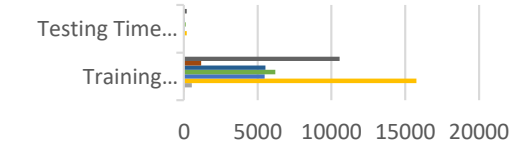
| No. | Type | Number of Data |
|---|---|---|
| 1 | Training Data | 117,298 |
| 2. | Testing Data | 29,325 |

Based on the results of the experiments conducted, this study successfully implemented and evaluated three modelling approaches for sentiment analysis of Mobile Banking application reviews, as shown in Figure 7. In the traditional machine learning approach, the four algorithms tested showed competitive performance with a range of 81-83%. Logistic Regression recorded the highest accuracy of 0.8268, followed by SVM with 0.8285, while Naive Bayes and Random Forest achieved 0.8198 and 0.8182, respectively.

Deep architecture shows different performance in analysing the sentiment of user reviews of mobile banking applications in Indonesia. The LSTM and GRU models have the same performance with an accuracy value of 0.5278, precision of 0.2786, recall of 0.5278, and F1-score of 0.3647, which indicates the model's inability to effectively distinguish sentiment classes, suggesting the possibility of overfitting or class distribution imbalance in the training data. The Bi-LSTM model showed a significant improvement in performance compared to LSTM and GRU, with an accuracy of 0.8347 and an F1-score of 0.8078. The bidirectional structure of Bi-LSTM allows the model to understand the context of words more deeply, both from the front and back directions, enabling it to capture more complex sentiment meanings. Meanwhile, the CNN model also showed competitive performance with an accuracy of 0.8311 and an F1-score of 0.8044, slightly lower than Bi-LSTM but with much higher computational efficiency.

The implementation of the IndoBERT Transformer model showed excellent performance in sentiment analysis of user reviews of mobile banking applications in Indonesia. This model achieved an accuracy of 0.8173, precision of 0.7696, recall of 0.8173, and an F1-score of 0.7904. These values indicate that IndoBERT is capable of consistently classifying sentiment between positive and negative predictions with a low error rate. This performance proves that the Transformer-based approach with the Bidirectional Encoder Representations from Transformers (BERT) architecture, which has been adapted for the Indonesian language, is capable of understanding complex linguistic contexts and variations in language expression in user reviews.

Comprehensive evaluation results of the nine implemented models show a confusion matrix in Figure 8 or significant performance variations in terms of classification accuracy and computational efficiency. In terms of classification accuracy, the Bi-LSTM model achieved the highest performance with an accuracy of 83.47%, followed by CNN (83.11%) and SVM (82.85%), while the LSTM and GRU models recorded the lowest accuracy of 52.78%. Precision and recall analysis revealed a consistent pattern with the highest F1-Score achieved by Bi-LSTM (80.78%), demonstrating the best ability to balance precision and recall across the three sentiment classes.




| | Training Time (s) | Testing Time (s) |
|---|---|---|
| IndoBERT | 10546,8 | 191,2373 |
| CNN | 1167,11 | 20,5426 |
| GRU | 5526,28 | 41,1671 |
| Bi-LSTM | 6202,52 | 123,1687 |
| LSTM | 5469,91 | 44,4315 |
| Support Vector Machine | 15766,8577 | 189,5218 |
| Random Forest | 538,4656 | 3,1569 |
| Logistic Regression | 27,7089 | 0,0153 |
| Naive Bayes | 0,8495 | 0,0076 |

Figure 9. Model Training and Testing Time Measurements

From a computational efficiency perspective, as shown in Figure 9, the time measurement results reveal a striking disparity. The Naive Bayes model recorded the best efficiency with a training time of only 0.85 seconds and a testing time of 0.008 seconds, making it the most computationally lightweight solution. In contrast, the SVM model required

the longest training time, reaching 15.766 seconds (≈4.4 hours), accompanied by a testing time of 189.52 seconds. The Transformer (IndoBERT) model also required significant computational resources, with a training time of 10.546 seconds and a testing time of 191 seconds.

