Vol. 8, No. 2, April 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Time-Series Monitoring of Sentiment Dynamics in Reviews of Four Indonesian E-Wallet Applications Using a Hybrid TF-IDF and Bi-LSTM Framework

Noor Latifah[1], Dias Henandra Eka Putra[2], Fajar Nugraha[3]

[1,2,3]Department of Information Systems, Muria Kudus University, Kudus, Central Java, Indonesia

Received:
September 5, 2025
Revised:
March 10, 2026
Accepted:
April 5, 2026
Published:
April 12, 2026

Corresponding Author:
Author Name*:
Noor Latifah
Email*:
noor.latifah@umk.ac.id

DOI:
10.63158/journalisi.v8i2.1488

Abstract. This study proposes a hybrid sentiment analysis framework to examine user perceptions of four Indonesian e-wallet applications using Google Play Store reviews. The framework combines TF-IDF features reduced through Truncated SVD with a Bidirectional Long Short-Term Memory (Bi-LSTM) model within a two-stage evaluation design consisting of holdout classification and external temporal inference. For supervised classification, 20,000 raw reviews were filtered and labeled using a rating-based strategy, resulting in 13,823 labeled reviews. Reviews with ratings of 4–5 stars were assigned to the positive class and 1–2 stars to the negative class; these labels should be interpreted as sentiment proxies rather than fully human-validated ground truth. A second dataset of 24,000 reviews was constructed for balanced cross-application temporal comparison across 2024–2026. On the holdout test set, the proposed model achieved an accuracy of 0.881, with macro-F1 and weighted-F1 scores of 0.881. Under the external temporal setting, DANA remained relatively stable, GoPay improved markedly in 2025 and remained high in 2026, ShopeePay showed a gradual decline, and OVO exhibited the strongest negative trend. These results indicate that the proposed framework is useful not only for supervised sentiment classification but also for structured temporal monitoring across e-wallet platforms.

Keywords: app review mining, Bi-LSTM, Google Play Store reviews, Indonesian e-wallet, temporal sentiment monitoring, TF-IDF.

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

1. INTRODUCTION

The rapid growth of the digital economy has accelerated the adoption of mobile-based financial services, making e-wallet applications an important part of everyday digital transactions in Indonesia [1]. Along with this growth, user expectations regarding transaction reliability, interface usability, security, feature performance, and service responsiveness have also increased. Users frequently express their unmet expectations through public review platforms such as the Google Play Store. As a result, user reviews have become an important source of real-world feedback for understanding how digital payment applications are perceived in practical usage contexts [2].

Google Play Store reviews are valuable because they contain large volumes of spontaneous textual feedback written by users based on actual experiences. However, these reviews are unstructured, noisy, and difficult to evaluate manually at scale. Prior studies have shown that review data can support application evaluation, user feedback analysis, and service improvement when processed systematically [2], [3]. In this context, sentiment analysis provides an effective analytical approach for transforming raw review text into structured indicators of positive and negative user perception, thereby supporting more data-driven decision-making for developers and digital service stakeholders [3].

Recent studies have applied various sentiment analysis approaches to user-generated reviews, ranging from conventional machine learning to deep learning and hybrid architectures [3]. In the financial technology context, [4] demonstrated that customer reviews can be used to analyze sentiment patterns across e-wallet companies and reveal user perceptions through review-based evidence. In the Indonesian context, local studies on OVO and GoPay application reviews further confirm that sentiment analysis is useful for identifying user opinions in digital wallet services [5], [6]. However, these studies still primarily focus on classification performance, algorithm comparison, or platform-specific sentiment profiling.

