Journal of Computing and Smart Ecosystem
Universitas Muhammadiyah Semarang
Vol. 2 No.2, 2026: 53-67

J-CaSE
e-ISSN 3110-5777

Comparative Sentiment And Complaint Pattern Analysis Of Indonesian E-Wallet Application Reviews Using Indobert

Firstiawan Fadhil Thobibuddin[1*], Gita Ayunda Dewi[2]
[1] Department of Accounting and Information System, Faculty of Computer Science, Soegijapranata Catholic University, Indonesia
[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, State university of Surabaya, Indonesia

***Corresponding author**: *firstiawanft07@gmail.com*

Article Info:
Received: May 20, 2026
Accepted: May 30, 2026
Available Online: June 30, 2026

Keywords:
Sentiment Analysis; IndoBERT; FinTech; Comparative Analysis; Natural Language Processing; E-wallet.

Abstract: The rapid adoption of e-wallet services in Indonesia has generated a large volume of user-generated reviews that provide valuable information about service quality, user satisfaction, and recurring technical problems. However, previous studies have commonly examined individual applications or emphasized classification performance, while cross-platform comparisons using a consistent Indonesian language model remain limited. This study conducts a comparative sentiment and exploratory complaint-term analysis of five widely used Indonesian e-wallet applications—DANA, OVO, GoPay, ShopeePay, and LinkAja—using IndoBERT. A total of 5,000 Google Play Store reviews were initially collected, of which 4,669 reviews remained after filtering and preprocessing. Ratings of 1–2 were used as negative proxy labels, whereas ratings of 4–5 were used as positive proxy labels, while three-star reviews were excluded. The dataset was divided into 80% training and 20% testing data using stratified sampling. The indobenchmark/indobert-base-p1 model was fine-tuned for three epochs using a learning rate of 2×10^{-5} and a batch size of 16. The best model achieved 84.36% accuracy and a positive-class F1-score of 73.45%. At the platform level, DANA recorded the highest proportion of positive reviews at 51.84%, whereas OVO showed the highest negative proportion at 84.77%. Exploratory complaint-term analysis identified recurring issues involving failed transactions, account balances, login and verification problems, application errors, and customer service. The findings demonstrate the usefulness of IndoBERT for analyzing Indonesian e-wallet reviews and provide a cross-platform perspective for identifying service improvement priorities in Indonesia's digital payment ecosystem.

[DOI: 10.14710/JCASE. vol2.iss1.1143]
OPEN ACCESS
CC BY SA

1. INTRODUCTION

The rapid development of information and communication technology has significantly transformed various sectors, including finance. One of the fastest-growing innovations is Financial Technology (fintech), which integrates financial services with digital technology to improve efficiency, accessibility, flexibility, and convenience in financial transactions [1]. In Indonesia, fintech growth has accelerated due to increasing internet penetration, smartphone usage, and the shift toward digital and cashless payments. Digital wallet (e-wallet) services have become an essential part of modern economic activities, including online payments, fund transfers, digital purchases, and e-commerce transactions [2].

Currently, e-wallet applications such as DANA, OVO, GoPay, ShopeePay, and LinkAja are among the most widely used digital payment platforms in Indonesia. These platforms compete to provide fast, secure, and user-friendly transaction services integrated with digital ecosystems such as online transportation, marketplaces, bill payment systems, and retail merchants. As competition intensifies, service quality and user satisfaction become critical factors in maintaining customer loyalty and competitiveness in the digital era [3].

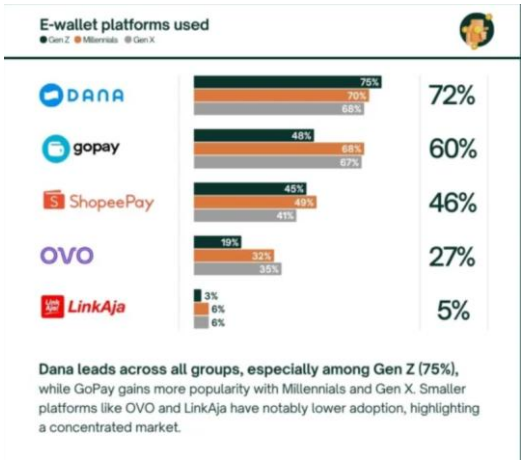


Fig 1. E-Wallet Platform Usage in Indonesia Across Generational Groups (Source: Jakpat, 2026)

