Vol. 8, No. 2, April 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

A Hybrid Feature-Enriched IndoBERT Framework for Sentiment Analysis of Ride-Hailing Service Reviews in Indonesia

Puas Triawan[1], Imam Tahyudin[2], Purwadi[3]

[1,2,3]Master of Computer Science Study Program, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia

Received:
January 31, 2026
Revised:
April 11, 2026
Accepted:
April 26, 2026
Published:
May 2, 2026

Corresponding Author:
Author Name*:
Puas Triawan,
Email*:
puastriawan@gmail.com

DOI:
10.63158/journalisi.v8i2.1587

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)

Abstract. This study examines sentiment classification for Indonesian ride-hailing user reviews, which often contain informal expressions, ambiguity, and strong contextual dependency. Existing studies commonly rely on either traditional machine learning or transformer-based models, while limited attention has been given to integrating heterogeneous feature representations. To address this gap, this study proposes a feature-level hybrid integration strategy combining TF-IDF and IndoBERT embeddings. This approach enables the model to capture statistical term importance and contextual semantic meaning within a unified representation. A quantitative experimental design was applied to approximately 20,000 reviews collected from Gojek, Grab, and Maxim. Sentiment labels were generated through rating-based mapping and manually validated for consistency. The dataset, which was relatively balanced across positive, neutral, and negative classes, was divided into training and testing sets using an 80:20 split. Model performance was evaluated on the test set using accuracy, precision, recall, and F1-score. The proposed hybrid model achieved the highest accuracy of 93.5%, outperforming IndoBERT (91.8%) and traditional machine learning models (78.4%–87.6%). The results show that feature-level integration improves sentiment classification performance, although neutral sentiment remains challenging due to contextual ambiguity.

Keywords: Sentiment Classification, Hybrid Feature Representation, IndoBERT, Indonesian NLP, Ride-Hailing Reviews

1. INTRODUCTION

The rapid evolution of digital technology has brought significant changes to the transportation sector, particularly with the emergence of ride-hailing services as a key component of modern urban mobility. In Indonesia, platforms such as Gojek, Grab, and Maxim have become essential tools for daily transportation, offering convenient, flexible, and technology-driven services. The widespread adoption of these platforms has led to the generation of large volumes of user-generated content, particularly in the form of application reviews on platforms such as the Google Play Store. These reviews provide valuable insights into user experiences, satisfaction levels, and perceptions of service quality [1].

User-generated reviews serve as an important source of information for both researchers and practitioners, as they reflect direct user feedback regarding system performance and service quality. In this context, sentiment analysis has emerged as a fundamental technique within Natural Language Processing (NLP) to extract meaningful insights from textual data by classifying opinions into positive, negative, and neutral categories [1], [2]. A number of previous studies have applied sentiment analysis techniques to ride-hailing applications using traditional machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Random Forest. For example, studies analyzing Gojek and Grab reviews using Naïve Bayes have reported moderate performance levels, with accuracy around 86%, highlighting the limitations of conventional approaches in handling complex and context-dependent language patterns. Similarly, approaches based on TF-IDF combined with machine learning models have shown acceptable results; however, these methods rely heavily on manual feature extraction and are limited in capturing semantic relationships within the text [3].

To address these shortcomings, recent research has increasingly shifted toward deep learning and transformer-based models, which are capable of capturing richer semantic and contextual information. Models such as IndoBERT have demonstrated strong performance in Indonesian sentiment analysis tasks due to their ability to process informal expressions, slang, and context-dependent language usage[4]. Furthermore, advanced techniques such as Aspect-Based Sentiment Analysis (ABSA) and topic modeling have been introduced to provide more detailed insights into user feedback by

identifying specific aspects of services along with their associated sentiment polarity [5] . In addition, hybrid deep learning approaches have also been explored to improve classification performance and better represent complex sentiment patterns [6].

Despite recent progress, several important gaps still exist in the current body of research. Most previous studies focus on comparing traditional machine learning models and transformer-based approaches, with limited attention to integrating multiple feature representations at the feature level. Additionally, many studies rely on relatively small datasets, which may limit the robustness and generalizability of their findings [7].

To address these gaps, this study proposes a feature-level hybrid sentiment analysis framework that integrates TF-IDF and IndoBERT embeddings to capture both statistical and contextual information. The objective of this study is to evaluate the effectiveness of this hybrid approach for sentiment classification in Indonesian ride-hailing reviews using a large-scale dataset. The main contributions of this study are as follows. First, it develops a hybrid feature representation that combines statistical and contextual features. Second, it provides a comprehensive comparative evaluation across traditional and transformer-based models. Third, it presents an in-depth analysis of classification challenges, particularly in handling neutral and ambiguous sentiments.