Despite these advances, an important gap remains in the literature. Most previous studies still treat app reviews as a relatively static corpus and primarily evaluate sentiment classification models using random train-test partitioning, algorithm comparison, or

platform-specific classification settings [4]–[6], [8], [9]. While such settings are useful for benchmark evaluation, they are less appropriate for real-world monitoring because user sentiment may shift over time due to application updates, transaction problems, policy changes, promotional campaigns, and evolving service expectations [9]. In the context of digital financial services, review-based sentiment analysis has been shown to provide meaningful insights for service evaluation and user-opinion monitoring, which further highlights the need for evaluation designs that remain informative under changing user conditions [8], [10].

Based on this gap, this study proposes a hybrid sentiment analysis framework for reviews of four Indonesian e-wallet applications by combining TF-IDF and Bidirectional Long Short-Term Memory (Bi-LSTM) within a two-stage evaluation design. The first stage applies holdout evaluation to measure supervised sentiment classification performance, while the second stage uses a balanced external dataset for structured temporal comparison across applications and years [9]. Therefore, the main contribution of this study lies in extending app-review sentiment analysis from static classification toward structured cross-application temporal sentiment monitoring through a two-stage evaluation design in the Indonesian e-wallet context [8], [9].

2. METHODS

This study proposes a hybrid sentiment analysis framework for Indonesian e-wallet user reviews collected from the Google Play Store. The framework integrates statistical lexical representation using Term Frequency–Inverse Document Frequency (TF-IDF) with sequential contextual modeling using Bidirectional Long Short-Term Memory (Bi-LSTM). In addition to supervised sentiment classification, the study incorporates a temporally structured evaluation setting to compare sentiment dynamics across applications during the 2024–2026 period. Accordingly, the research design consists of two complementary stages, namely holdout evaluation for supervised sentiment classification and external temporal inference for cross-application sentiment monitoring [9], [10]. The overall workflow of the study is presented in Figure 1.