According to the Jakpat survey shown in Figure 1, DANA recorded the highest usage rate at 72%, followed by GoPay (60%), ShopeePay (46%), OVO (27%), and LinkAja (5%). The survey also revealed that DANA dominates among Generation Z users, while GoPay is more popular among Millennials and Generation X. Meanwhile, ShopeePay has experienced significant growth through its integration with e-commerce platforms [4]. These findings indicate that the five selected e-wallets represent the major digital payment ecosystem in Indonesia and therefore serve as suitable research objects for analyzing fintech user perceptions.

As fintech adoption increases, the number of user reviews on digital platforms such as the Google Play Store has also grown significantly. User reviews contain valuable information reflecting customer experiences, satisfaction, complaints, expectations, and perceptions regarding fintech service quality [5]. Positive reviews are generally associated with transaction convenience, fast service, attractive promotions, and user-friendly interfaces. In contrast,

[DOI: 10.14710/JCASE. vol2.iss1.1143]

negative reviews often relate to failed transactions, application errors, login problems, delayed balances, and poor customer service.

However, the large volume of reviews makes manual analysis inefficient, time-consuming, and subjective. Therefore, automated approaches are needed to process and analyze textual data efficiently. One widely used approach is sentiment analysis, a branch of Natural Language Processing (NLP) used to identify opinions, emotions, or sentiment polarity within text, such as positive, negative, or neutral sentiment [6]. Through sentiment analysis, companies can systematically understand public perception toward products or services.

Previous sentiment analysis studies have largely relied on traditional machine learning methods such as Naive Bayes, Support Vector Machines (SVM), Decision Tree, and Random Forest [7], [8], [9], [10]. These methods are quite effective for simple text classification, but they still have limitations in understanding contextual relationships and semantic meanings between words because they rely heavily on feature extraction techniques such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) [11]. Such static feature-based approaches struggle to handle text with complex sentence structures, informal language, abbreviations, slang, and linguistic variations commonly found in user reviews of digital applications.

Recent advancements in deep learning and transformer architectures have significantly improved the performance of Natural Language Processing (NLP). One of the most influential transformer models is Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. [12]. BERT is capable of understanding bidirectional contextual relationships, thereby generating better semantic representations than traditional machine learning methods. This capability makes BERT superior in various NLP tasks such as text classification, machine translation, question answering, named entity recognition, and sentiment analysis [13].

In the context of the Indonesian language, IndoBERT has become one of the most widely used transformer models because it was specifically trained using a large-scale Indonesian language corpus [8]. IndoBERT demonstrates superior capabilities in understanding Indonesian language structure, semantic context, and variations in informal language compared to general multilingual models [14]. This architectural advantage has been proven effective in mapping customer feedback for financial applications, showing a significant accuracy leap over static word embeddings[15]. Previous research has shown that IndoBERT is capable of achieving higher sentiment classification performance compared to traditional machine learning methods due to its ability to capture deep contextual text representations [16].

Research by Ardiansyah et al. used the Naive Bayes and Support Vector Machine methods to analyze the sentiment of users of the GoPay digital wallet app. The results showed that the Naive Bayes method was capable of classifying sentiment quite well in user review data, although its performance was still inferior to more complex methods in understanding diverse text patterns [10]. In addition, Indriani and Wiranata compared several machine learning algorithms such as SVM, Decision Tree, and Random Forest in their sentiment analysis of GoPay app reviews. The study showed that the choice of algorithm affects the accuracy of sentiment classification in fintech app user review data[9]. Hidayatullah and Ma'arif found that the preprocessing stage significantly influenced the performance of Indonesian text classification, particularly in data containing informal language and

[DOI: 10.14710/JCASE. vol2.iss1.1143]

abbreviations [7]. Research by Koto et al. shows that a BERT-based transformer model specifically for Indonesian performs better than the TF-IDF and SVM methods because it is able to understand contextual relationships between words more effectively [8].

Although previous studies have investigated sentiment analysis of Indonesian application reviews using both traditional machine-learning algorithms and transformer-based models, much of the existing literature focuses on individual applications or primarily evaluates classification performance. Consequently, applying IndoBERT to a single e-wallet application is no longer sufficient by itself to constitute a strong methodological contribution.