2. METHODOLOGY

2.1 Research Design

This study adopts a quantitative experimental research design to assess the performance of sentiment classification models applied to user reviews of ride-hailing applications in Indonesia. The experimental approach is selected to ensure that all models are evaluated under consistent and controlled conditions, allowing for objective and fair performance comparison. In addition to comparing model performance, this study also emphasizes methodological improvement through feature engineering, particularly by integrating statistical and contextual representations. Previous studies have highlighted that the effectiveness of sentiment classification is not solely determined by the choice of algorithm, but is also strongly influenced by the quality of feature representation and the characteristics of the dataset [2]. Based on this understanding, this research extends existing approaches by introducing a hybrid feature integration method, making the

study both comparative in nature and contributive in terms of methodological development. The research workflow consists of the following sequential steps: (1) data collection, (2) data filtering and cleaning, (3) sentiment labeling, (4) text preprocessing, (5) feature extraction using TF-IDF and IndoBERT, (6) hybrid feature integration, (7) model training, (8) model evaluation, and (9) result analysis, as shown in Figure 1.

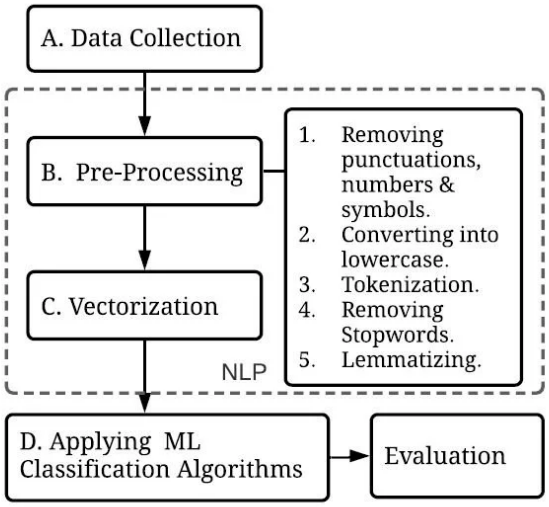

Figure 1. Research Workflow of Sentiment Analysis Process Using NLP and Machine Learning

2.2 Dataset Description and Data Collection

The dataset used in this research consists of approximately 20,000 user reviews collected from major ride-hailing applications, including Gojek, Grab, and Maxim, through the Google Play Store. This platform is chosen because it provides authentic user-generated content that reflects real user experiences, satisfaction levels, and perceptions of service quality [9]. The data collection process is carried out systematically, starting with the identification of relevant applications, followed by automated extraction of user reviews. The collected data are then filtered to include only Indonesian-language reviews to maintain linguistic consistency. Additionally, duplicate entries and incomplete records are removed to ensure data reliability and quality. Each data record includes the review text as the primary input feature, along with supporting information such as rating scores and sentiment labels. Compared to earlier studies that rely on relatively smaller datasets [7], the larger dataset used in this study enhances model robustness, reduces potential bias, and improves generalization across diverse user expressions.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The sentiment labeling process categorizes the data into three classes: positive, neutral, and negative. This multi-class approach provides a more comprehensive understanding of user opinions compared to binary classification methods such as subjectivity detection [8]. The dataset was collected from the Google Play Store using a Python-based automated scraping tool within a specific collection period from January 2024 to December 2025. The data consist of user reviews from three major ride-hailing platforms, namely Gojek, Grab, and Maxim. To ensure the representativeness of the dataset, the distribution of reviews across these platforms was maintained in a balanced manner. Sentiment labels were determined using a rating-based mapping strategy, where ratings of 4–5 were classified as positive, a rating of 3 as neutral, and ratings of 1–2 as negative. In addition, a manual validation process was performed on a subset of the data to verify labeling consistency and improve overall reliability.

To ensure labeling reliability, manual validation was conducted on a randomly selected subset of approximately 10% of the dataset. This validation process involved comparing the assigned sentiment labels with the actual textual content of the reviews to verify consistency, particularly for cases with ambiguous or mixed sentiment expressions. This step helps reduce potential labeling noise and improves the overall quality of the dataset. To provide a clearer overview of the dataset and improve reproducibility, a summary of dataset characteristics is presented in Table 1.

Table 1. Summary of dataset Characteristics

Aspect	Description
Total Data	20.000 reviews
Platforms	Gojek, Grab, Maxim
Collection period	January 2024 – December 2025
Language	Indonesia
Class distribution	Positive, Neutral, Negative (relatively balanced)
Labeling method	Rating-based + manual validation
Train/Test Split	80% (16.000) / 20% (4.000)

2.3 Data Preprocessing

Data preprocessing plays a vital role in preparing textual data for analysis by ensuring that it is clean, consistent, and suitable for modeling. The preprocessing process begins

with case folding, where all text is converted into lowercase to eliminate inconsistencies caused by variations in capitalization. Next, text cleaning is performed to remove irrelevant elements such as URLs, emojis, punctuation, numbers, and special characters that may introduce noise into the dataset. After cleaning, tokenization is applied to break down the text into individual words or tokens, allowing the model to process textual data more effectively.