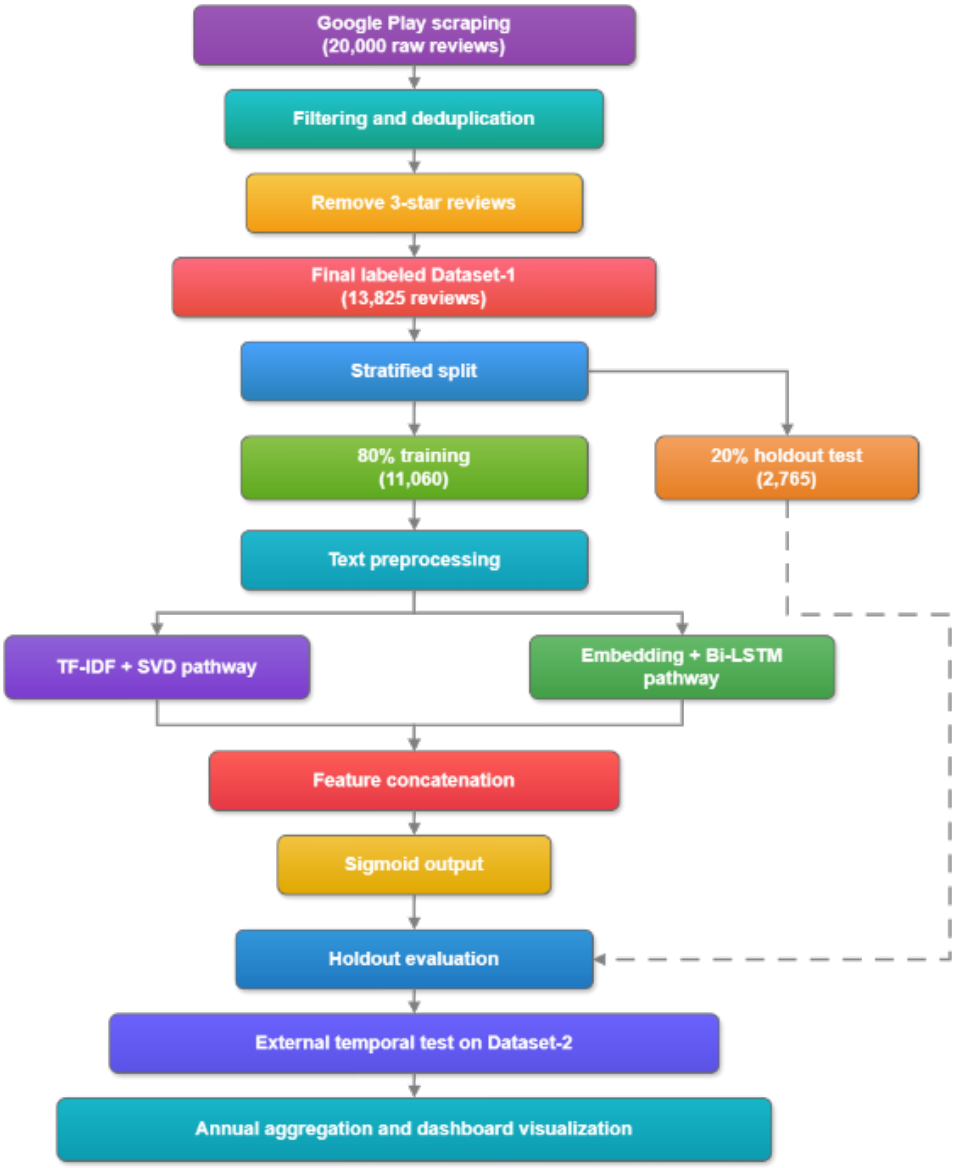

Figure 1. Research Method Flow

Table 1. Summary of the experimental workflow

Stage	Dataset	Main Purpose	Main Procedure	Output
Stage 1	Dataset-1	Supervised sentiment classification	Filtering, preprocessing, rating-based labeling, stratified holdout split, model training and holdout testing	Supervised classification performance

Stage	Dataset	Main Purpose	Main Procedure	Output
Stage 2	Dataset-2	External temporal monitoring	Filtering, duplicate control, balanced sampling by application-year, model inference	Yearly comparative sentiment indicators
Baseline comparison	Dataset-1	Model benchmarking	Same holdout setting applied to TF-IDF + Logistic Regression and Bi-LSTM-only	Comparative evaluation across modeling strategies

2.1 Tools and Libraries

The proposed hybrid sentiment analysis system was implemented in Python. Data scraping from the Google Play Store was conducted using the google-play-scraper library. Dataset manipulation and preprocessing were performed using pandas and NumPy. Statistical text feature extraction based on TF-IDF was implemented using scikit-learn. The deep learning architecture, including tokenization and Bidirectional LSTM modeling, was developed using TensorFlow/Keras. In addition, sentiment classification outputs and comparative monitoring results were presented through an interactive evaluation dashboard built with Streamlit.

2.2 Dataset Construction and Data Collection

Review data were collected from the Google Play Store through a Python-based scraping process for app-review mining [12], [8]. The data collection process focused on Indonesian market reviews (locale='id') from the Google Play Store. The reviews analyzed in this study were posted during 2024–2026, and the scraping process was conducted in January 2026. To maintain linguistic consistency during preprocessing and classification, the analysis was restricted to Indonesian-language user reviews. Each review record contained at least the review text, star rating, review date, and application identity. The study focused on four Indonesian e-wallet applications, namely DANA, GoPay, OVO, and ShopeePay.

To support the two-stage evaluation design, two datasets were constructed. Dataset-1 was used for supervised model development and holdout evaluation. The initial raw collection for Dataset-1 consisted of 20,000 reviews. Dataset-2 was constructed

separately for external temporal inference and cross-application comparison. This dataset covered the 2024–2026 period and was balanced with 2,000 reviews per application per year, resulting in 24,000 reviews in total. After filtering and duplicate removal, reviews in Dataset-2 were grouped by application and year, and 2,000 reviews were randomly selected without replacement for each application-year cell.