A systematic comparison of user sentiment across several major Indonesian e-wallet platforms using the same data collection, labeling, preprocessing, and classification framework remains comparatively limited. Furthermore, classification performance alone does not explain the operational issues underlying negative user perceptions. Therefore, this study addresses this gap by combining IndoBERT-based sentiment classification with a standardized cross-platform comparison and exploratory analysis of complaint-related terms across five major Indonesian e-wallet applications.

The novelty of this study lies not merely in applying IndoBERT to e-wallet reviews, but in providing a consistent comparative framework for analyzing sentiment distributions and platform-specific complaint characteristics across DANA, OVO, GoPay, ShopeePay, and LinkAja. Based on the identified research gap, this study addresses the following research questions:

1. RQ1: How effectively can IndoBERT classify positive and negative sentiment in Indonesian e-wallet reviews labeled using Google Play Store ratings?
2. RQ2: How does the distribution of positive and negative sentiment differ across DANA, OVO, GoPay, ShopeePay, and LinkAja?
3. RQ3: What recurring complaint-related terms characterize negative user reviews for each e-wallet platform?

Based on these issues, this study aims to conduct a comparative sentiment analysis of Indonesian fintech application reviews using the IndoBERT model. The research utilizes user reviews from the Google Play Store for DANA, OVO, GoPay, ShopeePay, and LinkAja applications. The research process includes web scraping, text preprocessing, sentiment classification using IndoBERT, comparative analysis across applications, and visualization using charts and Wordclouds.

This study provides four main contributions:

1. It evaluates IndoBERT for binary sentiment classification of informal Indonesian e-wallet reviews.
2. It provides a standardized comparison of sentiment distributions across five major Indonesian e-wallet platforms using the same analytical framework.
3. It identifies recurring and platform-specific complaint-related terms from negative user reviews.
4. It translates user-review patterns into practical service-improvement priorities involving transaction reliability, account access, system stability, and customer support.

[DOI: 10.14710/JCASE. vol2.iss1.1143]

2. METHODOLOGY

This study employed a quantitative approach using Natural Language Processing (NLP) techniques to perform sentiment analysis on Indonesian fintech application reviews. The research process consisted of several stages, including data collection, sentiment labeling, text preprocessing, model training, evaluation, and comparative analysis. The overall methodology was designed to identify user perceptions toward fintech services and compare sentiment patterns across major Indonesian e-wallet applications.

The dataset was collected from the Google Play Store using the google-play-scraper library in Python. Five Indonesian e-wallet applications were selected as research objects, namely DANA, OVO, GoPay, ShopeePay, and LinkAja. Approximately 1,000 recent reviews were retrieved for each application, resulting in an initial dataset of 5,000 reviews. Indonesian language and regional parameters were used during data collection to obtain reviews relevant to Indonesian users.

For applications integrated into broader digital ecosystems, particularly GoPay and ShopeePay, keyword-based filtering was applied to retain reviews related to payment or e-wallet functionality rather than unrelated application features. Reviews containing fewer than three words were excluded because they provided insufficient textual information for sentiment analysis.

To improve dataset relevance, filtering techniques were applied during the data collection process. Applications integrated into broader digital ecosystems, such as GoPay and ShopeePay, required keyword-based filtering to ensure that the collected reviews specifically discussed e-wallet services rather than unrelated platform features. In addition, reviews containing fewer than three words were removed because they lacked sufficient semantic information for sentiment analysis.

This study employed supervised binary sentiment classification. Google Play Store ratings were used as proxy sentiment labels. Reviews with ratings of one or two stars were categorized as negative sentiment, while reviews with ratings of four or five stars were categorized as positive sentiment. Reviews with three-star ratings were excluded because their sentiment was considered comparatively ambiguous. Negative sentiment was encoded as label 0, while positive sentiment was encoded as label 1.

Rating-based labeling enables efficient annotation of large review datasets; however, these labels should be interpreted as proxy labels rather than manually validated sentiment annotations. A user rating may occasionally conflict with the textual opinion expressed in a review. For example, a user may provide a relatively high rating while still mentioning a specific complaint. Such rating–text inconsistencies may introduce label noise and are therefore considered a limitation of this study

Although rating-based labeling provides efficient large-scale annotation, it may not always perfectly represent the actual sentiment expressed within textual reviews. Some users may provide high ratings while still expressing complaints in the review content, which could introduce labeling inconsistencies.