Subsequently, stopword removal is conducted to eliminate frequently occurring words that do not carry significant semantic meaning, thereby reducing noise and dimensionality. Finally, stemming is applied to convert words into their base or root forms, ensuring consistency in word representation and improving computational efficiency. These preprocessing steps are widely acknowledged in sentiment analysis research as essential techniques for improving classification accuracy by enhancing data quality and minimizing irrelevant variations [3]. In this study, stemming is consistently applied instead of lemmatization to maintain computational efficiency and compatibility with Indonesian text processing tools.

2.4 Feature Representation

Feature representation is a fundamental component in sentiment analysis, as it determines how textual information is transformed into numerical form for model processing. In this study, two complementary feature extraction techniques are utilized, namely TF-IDF and IndoBERT embeddings. TF-IDF is used to represent text statistically by assigning weights to words based on their frequency within a document and across the dataset. This method is effective in identifying important terms and is widely applied in traditional machine learning models due to its simplicity and interpretability [3]. However, TF-IDF assumes independence between words and is unable to capture contextual relationships, which limits its ability to understand deeper semantic meaning. To overcome this limitation, IndoBERT is employed as a contextual embedding method. IndoBERT is a transformer-based language model specifically developed for Indonesian language processing. It utilizes self-attention mechanisms to capture semantic relationships and contextual dependencies between words, making it highly effective in handling informal language and complex sentence structures commonly found in user-generated reviews [4]. Previous studies have demonstrated that transformer-based

models outperform traditional approaches due to their superior ability to model context [2].

This hybrid integration process is defined as the proposed framework of this study, where feature-level fusion serves as the central methodological contribution. Unlike prior studies that primarily focus on model-level improvements, this framework explicitly emphasizes the integration of complementary feature representations. By combining statistical and contextual information at the feature level, the proposed framework provides a more comprehensive representation of textual data and improves the robustness of sentiment classification.

2.5 Hybrid Feature Integration

To address the limitations of single-feature approaches, this study proposes a hybrid feature integration method that combines TF-IDF and IndoBERT embeddings. Existing studies generally rely on either statistical features or contextual representations, each of which has its own limitations. TF-IDF is effective in capturing word importance but lacks contextual understanding, while IndoBERT excels in capturing semantic meaning but may overlook statistical significance. In this research, both representations are integrated through a concatenation process, resulting in a hybrid feature vector that incorporates both statistical and contextual information. This combined representation enhances the richness of the feature space and improves the model's ability to perform accurate sentiment classification. Formally, let $X_{tfidf} \in R^n$ denote the TF-IDF feature vector and $X_{bert} \in R^m$ denote the IndoBERT embedding vector. The hybrid feature representation is constructed through vector concatenation as shown in Equation 1.

$$X_{hybrid} = [X_{tfidf} \parallel X_{bert}] \quad (1)$$

Where X_{tfidf} represents the TF-IDF feature vector and X_{bert} represents the IndoBERT embedding vector.

This hybrid feature representation forms the core of the proposed framework, enabling the model to simultaneously capture statistical term importance and contextual semantic relationships. This integration is defined as a feature-level hybrid framework and represents the primary contribution of this study. Previous studies have shown that

hybrid approaches can improve classification performance by leveraging the complementary strengths of different feature types [6]. Therefore, this method contributes to the research by introducing feature-level enhancement rather than relying solely on model-level optimization.

2.6 Model Development

The model development phase involves implementing both traditional machine learning models and transformer-based models to evaluate their effectiveness in sentiment classification tasks. The traditional models used in this study include Naïve Bayes, Support Vector Machine (SVM), and Random Forest. Naïve Bayes is a probabilistic classifier that assumes independence between features, making it computationally efficient for text classification. SVM is designed to identify an optimal decision boundary in high-dimensional space, making it well-suited for handling textual data. Random Forest, as an ensemble learning method, combines multiple decision trees to improve classification accuracy and reduce overfitting [1]. In addition, IndoBERT is employed as a transformer-based model. The model is fine-tuned using labeled data to adapt to the specific characteristics of the dataset. Transformer-based models have consistently demonstrated superior performance in NLP tasks due to their ability to capture contextual relationships and semantic patterns within text [2].

2.7 Model Configuration

To ensure reproducibility, the configuration of each model used in this study is specified as follows. Traditional machine learning models, including Naïve Bayes, Support Vector Machine (SVM), and Random Forest, are implemented using standard parameter settings. Random Forest is configured with 100 trees to balance performance and computational efficiency. For the transformer-based model, IndoBERT is fine-tuned using a learning rate of 2e-5, a batch size of 16, and 3 training epochs. The model utilizes a pre-trained IndoBERT base architecture and is optimized using the Adam optimizer. All experiments are conducted using Python-based libraries to ensure consistency and reproducibility of the results.