Dataset-1 and Dataset-2 were kept separate to preserve experimental independence. Cross-dataset duplicate cleaning was conducted using review text, review date, and application identity [13], [14]. For Dataset-1, the filtering procedure followed the exact order implemented in the original preprocessing script. Starting from 20,000 raw reviews, 818 reviews with a 3-star rating were first removed to avoid neutral-label ambiguity in binary sentiment learning, leaving 19,182 reviews. Next, 748 invalid or non-informative reviews were excluded after text cleaning, resulting in 18,434 reviews. Finally, 4,611 duplicate reviews with identical cleaned textual content were removed, leaving 13,823 labeled reviews for supervised classification. This procedure was intended to preserve more organic review patterns and to reduce redundancy in the supervised learning stage. A summary of the dataset structure is shown in Table 2, while the temporal sampling design is illustrated in Figure 2.

To improve methodological transparency, duplicate handling was applied at two levels. Within Dataset-1, duplicate reviews were identified when two records produced exactly identical cleaned textual content after preprocessing. Across Dataset-1 and Dataset-2, independence checking additionally considered the combination of review text, review date, and application identity in order to minimize cross-dataset overlap [13], [14]. For Dataset-2, temporal balancing was conducted only after filtering and duplicate removal, by randomly selecting 2,000 cleaned reviews without replacement for each application-year cell. This procedure was intended to improve comparability across applications and years while reducing distortion caused by unequal review volume.

Table 2. Dataset summary

Dataset	Purpose	Period	Applications	Sampling Rule	Total
Dataset-1	Training and holdout evaluation	Mixed collection period	DANA, GoPay, OVO, ShopeePay	Raw collection of 20,000 reviews; 818 3-star reviews removed; 748 invalid/short	13,823 final

Dataset	Purpose	Period	Applications	Sampling Rule	Total
				reviews removed; 4,611 cleaned-text duplicates removed; final 13,823 labeled reviews; stratified 80:20 split	labeled reviews
Dataset-2	External temporal comparison	2024–2026	DANA, GoPay, OVO, ShopeePay	2,000 reviews per application per year	24,000 reviews

Figure 2. Balanced external temporal sampling design

2.3 Data Filtering and Rating-Based Sentiment Labeling

Before model development, Dataset-1 underwent a staged filtering procedure to remove ambiguous labels, invalid textual entries, and duplicate content [15]. The filtering order followed the original preprocessing script. From the initial 20,000 raw reviews, 818 reviews with a 3-star rating were first removed to avoid neutral-label ambiguity in binary sentiment learning, leaving 19,182 reviews. In the second stage, 748 invalid or non-informative reviews were excluded after regex-based cleaning. These included entries with no interpretable semantic content, such as symbol-dominated text, repetitive fragments, or cleaned text with a remaining length of two characters or fewer. This step

reduced the dataset to 18,434 reviews. In the final stage, 4,611 duplicate reviews with identical cleaned textual content were removed, leaving exactly 13,823 labeled reviews for supervised classification. Dataset-1 was labeled using a rating-based sentiment strategy. Reviews with ratings of 4–5 stars were assigned to the positive class, whereas reviews with ratings of 1–2 stars were assigned to the negative class [15]. This strategy enabled large-scale supervised learning without full manual annotation. However, the resulting labels should be interpreted as rating-derived sentiment proxies rather than fully human-validated ground truth [15].

2.4 Text Preprocessing

After filtering and label construction, the review texts underwent preprocessing to reduce noise and standardize the textual input before feature extraction. First, cleansing was applied to remove URLs, numbers, selected punctuation, emojis, and excessive whitespace. Second, case folding converted all characters into lowercase form. Third, tokenization segmented each review into word-level units. Fourth, token normalization and stopword removal were conducted to standardize informal expressions and eliminate low-information words [16],[17].