Before the classification process, text preprocessing was conducted to reduce noise and improve text quality. User reviews from digital platforms generally contain slang expressions, abbreviations, symbols, emojis, repeated characters, and informal language structures that may affect NLP model performance. Therefore, several preprocessing techniques were applied,

[DOI: 10.14710/JCASE. vol2.iss1.1143]

including case folding, URL removal, mention and hashtag removal, number removal, symbol and emoji cleaning, repeated character normalization, slang word normalization, stopword removal, stemming, and short token filtering. Slang normalization was particularly important because Indonesian app reviews frequently contain informal expressions such as “gk”, “gabisa”, and “lemot”, which were converted into their standard forms. Stopword removal and stemming were performed using the Sastrawi library to improve semantic consistency.

Because IndoBERT is a contextual transformer model, aggressive preprocessing may remove linguistic information that could otherwise contribute to contextual representation. Therefore, the preprocessing procedures used in this study should be regarded as an experimental design choice rather than a universally optimal preprocessing strategy for transformer-based models. A dedicated ablation experiment would be required to determine the individual contribution of stemming, stopword removal, or slang normalization to model performance.

After preprocessing and the exclusion of three-star reviews, 4,669 reviews remained in the analytical dataset, consisting of 3,174 negative reviews and 1,495 positive reviews. The dataset was divided into 80% training data and 20% testing data using stratified sampling to preserve the original sentiment distribution across both subsets. This resulted in 3,735 training reviews and 934 testing reviews. Stratification was particularly important because the dataset was imbalanced toward the negative class.

IndoBERT was selected because it has been specifically pretrained on large-scale Indonesian corpora and has demonstrated superior contextual understanding for Indonesian NLP tasks, particularly for informal and slang-based text commonly found in user-generated reviews. Compared to traditional machine learning approaches, transformer-based architectures are more effective in capturing semantic relationships and contextual meaning within sentences.

The sentiment classification process used the indobenchmark/indobert-base-p1 model from the HuggingFace Transformers library. Textual data were tokenized using the corresponding IndoBERT tokenizer and converted into numerical representations consisting of input_ids and attention_mask. The maximum sequence length was set to 64 tokens. Longer sequences were truncated, while shorter sequences were padded to obtain uniform input dimensions.

Model fine-tuning was conducted using the HuggingFace Trainer framework. The learning rate was set to 2×10^{-5}, the batch size to 16, the number of training epochs to three, and the weight decay to 0.01. Model performance was evaluated after each epoch using accuracy, precision, recall, and F1-score. Because the dataset was imbalanced, class-specific precision, recall, and F1-score were considered together with overall accuracy.

After training and evaluation, the model was used to classify sentiment across the entire dataset. Comparative sentiment analysis was then conducted to examine user perception differences among fintech applications. The comparison process involved sentiment distribution analysis, cross-tabulation, and visualization using bar charts and Wordclouds. Wordcloud analysis was employed to identify dominant complaint patterns and frequently occurring terms in user reviews. Through this approach, the study aimed not only to evaluate sentiment polarity but also to uncover key service issues affecting user satisfaction within Indonesian fintech platforms.

[DOI: 10.14710/JCASE. vol2.iss1.1143]

In this study, the F1-score reported during model selection refers to the positive-class F1-score rather than a macro-averaged F1-score. Therefore, accuracy and class-specific evaluation metrics are reported together to provide a more complete interpretation of model performance.

3. RESULTS AND DISCUSSION

3.1 Dataset Overview

This study collected 5,000 user reviews from five major Indonesian e-wallet applications: DANA, OVO, GoPay, ShopeePay, and LinkAja from the Google Play Store using Python-based web scraping and keyword filtering. After preprocessing and removing neutral reviews, 4,669 reviews were retained for analysis.