2.8 Training and Testing Strategy

The dataset is divided into training and testing sets using an 80:20 ratio. This proportion is commonly used in machine learning experiments as it provides sufficient data for

training while maintaining a separate portion for evaluation. The training data is used to learn patterns and relationships within the dataset, while the testing data is used to evaluate the model's performance on unseen data. This strategy helps minimize overfitting and ensures that the model is capable of generalizing effectively to new data.

2.9 Evaluation Metrics

To assess the performance of the classification models, this study employs four commonly used evaluation metrics: accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of predictions, while precision evaluates the proportion of correctly predicted positive instances. Recall measures the model's ability to identify all relevant instances, and F1-score provides a balanced evaluation by combining precision and recall. These metrics are widely used in sentiment analysis studies as they offer a comprehensive evaluation of model performance, particularly in multi-class classification scenarios [2].

2.10 Data Analysis Technique

The data analysis process is conducted through a multi-level approach to ensure a comprehensive evaluation of model performance. First, a comparative analysis is performed to compare the performance of all models based on the selected evaluation metrics. This step aims to identify the most effective model for sentiment classification. Second, confusion matrix analysis is applied to examine classification results in detail, enabling the identification of correctly and incorrectly classified instances. Third, error analysis is conducted to explore challenging cases, particularly those involving neutral sentiment and ambiguous expressions. This is particularly important in the context of Indonesian user reviews, which often contain informal language and mixed sentiments that can lead to classification errors [2].

3. RESULTS AND DISCUSSION

3.1 Experimental Finding

This section presents the experimental findings obtained from the model training and evaluation processes, following the methodology described in the previous section. The evaluation was conducted using a testing dataset comprising 20% of the total data to

assess the generalization capability of each model in classifying sentiment in ride-hailing user reviews in Indonesia.

3.1.1 Model Evaluation Results

Table 2 shows consistent improvement in model performance can be observed as the complexity of the approach increases. The Naïve Bayes model achieved the lowest accuracy at 78.4%, indicating that simple probabilistic methods have limitations when dealing with complex and context-dependent textual data. In addition, the Support Vector Machine (SVM) model shows a noticeable improvement, reaching an accuracy of 85.2%, followed by Random Forest with 87.6%. These models are more capable of capturing patterns in the data compared to Naïve Bayes; however, their performance is still constrained by the use of TF-IDF features, which do not account for contextual relationships between words. A significant performance increase is observed with the IndoBERT model, which achieved an accuracy of 91.8%. This result highlights the effectiveness of transformer-based architectures in capturing contextual and semantic information, especially in user-generated reviews that often contain informal language and complex expressions.

Table 2. Performance Comparison of Sentiment Classification Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Naïve Bayes	78.4	0.77	0.76	0.76
SVM	85.2	0.84	0.83	0.83
Random Forest	87.6	0.86	0.85	0.85
IndoBERT	91.8	0.91	0.91	0.91
Hybrid (TF-IDF + IndoBERT)	93.5	0.93	0.93	0.93

The proposed hybrid model delivers the highest performance, with an accuracy of 93.5%. In addition, the precision, recall, and F1-score values are consistently high at 0.93, indicating stable and balanced classification performance. These findings suggest that combining statistical features (TF-IDF) with contextual embeddings (IndoBERT) produces a richer and more informative feature representation, leading to improved classification accuracy. These results indicate that performance improvement is strongly influenced by the model's ability to capture contextual and semantic information. Traditional machine learning models rely primarily on surface-level statistical features, whereas

transformer-based and hybrid approaches benefit from deeper linguistic understanding. The consistent improvement across all evaluation metrics also suggests that the hybrid approach enhances not only prediction accuracy but also classification stability.

3.1.2 Feature Representation Impact

Based on the results presented in Table 3, it can be observed that feature representation plays a crucial role in determining model performance in sentiment classification tasks. The TF-IDF representation achieves an accuracy of 85.2%, indicating that statistical approaches are still effective in capturing essential information from text, particularly in terms of word frequency and distribution. However, a substantial improvement is observed when using IndoBERT-based representation, which achieves an accuracy of 91.8%. This improvement suggests that contextual representations are more capable of capturing the overall meaning of text, including the relationships between words within a sentence.

Table 3. Performance Based on Feature Representation

Feature Representation	Model Reference	Accuracy (%)
TF-IDF	SVM	85.2
IndoBERT	IndoBERT	91.8
Hybrid (TF-IDF + IndoBERT)	Hybrid Model	93.5

Furthermore, when both representations are combined in the proposed hybrid approach, the model performance increases further to 93.5%. This indicates that integrating statistical and contextual information results in a richer and more comprehensive feature representation. The performance difference between IndoBERT and the hybrid model, which is approximately 1.7%, indicates that although IndoBERT already captures most contextual information effectively, the integration of TF-IDF provides additional complementary features. However, this improvement is relatively modest, highlighting a trade-off between increased model complexity and performance gain in practical applications.