During text preparation, web addresses (URLs), emojis, and user mentions were removed using Regular Expression (Regex)-based cleansing. The cleaned text was then mapped into a custom lexical normalization dictionary to unify informal and non-standard expressions. For example, lexical variants referring to slow performance, such as *lemot*, *lelet*, *delay*, *pending*, and *nge-lag*, were normalized into the unified feature form *lambat*. In addition, contrast markers such as tapi were preserved using the token <BUT>, and negation expressions were retained using _NEG to maintain sentiment-shifting cues [18]. The resulting corpus was then used as the common input for both the TF-IDF pathway and the Bi-LSTM pathway. Examples of review text transformation are presented in Table 3.

Table 3. Examples of review text transformation across preprocessing stages (Indonesian Text)

Raw Review	Cleansed Text	Lowercased Text	Tokenized Text	Final Text
"Aplikasinya bagus!!! tapi kadang error 😭"	*"Aplikasinya bagus tapi kadang error"*	*"aplikasinya bagus tapi kadang error"*	*[aplikasinya, bagus, tapi, kadang, error]*	*"aplikasinya bagus <BUT> kadang error"*

Raw Review	Cleansed Text	Lowercased Text	Tokenized Text	Final Text
"Transfer lama bgt!!! saldo kepotong, tapi uang belum masuk."	*"Transfer lama bgt saldo kepotong tapi uang belum masuk"*	*"transfer lama bgt saldo kepotong tapi uang belum masuk"*	*[transfer, lambat, bgt, saldo, kepotong, tapi, uang, tidak, masuk]*	*"transfer lambat bgt saldo kepotong <BUT> uang tidak masuk_NEG"*
"Promo banyak, tp pas bayar QR tdk bisa dipakai."	*"Promo banyak tp pas bayar QR tdk bisa dipakai"*	*"promo banyak tp pas bayar qr tdk bisa dipakai"*	*[promo, banyak, tapi, bayar, qr, tidak, dipakai]*	*"promo banyak <BUT> bayar qr tidak dipakai_NEG"*

2.5 Proposed Hybrid TF-IDF – Bi-LSTM Framework

The proposed framework integrates two complementary representation pathways. The first pathway is based on Bi-LSTM and is intended to capture contextual dependencies in review text from both forward and backward directions [19]. The second pathway is based on TF-IDF and is intended to capture discriminative lexical importance across the corpus [16]. In the sequential pathway, preprocessed reviews were transformed into token sequences and converted into fixed-length inputs using padding or truncation. The vocabulary size was limited to the most frequent 12,000 words, and the maximum sequence length was set to 120 tokens. These sequences were passed through an embedding layer with a dimension of 128 and then processed by a Bidirectional LSTM layer with 64 units to produce contextual sentence-level representations [19].

In the statistical pathway, the same preprocessed reviews were represented using TF-IDF vectors. Lexical features were extracted using TfidfVectorizer with an n-gram range of (1,2). The maximum vocabulary size was limited to 12,000 features in order to retain the most informative lexical patterns while controlling dimensional growth. Because TF-IDF features are sparse and high-dimensional, Truncated Singular Value Decomposition (SVD) was applied to reduce the feature space to 128 dimensions while retaining the dominant latent structure [19]. The outputs from the Embedding–Bi-LSTM pathway and the TF-IDF–SVD pathway were then concatenated and passed to a dense output layer with sigmoid activation for binary sentiment classification [19]. A review was classified as positive when the predicted probability was equal to or greater than 0.5, and negative otherwise.

The methodological contribution of this study lies in integrating statistical lexical representation and sequential contextual representation within a two-stage evaluation design that includes both holdout testing and external temporal generalization. The proposed architecture is shown in Figure 3, and the training configuration is summarized in Table 4.