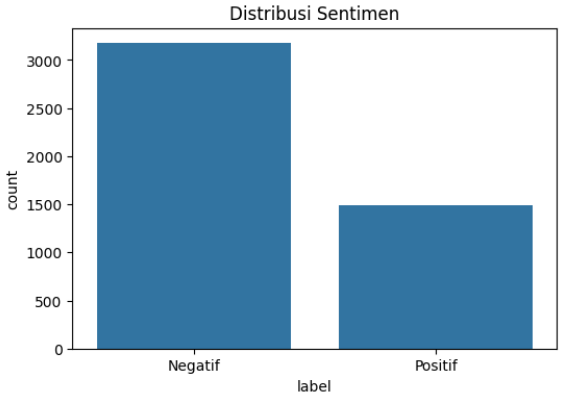
The dataset contains diverse opinions ranging from positive experiences about transaction convenience to complaints regarding transaction failures and application errors. Initial distribution analysis shows that negative sentiment dominated the dataset, with 3,174 negative reviews and 1,495 positive reviews. The predominance of negative reviews indicates that dissatisfaction and technical problems were strongly represented in the collected dataset. However, this distribution should be interpreted as a characteristic of the study sample and does not by itself establish that users in general are more likely to submit reviews after negative experiences.

3.2 Data Preprocessing Results

Preprocessing was conducted to remove noise such as URLs, symbols, numbers, emojis, repeated characters, and informal language commonly found in Google Play Store reviews. Slangword normalization was also applied to convert non-standard words into standard Indonesian. The preprocessing procedure produced a more standardized textual representation by reducing noise and lexical variation. However, the present study did not conduct an ablation experiment to measure the independent effect of each preprocessing operation. Therefore, no causal conclusion can be made regarding the contribution of individual preprocessing steps to IndoBERT performance.

3.3 Visualisasi Distribusi Sentimen

Figure 2 illustrates the distribution of positive and negative sentiments in the dataset. Negative reviews clearly outnumber positive reviews, indicating that users tend to provide feedback when facing issues such as failed transactions, login problems, or system errors.



[DOI: 10.14710/JCASE. vol2.iss1.1143]

Fig. 2. Sentiment Distribution of Fintech Dataset

This finding suggests that negative experiences trigger stronger emotional responses compared to positive experiences, which is common in digital platform reviews.

3.4 Sentiment Wordcloud Analysis

Figure 3 presents the positive sentiment Wordcloud. Frequently occurring words include "mudah" (easy), "bagus" (good), "cepat" (fast), "praktis" (practical), and "membantu" (helpful), indicating that users appreciate transaction efficiency and convenience.



Fig. 3. Positive Sentiment Wordcloud

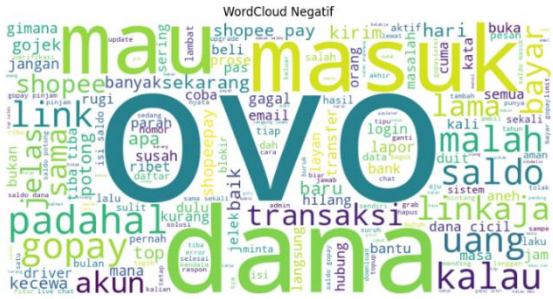


Fig. 4. Negative Sentiment Wordcloud

Figure 4 presents the negative sentiment Wordcloud, where dominant words include "error," "gagal" (failed), "saldo" (balance), "login," "verifikasi" (verification), and "kecewa" (disappointed). The results show that technical issues, failed transactions, account verification, and system stability remain the primary concerns among fintech users.

3.5 IndoBERT Model Training Results

This study employed the indobenchmark/indobert-base-p1 model for sentiment classification. The dataset was divided into 80% training data (3,735 reviews) and 20% testing data (934 reviews). Training was conducted for three epochs using a learning rate of 2e-5 and batch size of 16.

Table 1. IndoBERT Training Results

Epoch	Validation Loss	Accuracy	F1-Score
1	0.3887	0.8329	0.6917
2	0.4188	0.8436	0.7345
3	0.4894	0.8361	0.7320

The highest classification performance was obtained at Epoch 2, with an accuracy of 84.36% and a positive-class F1-score of 73.45%. Although Epoch 1 produced a lower validation loss, both accuracy and positive-class F1 improved at Epoch 2. At Epoch 3, validation loss increased to 0.4894 while accuracy and F1-score slightly declined. This pattern suggests the possible onset of overfitting after the second epoch. Therefore, the Epoch 2 checkpoint was selected as the best-performing model based on classification performance.