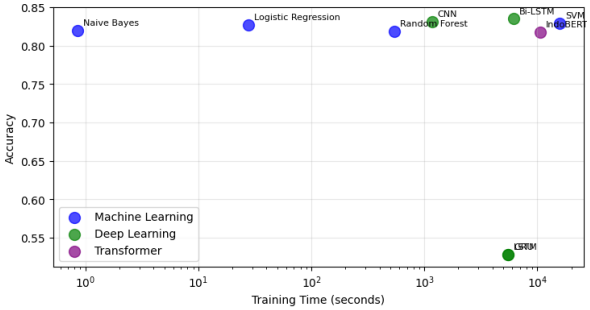



Figure 10. Trade-off: Accuracy vs Training time

The trade-off analysis between accuracy and efficiency in Figure 10 reveals that Machine Learning models are generally more efficient, with Logistic Regression offering an optimal balance between accuracy (82.68%) and reasonable training time (27.7 seconds). While Deep Learning models such as Bi-LSTM and CNN achieve slightly higher accuracy, they require 40-200 times longer training time compared to Logistic Regression. These results recommend model selection based on resource constraints: Naive Bayes for real-time applications with limited resources, Logistic Regression for practical implementations with adequate accuracy, and Bi-LSTM for scenarios that prioritise maximum accuracy at the expense of significant computational resources.

## CONCLUSION

The research results show that the *deep learning* approach, particularly the Bi-LSTM model, provides the best performance in analysing the sentiment of *mobile banking* application user reviews, with an accuracy of 83.47% and an F1-score of 80.78%. The bidirectional structure of Bi-LSTM is effective in understanding the complex linguistic context of the Indonesian language, while the CNN model shows competitive performance with higher training efficiency. Meanwhile, *machine learning* models such as Logistic Regression and SVM remain relevant for practical implementation thanks to their balance between accuracy and computational time efficiency. The IndoBERT *Transformer* model has proven capable of capturing deeper semantic context, although it requires greater computational resources.

In terms of efficiency, Logistic Regression is recommended for applications with limited resources, while Bi-LSTM is suitable for scenarios that demand maximum accuracy. This research provides an empirical contribution to understanding cross-paradigm performance in Indonesian text sentiment analysis, particularly in the domain of *mobile banking*. Further research could expand the approach by integrating inter-paradigm *ensemble* models such as *Hybrid Transformer–BiLSTM* to improve generalisation and accuracy. In addition, it is recommended to expand the data corpus to other platforms such as the App Store and social media, and to apply an *aspect-based sentiment analysis* approach to identify specific aspects (application features, security, performance) that most influence user perception.

## REFERENCES

Adrian, A., & Verawati, I. (2025). Analisis Performa *Logistic Regression* Dan *Random Forest* Dalam Klasifikasi Kelayakan Penerimaan Kredit. *Ijcsr: The Indonesian Journal Of Computer Science Research*, *4*(2), 148–158. Https://Subset.Id/Index.Php/Ijcsr

Aldren Marpaung, J., Devega, M., & Yulhemi. (2024). Analisis Sentimen Kepuasan Pengguna Aplikasi Bca Mobile Menggunakan Metode Naïve Bayes Dan Support Vector Machine (*SVM*). *Prosiding-Seminar Nasional Teknologi Informasi & Ilmu Komputer (Semaster)*, *3*(1), 249–261.

Anggraeni, R., & Khadafi, M. (2022, August 3). *Transaksi Digital Ngebut, Mobile Banking Jadi Primadona*. Finasial. Https://Finansial.Bisnis.Com/Read/20221003/90/1583574/Transaksi-Digital-Ngebut-Mobile-Banking-Jadi-Primadona

Arifiyanti, A. A., Shantika, N. R., & Syafira, A. O. (2023). Analisis Sentimen Ulasan Pengguna Bsi Mobile Pada Google Play Dengan Pendekatan Supervised Learning. *Jip (Jurnal Informatika Polinema)*, *9*(3), 283–288.

Aryanusa, A., Akbar, Ronny M., & Nur, Y. (2025). Evaluasi Kinerja *Logistic Regression* Dan Multinomial Naïve Bayes Dalam Klasifikasi Sentimen *Mobile Banking*. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(3s1). Https://Doi.Org/10.23960/Jitet.V13i3s1.7687

Astuti, A. P., Alam, S., & Jaelani, I. (2022). Komparasi Algoritma Support Vector Machine Dengan *Naive Bayes* Untuk Analisis Sentimen Pada Aplikasi Brimo. *Bangkit Indonesia*, *Xi*(02).