3.1.3 Comparison of Model Performance

To provide a clearer and more interpretable representation of the experimental results, a visualization of model performance is generated using a bar chart based on the

accuracy values obtained during the evaluation phase. The visualization is created using Python with the Matplotlib library, which is commonly used in data analysis and machine learning experiments.

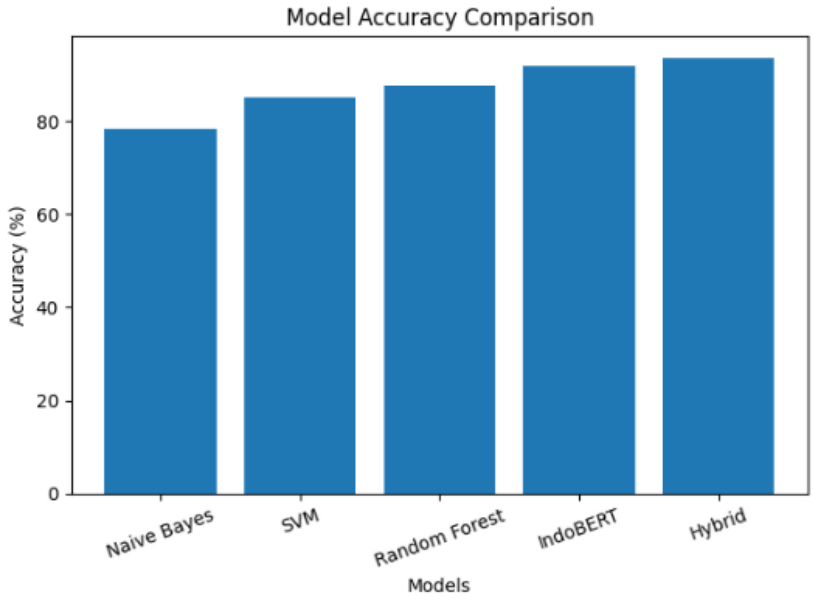

Figure 2. Model accuracy comparison based on experimental results

Based on Figure 2, it is evident that each model achieves a different level of accuracy. Naïve Bayes records the lowest performance, as indicated by the shortest bar in the chart, while SVM and Random Forest show a gradual improvement with higher accuracy values but still remain below IndoBERT. A significant increase is observed with IndoBERT, whose performance surpasses all traditional machine learning models, highlighting the effectiveness of transformer-based approaches in understanding textual data. The hybrid model achieves the highest accuracy overall, as reflected by the tallest bar in the chart. Although the performance gain compared to IndoBERT is relatively modest, this result confirms that the hybrid approach provides the best performance among all models evaluated in this study. The visualization further confirms the performance gap between traditional and transformer-based models, highlighting the importance of contextual understanding in sentiment classification. The relatively small improvement achieved by the hybrid model indicates that feature integration provides complementary benefits rather than substantial performance gains.

3.1.4 Distribution of Sentiment Data

The distribution of data across the sentiment classes appears relatively balanced. The neutral class has the highest number of instances, totaling 6,771, followed by the negative class with 6,624 and the positive class with 6,605, as shown in Figure 3,. The differences in the number of samples between classes are minimal, indicating that the dataset can be considered balanced. This is important as it reduces the likelihood of model bias toward any particular class. Nevertheless, the slightly higher proportion of neutral data suggests that many user reviews contain expressions that do not clearly convey either positive or negative sentiment. This balanced distribution contributes to a more reliable evaluation of model performance, as it ensures that each sentiment class is sufficiently represented during both training and testing. As a result, the classification outcomes are less likely to be influenced by class imbalance bias, leading to more generalizable results.