[DOI: 10.14710/JCASE. vol2.iss1.1143]

3.6 Evaluasi Model

Table 2. Classification Report

Kelas	Precision	Recall	F1-Score
Negatif	0.86	0.92	0.89
Positif	0.80	0.68	0.73

The model achieved excellent performance in detecting negative sentiment, with a recall score of 92%. However, positive sentiment detection remained lower due to class imbalance.

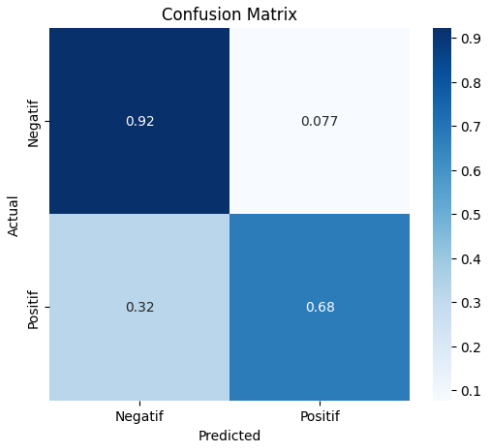


Fig. 5. Confusion Matrix Model IndoBERT

The model showed stronger performance in identifying negative sentiment than positive sentiment. Negative reviews achieved a precision of 0.86, recall of 0.92, and F1-score of 0.89. In contrast, positive reviews achieved a precision of 0.80, recall of 0.68, and F1-score of 0.73. The lower recall for the positive class indicates that a considerable proportion of actual positive reviews were incorrectly classified.

One plausible contributing factor is the class imbalance in the dataset, in which negative reviews substantially outnumbered positive reviews. Consequently, the overall accuracy of 84.36% should not be interpreted independently but should be considered together with class-level precision, recall, and F1-score.

3.7 Comparative Analysis Among Fintech Applications

Table 3. Sentiment Comparison Across Applications

Aplikasi	Negatif	Positif
DANA	444	478
GoPay	756	178
LinkAja	745	200
OVO	802	144
ShopeePay	579	343

[DOI: 10.14710/JCASE. vol2.iss1.1143]

DANA recorded the highest number of positive reviews, while OVO had the highest negative sentiment count.

3.8 Sentiment Percentage Across Applications

Table 4. Sentiment Percentage Across Applications

Application	Negative	Positive
DANA	48.15%	51.84%
GoPay	80.94%	19.05%
LinkAja	78.83%	21.16%
OVO	84.77%	15.22%
ShopeePay	62.79%	37.20%

The sentiment distributions differed substantially across the five applications. DANA was the only platform in the dataset for which positive reviews slightly exceeded negative reviews, with 51.84% positive and 48.15% negative sentiment. In contrast, OVO recorded the highest proportion of negative reviews at 84.77%, followed by GoPay at 80.94% and LinkAja at 78.83%. These results should be interpreted as sentiment distributions within the collected Google Play Store sample rather than as direct measurements of the overall customer satisfaction level of each platform. Differences in review timing, application versions, user characteristics, and reviewing behavior may affect the observed proportions.

3.9 Negative Wordcloud Analysis by Application

3.9.1 DANA

Common complaint keywords include “saldo,” “uang,” “hilang,” and “limit,” indicating issues related to balance security, failed transactions, and account verification.



Fig. 6. Negative Wordcloud of DANA



Fig. 7. Negative Wordcloud of OVO

3.9.2 OVO

Dominant words such as “top up,” “gagal,” “login,” and “error” indicate major problems involving top-up transactions, failed payments, and application instability.

3.9.3 GoPay

The word “driver” appeared prominently, showing that many complaints were related to integration with Gojek services such as GoRide and GoFood. Other concerns included blocked accounts and failed transfers.

[DOI: 10.14710/JCASE. vol2.iss1.1143]

J-CaSE
e-ISSN 3110-5777



Fig. 8. Negative Wordcloud of GoPay



Fig. 9. Negative Wordcloud of ShopeePay

3.9.4 ShopeePay

Negative sentiment was strongly associated with vouchers, promotions, payment failures, and delayed balance updates within the e-commerce ecosystem.

3.9.5 LinkAja



Fig. 10. Negative Wordcloud of Linkaja

Frequent keywords included “saldo,” “login,” “potong,” and “live chat,” indicating issues related to balance deductions, login problems, failed transactions, and slow customer service response.