Azarya, Y., & Budi, I. (2025). Analisis Sentimen Berbasis Aspek Aplikasi Brimo Berdasarkan Ulasan Pengguna Di Google Playstore. *The Indonesian Journal Of Computer Science*, *14*(1), 1111–1125. Https://Doi.Org/10.33022/Ijcs.V14i1.4613

Dennis, M., Rahmaddeni, R., Zoromi, F., & Anam, M. K. (2022). Penerapan Algoritma Naïve Bayes Untuk Pengelompokkan Predikat Peserta Uji Kemahiran Berbahasa Indonesia. *Jurnal Media Informatika Budidarma*, *6*(2), 1183. Https://Doi.Org/10.30865/Mib.V6i2.3956

Dwicahyo, K., & Indah Ratnasari, C. (2023). Perbandingan Metode Web Scraping Dalam Pengambilan Data: Kajian Literatur. *Automata*, *4*(2). Https://Journal.Uii.Ac.Id/Automata/Article/View/28635/

Flores, V. A., Permatasari, P. A., & Jasa, L. (2020). Penerapan Web Scraping Sebagai Media Pencarian Dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword. *Majalah Ilmiah Teknologi Elektro*, *19*(2), 157. Https://Doi.Org/10.24843/Mite.2020.V19i02.P06

Hafizh Mahendra, M., Triantoro Murdiansyah, D., & Muslim Lhaksmana, K. (2023). *Analisis Sentimen Tweet Covid-19 Menggunakan Metode K-Nearest Neighbors Dengan Ekstraksi Fitur Tf-Idf Dan Countvectorizer*.

Hosseinzadeh, M., Azhir, E., Ahmed, O. H., Ghafour, M. Y., Ahmed, S. H., Rahmani, A. M., & Vo, B. (2023). Data Cleansing Mechanisms And Approaches For Big Data Analytics: A Systematic Study. *Journal Of Ambient Intelligence And Humanized Computing*, *14*(1), 99–111. Https://Doi.Org/10.1007/S12652-021-03590-2

Kusnaidi, M. R., Gulo, T., & Aripin, S. (2022). Penerapan Normalisasi Data Dalam Mengelompokkan Data Mahasiswa Dengan Menggunakan Metode K-Means Untuk Menentukan Prioritas Bantuan Uang Kuliah Tunggal. *Journal Of Computer System And Informatics (Josyc)*, *3*(4), 330–338. Https://Doi.Org/10.47065/Josyc.V3i4.2112

Mifathusalam, A., Pratiwi, H., & Slamet Isnandar. (2023). Perbandingan Metode *Random Forest* Dan *Naive Bayes* Pada Analisis Sentimen Review Aplikasi Bca Mobile. *"Peran Teknologi Pendidikan Menuju Pembelajaran Masa Depan: Tantangan Dan Peluang,"* 1–8.

Mola, S. A. S., Baun, D. L. B., Nunes, I. O., & Sani, M. M. A. R. (2024). Analisis Sentimen Aplikasi Halo Bca Di *Google Play Store* Menggunakan Metode *Naive Bayes*, Support Vector Machine Dan *Random Forest*. *Hoaq (High Education Of Organization Archive Quality) : Jurnal Teknologi Informasi*, *15*(2), 69–79. Https://Doi.Org/10.52972/Hoaq.Vol15no2.P69-79

Nadira, A., Setiawan, N. Y., & Purnomo, W. (2023). Analisis Sentimen Pada Ulasan Aplikasi *Mobile Banking* Menggunakan Metode Naïve Bayes Dengan Kamus Inset. *Indexia : Informatic And Computational Intelligent Journal*, *5*(1), 35–47. Https://Journal.Umg.Ac.Id/Index.Php/Indexia/Article/View/5138/3113

Pardede, J., & Darmawan, D. (2025). *Perbandingan Algoritma Stemming Porter, Sastrawi, Idris, Dan Arifin & Setiono Pada Dokumen Teks Bahasa Indonesia*. *12*(1), 69–76. Https://Doi.Org/10.25126/Jtiik.2025128860

Purwanto, H., Yandri, D., & Yoga, M. P. (2022). Perkembangan Dan Dampak Financial Technology (Fintech) Terhadap Perilaku Manajemen Keuangan Di Masyarakat. *Kompleksitas*, *11*(1), 80–92.