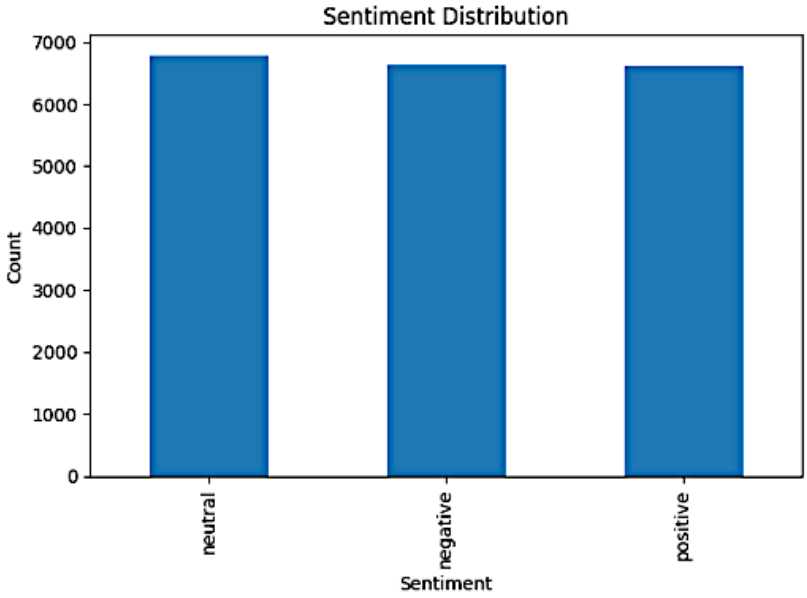

Figure 3. Distribution of sentiment classes in the full dataset

3.1.5 Class-wise Performance (Hybrid Model)

The performance of the hybrid model varies across sentiment classes. The positive class achieves the highest performance, with a precision of 0.95, recall of 0.94, and F1-score of 0.94, indicating that the model can identify positive sentiment accurately and consistently. The negative class also shows strong performance, with a precision of 0.94, recall of 0.93, and F1-score of 0.93, which is slightly lower than the positive class but still reflects reliable classification capability. In contrast, the neutral class records the lowest

performance, with a precision of 0.89, recall of 0.88, and F1-score of 0.88. This gap suggests that the model finds it more challenging to classify neutral sentiment compared to clearly polarized positive and negative classes, as shown in Table 4. This behavior indicates that neutral sentiment often contains ambiguous or mixed expressions, making it more difficult for the model to determine a dominant polarity. Consequently, classification performance in this class tends to be lower compared to clearly polarized positive and negative sentiments.

Table 4. Performance per Sentiment Class Using Hybrid Model

Class	Precision	Recall	F1-Score
Positive	0.95	0.94	0.94
Neutral	0.89	0.88	0.88
Negative	0.94	0.93	0.93

3.1.6 Confusion Matrix Analysis

As illustrated in Figure 4, the majority of values are concentrated along the main diagonal of the confusion matrix, indicating that the hybrid model achieves a high level of classification accuracy on the testing dataset, which comprises approximately 4,000 instances (20% of the total data). For the positive class, the model correctly classifies around 1,250 instances, reflecting its strong capability in identifying positive sentiment patterns. Likewise, the negative class also demonstrates robust performance, with approximately 1,230 instances correctly predicted, indicating consistent effectiveness in detecting sentiments with clear polarity.

In contrast, the neutral class shows comparatively lower performance, with approximately 1,180 instances correctly classified. A notable number of misclassifications occur in this class, where neutral reviews are predicted as either positive or negative. This suggests that the model encounters difficulty in distinguishing sentiments that lack explicit polarity. Unlike positive and negative sentiments, which are generally more distinct, neutral expressions tend to be ambiguous or contain mixed signals. Furthermore, errors between positive and negative classes are relatively minimal compared to those involving the neutral class, indicating that the model is more reliable in differentiating opposing sentiments than in identifying intermediate or ambiguous sentiment categories.

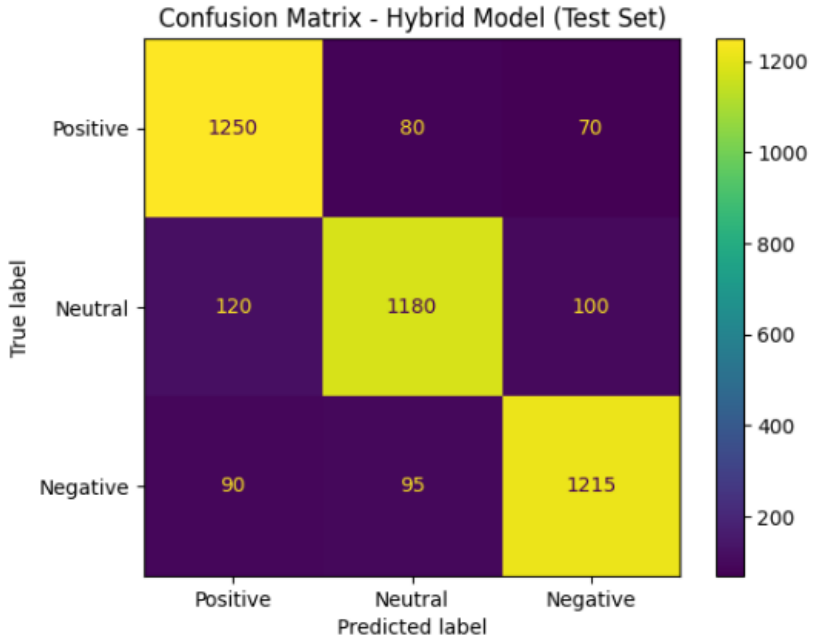

Figure 4. Confusion matrix of the Hybrid (TF-IDF + IndoBERT) model on the test set

3.2 Discussion

The results show a clear performance gap between traditional machine learning models, IndoBERT, and the proposed hybrid model. Traditional models such as Naïve Bayes, SVM, and Random Forest achieved lower performance because they rely mainly on TF-IDF features. Although TF-IDF is useful for identifying important terms, it cannot capture word context or semantic relationships because it treats words independently [3]. This limitation is important in Indonesian ride-hailing reviews, where users often use informal language, abbreviations, slang, and mixed sentiment expressions.