3.9.6 Comparison of Complaint Patterns

Although each application showed different complaint characteristics, common keywords such as “saldo,” “akun,” “bayar,” “transaksi,” and “error” appeared across all platforms. This indicates that user satisfaction in fintech services is strongly influenced by transaction stability, account security, payment reliability, application performance, and customer service quality. The findings demonstrate that sentiment analysis can effectively identify major service issues and provide strategic insights for fintech companies to improve user experience. The Wordcloud results should be interpreted as an exploratory representation of frequently occurring complaint-related vocabulary rather than as statistically validated complaint topics. Word frequency alone does not measure the severity, prevalence, or causal importance of a service problem. Nevertheless, the visualization provides an initial indication of recurring issues that can guide more detailed topic or aspect-based analyses in future research.

3.10 Discussion

This study produced two principal findings. First, IndoBERT achieved an accuracy of 84.36% for binary sentiment classification, with stronger performance in identifying negative

[DOI: 10.14710/JCASE. vol2.iss1.1143]
OPEN ACCESS
CC BY SA

sentiment than positive sentiment. Second, sentiment distributions and complaint-related vocabulary differed substantially across the five e-wallet applications.

The classification results are consistent with previous studies demonstrating the usefulness of contextual language models for Indonesian Natural Language Processing. IndoBERT's ability to represent contextual relationships makes it suitable for processing informal Indonesian reviews containing abbreviations, slang, and non-standard expressions. However, the performance obtained in this study should not be directly compared with accuracy values reported in other studies without considering differences in datasets, labeling procedures, preprocessing, sentiment classes, and evaluation protocols.

Unlike previous research that has focused on individual e-wallet applications or primarily evaluated classification algorithms, the present study applies one consistent analytical framework across five e-wallet platforms. Therefore, the principal contribution of this study is not a claim that IndoBERT universally outperforms conventional machine-learning algorithms. Because no controlled baseline comparison with SVM, Naive Bayes, Random Forest, or other transformer models was conducted on the same dataset, such a superiority claim cannot be established from the present experiment.

The cross-platform findings indicate that negative user experiences differ according to each application's service ecosystem. OVO reviews were strongly associated with top-up, login, and system-error vocabulary, whereas LinkAja complaints frequently involved balances, login problems, and customer-service interactions. ShopeePay complaints reflected payment processes, vouchers, and its integration with an e-commerce ecosystem. GoPay reviews also contained complaints connected with the broader Gojek service ecosystem.

DANA differed from the other platforms because positive sentiment slightly exceeded negative sentiment. Nevertheless, this result should not be interpreted as definitive evidence that DANA has higher overall customer satisfaction than all competing e-wallet platforms. It reflects only the distribution observed in the Google Play Store sample collected for this study.

From a practical perspective, the results indicate that transaction reliability, account accessibility, application stability, and customer-support responsiveness are important service-quality dimensions reflected in user feedback. Consequently, fintech providers should not rely solely on aggregate sentiment scores but should also examine platform-specific complaint characteristics when prioritizing service improvements.

3.11 Limitations

This study has several limitations. First, the dataset was collected exclusively from the Google Play Store; therefore, the findings may not represent users on the Apple App Store, social media, or other feedback channels.

Second, Google Play Store ratings were used as proxy sentiment labels rather than manually validated sentiment annotations. Rating–text inconsistencies may therefore introduce label noise into the classification process.

Third, the dataset was imbalanced, with substantially more negative than positive reviews. This imbalance may have contributed to the lower recall observed for positive sentiment.

[DOI: 10.14710/JCASE. vol2.iss1.1143]

Fourth, complaint analysis primarily relied on Wordcloud visualization and frequent-term inspection. Although useful for exploratory analysis, this approach does not quantify topic prevalence or model semantic relationships among complaints.

Fifth, this study did not conduct a controlled comparison between IndoBERT and conventional machine-learning algorithms or other transformer models using the same dataset. Therefore, the findings demonstrate the effectiveness of IndoBERT for the present dataset but do not establish its statistical superiority over competing methods.

4. CONCLUSION

This study conducted a comparative sentiment and complaint-term analysis of Google Play Store reviews from five Indonesian e-wallet applications—DANA, OVO, GoPay, ShopeePay, and LinkAja—using IndoBERT.