Figure 3. Proposed hybrid model architecture

Table 4. Hyperparameter configuration of the proposed model

Parameter	Value
Max Words	12,000
Max Sequence Length	120
Embedding Dimension	128
Bi-LSTM Units	64
TF-IDF SVD Dimension	128
Dropout Rate	0.3
Batch Size	32
Epochs	Maximum 20 with early stopping (validation ratio = 0.10)

Parameter	Value
Learning Rate	0.001
Optimizer	Adam
Output Activation	Sigmoid
Classification Threshold	0.5

2.6 Holdout Evaluation Protocol

The first evaluation stage used Dataset-1 to measure sentiment classification performance under a supervised holdout setting. After the filtering and labeling stages were completed, the cleaned corpus of 13,823 reviews was organized as Dataset-1 for supervised historical sentiment classification. The dataset was divided proportionally into 11,058 reviews (80%) for training and 2,765 reviews (20%) for testing using a stratified holdout strategy. This split was designed to preserve class balance between positive and negative labels during model development and final evaluation [15].

Model training was conducted on the training partition. During training, 10% of the 11,058 training reviews were reserved internally as a validation subset, yielding 1,106 validation reviews and 9,954 reviews for model fitting. This validation subset was used to monitor performance at each epoch and to activate early stopping when no further improvement was observed. Model training was monitored using the Area Under the Precision-Recall Curve (PR-AUC) rather than relying only on standard loss values. To mitigate overfitting, an EarlyStopping strategy was applied by monitoring validation PR-AUC (val_prc) with a patience of 3 epochs. In parallel, ReduceLROnPlateau was used to reduce the learning rate by 50% (factor=0.5) after 2 epochs of stagnant validation improvement. The best-performing model parameters were restored automatically (restore_best_weights=True) and retained as the final .keras checkpoint for evaluation. The best model checkpoint was then evaluated once on the holdout test set. To prevent information leakage, data-dependent transformations such as vocabulary construction, token indexing, TF-IDF fitting, and SVD fitting were derived only from the training partition and then applied unchanged to the holdout test set [20]. In addition to the proposed framework, comparative evaluation was also conducted using selected baseline models, namely TF-IDF + Logistic Regression and Bi-LSTM-only configurations, under the same holdout partition.

The selected baselines were intended to represent the two principal modeling perspectives underlying the proposed framework. TF-IDF + Logistic Regression was chosen as a strong lexical-statistical baseline because it is widely recognized as effective for short-text sentiment classification and remains competitive under noisy review conditions [15]. Bi-LSTM-only was selected to represent a purely sequential contextual approach without explicit TF-IDF-based lexical weighting [19], [20]. By comparing the proposed hybrid framework with these two baselines under the same holdout partition, the study can more transparently assess whether the integration of lexical-statistical and sequential contextual representations provides added value beyond single-path modeling strategies.

2.7 External Temporal Inference and Cross-Application Comparison

After the best-performing model had been obtained from Dataset-1, it was applied to Dataset-2 as an external temporal inference setting for cross-application sentiment monitoring. Dataset-2 was organized into balanced application-year cells, with 2,000 reviews for each application in each year from 2024 to 2026. The balanced application-year structure was formed after filtering and duplicate removal so that each cell represented a comparable set of cleaned reviews for temporal inference.

Dataset-2 was used exclusively as an inference-oriented case study dataset for comparative dashboard analysis across temporal periods. The reviews contained in Dataset-2 were never introduced into the training process of the Hybrid TF-IDF-BiLSTM model. Therefore, all supervised performance metrics, including Accuracy, Macro F1-Score, and PR-AUC, were calculated solely on the independent testing set of 2,765 reviews from Dataset-1, while Dataset-2 was used only for external classification inference and cross-application temporal comparison.

For each application-year cell, the predicted labels were aggregated to compute the annual proportion of positive sentiment. This aggregation generated a comparable temporal sentiment indicator for DANA, GoPay, OVO, and ShopeePay. Because the dataset was balanced by design, the comparison was less affected by unequal review volume and more suitable for interpreting relative sentiment changes across applications and years [17]. The resulting yearly sentiment proportions were then visualized using line charts and grouped comparison plots.