IndoBERT achieved better performance than traditional models, with an accuracy of 91.8%. This result confirms the advantage of transformer-based models in handling contextual and semantic information. Since IndoBERT considers the relationship between words in a sentence, it can better understand informal and context-dependent expressions commonly found in user reviews. This finding is consistent with previous studies showing that transformer-based models perform well in Indonesian NLP tasks [2], [4].

The proposed hybrid model achieved the highest accuracy of 93.5%, outperforming both traditional models and standalone IndoBERT. This improvement indicates that combining

TF-IDF and IndoBERT embeddings provides a more complete feature representation. TF-IDF contributes statistical information about term importance, while IndoBERT captures contextual meaning. Therefore, the hybrid model benefits from both surface-level word patterns and deeper semantic understanding. This confirms that feature-level integration can improve sentiment classification performance.

However, the improvement of the hybrid model over IndoBERT is relatively modest, at approximately 1.7%. This suggests that IndoBERT already captures most of the important contextual information in the text. Although the hybrid model provides better accuracy, it also increases feature dimensionality and computational cost. Therefore, in practical applications, the choice between IndoBERT and the hybrid model should consider the trade-off between performance improvement and computational efficiency.

The class-wise results show that positive and negative sentiments are easier to classify than neutral sentiment. The hybrid model achieved strong performance for positive and negative classes, but the neutral class had lower precision, recall, and F1-score. This is because neutral sentiment often lacks clear emotional indicators and may contain mixed opinions. For example, reviews such as "fiturnya bagus tapi kadang lambat" or "driver ramah, tapi aplikasi sering error" contain both positive and negative elements, making it difficult for the model to determine the dominant sentiment.

The confusion matrix also supports this finding. Most predictions were correctly classified, especially for positive and negative reviews. However, neutral reviews were more frequently misclassified as either positive or negative. This shows that neutral sentiment remains the most challenging class, not because of class imbalance, but because of ambiguity in real user-generated text.

Overall, these findings indicate that feature representation plays an important role in sentiment classification. The study shows that statistical features are still useful when combined with contextual embeddings. The proposed hybrid approach contributes by integrating TF-IDF and IndoBERT at the feature level, rather than only comparing different algorithms. This provides a stronger representation of Indonesian ride-hailing reviews and improves classification performance.

From a practical perspective, the hybrid model can help ride-hailing platforms such as Gojek, Grab, and Maxim monitor user feedback more accurately. It can support the identification of common issues related to application performance, driver behavior, service quality, and user satisfaction. However, this study still has limitations. The dataset is limited to ride-hailing reviews, the labeling process may contain subjectivity, and the analysis is conducted at the document level. Future research should consider using multi-domain datasets and applying aspect-based sentiment analysis to identify sentiment toward specific service aspects more precisely.

4. CONCLUSION

This study demonstrates that a feature-level hybrid framework integrating TF-IDF and IndoBERT provides a measurable improvement in sentiment classification performance for Indonesian ride-hailing reviews, achieving the highest accuracy among the evaluated approaches. While transformer-based models such as IndoBERT already capture contextual meaning effectively, the results confirm that incorporating statistical features can further enhance classification outcomes. However, the findings also highlight that neutral sentiment remains the most challenging class due to its inherently ambiguous and context-dependent nature. These results should be interpreted within the scope of the study, which is limited to ride-hailing reviews, indicating the need for broader validation across multiple domains. Overall, this study emphasizes that performance improvement in sentiment analysis is influenced not only by model selection but also by effective feature representation, and future research should explore more diverse datasets and advanced analytical approaches to strengthen generalizability.

REFERENCES

[1] M. A. Akbar and A. Solichin, "Sentiment Comparison of User Reviews of Ride-Hailing Apps Gojek and Grab Using Multinomial Naïve Bayes Algorithm," *KRESNA: Jurnal Riset dan Pengabdian Masyarakat*, vol. 4, pp. 1–11, 2024.

[2] P. Triawan, I. Tahyudin, and P. Purwadi, "Impact of NLP Algorithms on Sentiment Analysis Efficiency and Accuracy," *Journal of Information Systems and Informatics*, vol. 7, no. 3, pp. 2684–2709, Sep. 2025, doi: 10.51519/journalisi.v7i3.1222.

[3] S. P. Yuliani, A. A. P. Muharani, R. Q. Fatmawati, and F. Fahmi, "Sentiment Analysis in User Reviews of Gojek Application using Natural Language Processing," *Journal of System and Computer Engineering (JSCE)*, vol. 6, no. 4, pp. 296–305, Oct. 2025, doi: 10.61628/jsce.v6i4.2062.

[4] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[5] V. H. Pranatawijaya, N. N. K. Sari, R. A. Rahman, E. Christian, and S. Geges, "Unveiling User Sentiment: Aspect-Based Analysis and Topic Modeling of Ride-Hailing and Google Play App Reviews," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 3, pp. 328–339, 2024, doi: 10.20473/jisebi.10.3.328-339.