Rachmadana Ismail, A., Bagus, R., Hakim, F., & Artikel, R. (2023). Implementasi Lexicon Based Untuk Analisis Sentimen Dalam Mengetahui Trend Wisata Pantai Di Di Yogyakarta Berdasarkan Data Twitter P-Issn E-Issn. In *Emerging Statistics And Data Science Journal* (Vol. 1, Issue 1).

Reyan, M., & Purwaningtyas, F. (2025). Analisis Sentimen Ulasan Aplikasi Ibi Library Pada *Google Play Store* Menggunakan *Naive Bayes* Classifier. *Jurnal Teknologi Informasi*, *6*(2). Https://Doi.Org/10.46576/Djtechno

Rifano, E. J., Fauzan, Abd. C., Makhi, A., Nadya, E., Nasikin, Z., & Putra, F. N. (2020). Text Summarization Menggunakan Library Natural Language Toolkit (Nltk) Berbasis Pemrograman Python. *Ilkomnika: Journal Of Computer Science And Applied Informatics*, *2*(1), 8–17. Https://Doi.Org/10.28926/Ilkomnika.V2i1.32

Rinandyaswara, R., Sari, Y. A., & Furqon, M. T. (2022). *Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi)*. *9*(4). Https://Doi.Org/10.25126/Jtiik.202294707

Rizky Pratama, M., Ramadhan, Y. R., & Komara, M. A. (2023). Analisis Sentimen Brimo Dan Bca Mobile Menggunakan Support Vector Machine Dan Lexicon Based. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, *3*(12), 1439–1450.

Rolangon, A., Weku, A., & Sandag, G. A. (2023). Perbandingan Algoritma *LSTM* Untuk Analisis Sentimen Pengguna Twitter Terhadap Layanan Rumah Sakit Saat Pandemi Covid-19. *Jurnal Teika*, *13*(1), 31–40.

Romli, I. (2021). Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Klasifikasi Penyakit Ispa. *Indonesian Journal Of Business Intelligence (Ijubi)*, *4*(1), 10. Https://Doi.Org/10.21927/Ijubi.V4i1.1727

Salman, M. (2023). Analisis Faktor-Faktor Yang Mempengaruhi Minat Nasabah Dalam Menggunakan Layanan *Mobile Banking* Pada Bank Syariah. *Jurnal Persya: Perbankan Syariah*, *1*(1), 31–37.

Sari, W. F., Rahim, R., & Adrianto, D. F. (2023). Analisis Sentiment Review Pengguna Bca Mobile Menggunakan Teks Mining. *Cakrawala*, *6*(2), 981–987.

Septiani, D., & Isabela, I. (2022). Sintesia: Jurnal Sistem Dan Teknologi Informasi Indonesia Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks. *Sintesia: Jurnal Sistem Dan Teknologi Informasi Indonesia*, *1*(2), 81–88.

Siambaton, M. Z., & Husein, A. M. (2022). Menganalisis Data Kesehatan Global : Pendekatan Analisis Data Eksplorasi Visual. *Data Sciences Indonesia (Dsi)*, *1*(2), 41–49. Https://Doi.Org/10.47709/Dsi.V1i2.1315

Wahid, A. M., & Utomo, F. S. (2024). Optimasi *Logistic Regression* Dan *Random Forest* Untuk Deteksi Berita Hoax Berbasis Tf-Idf. *Jurnal Pendidikan Dan Teknologi Indonesia (Jpti)*, *4*, 381–392. Https://Doi.Org/10.52436/1.Jpti.602

Widaad, N., & Anggraini, D. (2024). Sentiment Analysis Of Chatgpt App User Reviews Using *SVM* And *CNN* Methods. *Jurnal Teknik Informatika (Jutif)*, *5*(6), 1687–1700. Https://Doi.Org/10.52436/1.Jutif.2024.5.6.4010

Yoga Pratama, A., Ananda Sanjaya, G., Khairunisa Lubis, N., & Rangga Aditya, M. (2025). Analisis Sentimen Publik Terkait Danantara Menggunakan Algoritma *IndoBERT* Pada Platform Media Sosial. *Metik Jurnal*, *9*(1), 2025. Https://Doi.Org/10.47002/Metik.V9i1.1055

Zelina, N., & Afiyati, A. (2024). Analisis Sentimen Ulasan Pengguna Aplikasi M-Banking Menggunakan Algoritma Support Vector Machine Dan Decision Tree. *Jlk*, *7*(1).