2.8 Performance Metrics and Implementation Environment

Performance on Dataset-1 was evaluated using a confusion matrix and standard classification metrics, namely accuracy, precision, recall, and F1-score [15]. Macro-average and weighted-average F1-scores were also reported to provide a more complete view of model behavior across both sentiment classes [15], [21]. The complete research pipeline, including data collection, filtering, preprocessing, hybrid feature construction, model training, evaluation, and visualization, was implemented in Python. For monitoring purposes, the trained model was integrated into a Streamlit-based application to display sentiment distribution, confusion matrices, evaluation metrics, and temporal comparison results across applications.

3. RESULTS AND DISCUSSION

This section presents the empirical results of the proposed hybrid TF-IDF-SVD and Bi-LSTM framework and discusses their methodological and practical implications. The analysis is organized into two complementary parts. First, holdout evaluation on Dataset-1 is used to assess supervised sentiment classification performance on unseen reviews. Second, external temporal inference on Dataset-2 is used to examine yearly sentiment dynamics across four Indonesian e-wallet applications during 2024–2026. Beyond reporting quantitative outcomes, this section also interprets the observed patterns, relates them to the study design, and discusses their implications for sentiment monitoring in digital financial services.

3.1 Holdout Evaluation Results

The holdout evaluation was conducted to assess the supervised classification performance of the proposed hybrid TF-IDF-SVD and Bi-LSTM model on unseen data. After filtering and rating-based labeling, Dataset-1 contained 13,823 labeled reviews, which were divided into 11,058 training reviews and 2,765 testing reviews using a stratified split. The evaluation was performed on the 2,765-review holdout test set. The results indicate that the proposed model achieved an overall accuracy of 0.881 on the holdout test set, with macro-F1 and weighted-F1 values both equal to 0.881. This result suggests that the model maintained relatively balanced predictive performance across the positive and negative classes rather than concentrating performance on only one dominant label.

Table 5. Holdout evaluation metrics

Sentiment Class	Precision	Recall	F1-Score	Support
Negative	0.892	0.862	0.877	1351
Positive	0.872	0.900	0.886	1414
Accuracy			**0.881**	**2765**
Macro Avg	0.882	0.881	0.881	2765
Weighted Avg	0.882	0.881	0.881	2765

A closer inspection of class-level performance shows that the negative class obtained higher precision (0.892) but lower recall (0.862), whereas the positive class obtained slightly lower precision (0.872) but higher recall (0.900). This pattern indicates that the model was somewhat more conservative when assigning negative sentiment, resulting in fewer false positive negative predictions but a slightly greater tendency to miss some truly negative reviews. In contrast, the model was more inclusive in identifying positive sentiment, which improved recall but introduced a modest trade-off in precision. The relatively small gap between the two classes indicates that the hybrid architecture was able to combine lexical discrimination from TF-IDF-SVD with contextual sequence modeling from Bi-LSTM in a reasonably stable manner. In practical terms, this is important because app-review sentiment data often contain short, noisy, and informal expressions, where purely lexical or purely sequential representations may each capture only part of the available sentiment signal.

3.2 Comparative Performance with Baseline Models

To provide a clearer evaluation of the proposed model, a comparative analysis was conducted against several baseline models under the same holdout evaluation setting. This comparison aims to assess whether the proposed hybrid architecture offers measurable advantages over conventional approaches.

Table 6 compares the proposed hybrid model with selected baseline models under the same holdout evaluation setting. The results show that the proposed hybrid model did

not achieve the single highest accuracy among all compared models, as the TF-IDF + Logistic Regression baseline produced a slightly higher accuracy score. This finding indicates that lexical-statistical baselines remain highly competitive for short and noisy app-review texts. However, the hybrid model still demonstrated competitive overall performance across the evaluated metrics, while also offering a more structurally integrated framework that combines lexical salience and sequential contextual information within a unified architecture. In methodological terms, the value of the proposed framework does not lie solely in achieving the best raw classification score, but in supporting the broader objective of this study, namely linking supervised sentiment classification with structured cross-application temporal monitoring. Accordingly, the hybrid model should be interpreted as a competitive and methodologically aligned framework for temporally oriented review analysis rather than as a purely accuracy-driven alternative to simpler baselines.