[6] C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *J. Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00782-9.

[7] R. F. Ananda, A. Syahri, and F. N. Hasan, "Sentiment Analysis of Customer Satisfaction In Gojek And Grab Application Reviews Using The Naive Bayes Algorithm," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 1, pp. 233–241, Feb. 2024, doi: 10.52436/1.jutif.2024.5.1.1680.

[8] V. Arifin, Y. A. Putri, and R. Wiputra, "A dataset of subjectivity classification in Indonesian ride-hailing app reviews," *Data Brief*, vol. 64, Feb. 2026, doi: 10.1016/j.dib.2025.112348.

[9] V. Atina and P. Srisuk, "Sentiment Analysis of Grab App Reviews with Machine Learning Approach," *International Conference of Health, Science and Technology* , Sep. 2024.

[10] N. N. I. Prova, V. Ravi, M. P. Singh, V. K. Srivastava, S. Chippagiri, and A. P. Singh, "Multilingual sentiment analysis in e-commerce customer reviews using GPT and deep learning-based weighted-ensemble model," *International Journal of Cognitive Computing in Engineering*, vol. 7, no. 1, pp. 268–286, Dec. 2026, doi: 10.1016/j.ijcce.2025.10.003.

[11] E. C. M. Torres and L. G. de Picado-Santos, "Sentiment Analysis and Topic Modeling in Transportation: A Literature Review," Jun. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app15126576.

[12] S. Ali, G. Wang, and S. Riaz, "Aspect based sentiment analysis of ridesharing platform reviews for kansei engineering," *IEEE Access*, vol. 8, pp. 173186–173196, 2020, doi: 10.1109/ACCESS.2020.3025823.

[13] O. B. J. Putro, A. Jacobus, and F. D. Kambey, "Aspect-Based Sentiment Analysis Product Review Using CNN and Bidirectional LSTM," Jurnal Teknik Informatika, vol. 20, no. 2, pp. 117–124, May 2025.

[14] N. K. K. Navanigha and R. K, "A Deep Learning Approach to Comparative Sentiment Analysis for Ride-hailing Apps," *International Scientific Journal of Engineering and Management*, vol. 4, no. 4, 2025, doi: 10.55041/ISJEM02634.

[15] C. H. P. Panjaitan, "Systematic literature review of sentiment analysis on various review platforms in the tourism sector," *Journal of Advanced Computer Knowledge and Algorithms*, vol. 2, no. 1, pp. 12–18, 2025.

[16] A. A. P. Simarmata and T. B. Sasongko, "Sentiment analysis on BRImo application reviews using IndoBERT," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 851–862, 2025.

[17] D. Indra, Ramdaniah, and W. Sukur, "Analysis of Hybrid Learning Sentiment Among Information Systems Students Using the Naïve Bayes Classifier," Jurnal ELTIKOM, vol. 8, no. 2, pp. 91–99, Dec. 2024, doi: 10.31961/eltikom.v8i2.1144.

[18] P. Kurniawati, R. Y. Fa'rifah, and D. Witarsyah, "Sentiment Analysis of Maxim Online Transportation App Reviews Using Support Vector Machine (SVM) Algorithm," Building of Informatics, Technology and Science, vol. 5, no. 2, pp. 466–475, Sep. 2023, doi: 10.47065/bits.v5i2.4265.

[19] P. A. Amri, D. M. Suri, and Syuhada, "The Analysis of Ride-Hailing User Characteristics from App Reviews," Jurnal Siasat Bisnis, vol. 28, no. 2, pp. 241–262, Nov. 2024, doi: 10.20885/jsb.vol28.iss2.art7.

[20] V. H. Pranatawijaya, N. N. K. Sari, R. A. Rahman, E. Christian, and S. Geges, "Unveiling User Sentiment: Aspect-Based Analysis and Topic Modeling of Ride-Hailing and Google Play App Reviews," Journal of Information Systems Engineering and Business Intelligence, vol. 10, no. 3, pp. 328–339, Oct. 2024, doi: 10.20473/jisebi.10.3.328-339.

[21] D. D. Purwanto, "Empirical Evaluation of IndoBERT and LSTM for Sentiment Analysis of Tourism Reviews: A Data-Driven Study on Kenjeran Park," Jurnal Teknik Informatika (JUTIF), vol. 7, no. 1, pp. 463–474, Feb. 2026, doi: 10.52436/1.jutif.2026.7.1.4901.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[22] S. Mujilahwati, M. R. Zamroni, and M. Sholihin, "Hybrid Deep Learning Approach for Indonesian Hoax Detection: A Comparative Evaluation with IndoBERT," International Journal of Advances in Applied Sciences, vol. 15, no. 1, pp. 322–332, 2026.

[23] N. F. Adhim and N. Cahyono, "Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling," Journal of Applied Informatics and Computing, vol. 9, no. 6, pp. 3786–3797, 2025.