Of the 5,000 reviews initially collected, 4,669 remained after filtering and preprocessing. The indobenchmark/indobert-base-p1 model achieved its best performance at Epoch 2, with an accuracy of 84.36% and a positive-class F1-score of 73.45%. Class-level evaluation showed stronger detection of negative sentiment, which achieved a recall of 92%, compared with 68% recall for positive sentiment.

Cross-platform analysis showed substantial differences in the collected review distributions. DANA recorded the highest proportion of positive sentiment at 51.84%, whereas OVO recorded the highest proportion of negative sentiment at 84.77%. Exploratory complaint-term analysis further revealed recurring concerns involving transaction failures, account balances, login and verification problems, application errors, and customer-service issues.

The main contribution of this study lies in applying a consistent analytical framework to compare multiple Indonesian e-wallet applications rather than analyzing an individual platform in isolation. The findings demonstrate that sentiment classification can be complemented by platform-level complaint analysis to provide more actionable insights into user experience.

Nevertheless, the findings should be interpreted within the limitations of Google Play Store sampling, rating-based proxy labels, class imbalance, and exploratory Wordcloud analysis. Future research should incorporate manually validated sentiment labels, multi-platform datasets, data-balancing or class-weighting strategies, quantitative topic-modeling techniques, and controlled comparisons between IndoBERT and alternative classification models.

REFERENCES

[1] H. Kusuma and W. K. Asmoro, “Perkembangan Financial Technologi (Fintech) Berdasarkan Prespektif Ekonomi Islam,” *Istithmar J. Studi Ekon. Syariah*, vol. 4, no. 2, Dec. 2020, doi: 10.30762/istithmar.v4i2.14.

[2] A. A. Pratiwi, N. Aulia, N. Nikmah, D. Indriani, and K. Audy, “Perkembangan dan Implementasi Digital Payment dalam Mendukung Sistem Pembayaran Cashless di Indonesia,” vol. 3, no. 4, 2026.

[3] A. Koswara, “Persaingan Dompet Digital di Indonesia: Analisis Tren Google dan Machine Learning pada GoPay, OVO, dan DANA,” 2025.

[4] Jakpat, “E-Wallet Consumer Patterns by Generation,” Jakpat Insight. Accessed: May 17, 2026. [Online]. Available: https://insight.jakpat.net/e-wallet-consumer-patterns-by-generation/

[5] T. Victory and D. D. Wazaumi, "Analisis Sentimen Ulasan Pengguna pada Aplikasi Tunaiku dengan Pendekatan Machine Learning," *remik*, vol. 10, no. 1, pp. 315–324, Jan. 2026, doi: 10.33395/remik.v10i1.15839.
[6] M. A. Z. Abidin and A. Mulyani, "Evaluasi Sentimen Ulasan CGV Cinemas Indonesia dengan Naïve Bayes dan SVM," *Indones. J. Softw. Eng. IJSE*, vol. 11, no. 1, pp. 44–53, 2025.
[7] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *J. Phys. Conf. Ser.*, vol. 801, p. 012072, Jan. 2017, doi: 10.1088/1742-6596/801/1/012072.
[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.
[9] I. Indriani and Ade Davy Wiranata, "Comparison of Accuracy Levels of SVM, Decision Tree and Random Forest Algorithms in Sentiment Analysis of User Responses of The Gopay Application," *J. Tek. Inform. Jutif*, vol. 5, no. 3, pp. 777–787, May 2024, doi: 10.52436/1.jutif.2024.5.3.1885.
[10] W. Andriyani, Y. Astuti, B. A. Wisesa, and H. Hengki, "Analisis Sentimen pada Ulasan Produk dengan SVM dan Word2Vec," *JIKO J. Inform. Dan Komput.*, vol. 9, no. 1, p. 173, Feb. 2025, doi: 10.26798/jiko.v9i1.1498.
[11] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1.
[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
[13] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, Dec. 2020, doi: 10.1162/tacl_a_00349.
[14] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[15] F. R. Andhika, W. Witanti, and P. N. Sabrina, "Analisis Sentimen Menggunakan Metode IndoBERT pada Ulasan Aplikasi Zoom Menggunakan Fitur Ekstrasi GloVe," *Metik J.*, vol. 9, no. 2, 2025, doi: 10.47002/metik.v9i2.1098.