Table 6. Comparative performance of the proposed model and baseline models

Model	Accuracy	Precision	Recall	Macro-F1	Weighted-F1
TF-IDF + Logistic Regression	0.8833	0.8845	0.8841	0.8833	0.8833
Bi-LSTM only	0.8773	0.8798	0.8784	0.8773	0.8772
Proposed Hybrid TF-IDF-SVD + Bi-LSTM	0.881	0.8848	0.8839	0.881	0.881

3.3 Confusion Matrix Analysis

Figure 4 provides a more detailed view of the prediction distribution on the holdout test set. The confusion matrix shows that correct predictions dominate in both sentiment classes, confirming that the model did not collapse into one-sided classification behavior. However, classification errors still occurred in both directions, which is expected given the noisy nature of app-review text and the use of rating-derived labels as sentiment proxies rather than fully human-annotated ground truth.

These misclassifications may arise from several sources. First, short reviews often contain limited contextual information, making sentiment polarity harder to infer reliably. Second, some reviews may include mixed opinions, such as positive evaluations of one feature combined with complaints about another. Third, because the training labels were derived from star ratings, the textual sentiment and numerical rating may not always be

perfectly aligned. As a result, some errors may reflect label noise rather than pure model failure.

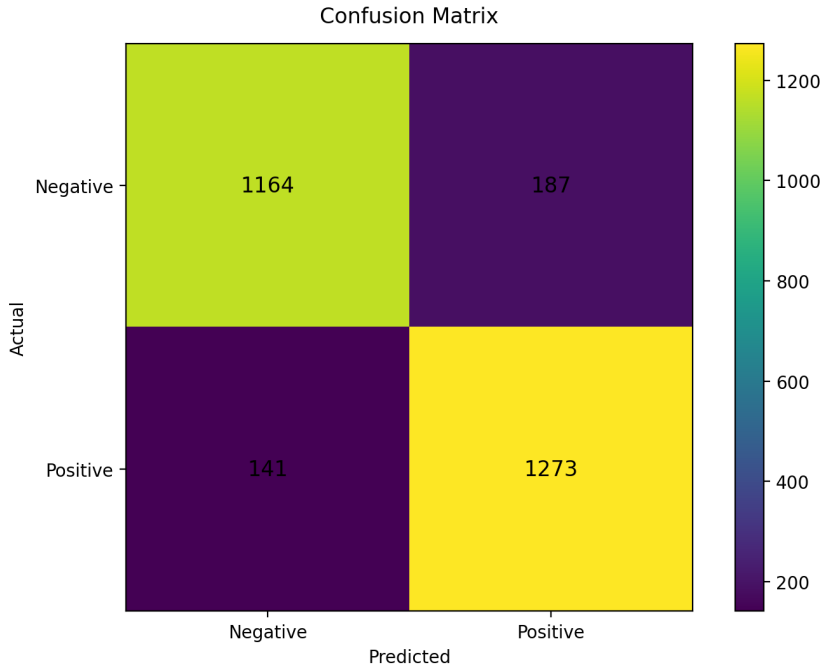

Figure 4. Confusion matrix on holdout test

A brief qualitative interpretation helps clarify this issue. For example, a user may assign a 4-star rating while still writing a complaint about delayed balance updates, failed transfers, or unstable login performance, causing the textual content to appear more negative than the assigned label. Conversely, a short review may contain a brief positive expression while receiving a low rating because the user was dissatisfied with one critical service incident. In such cases, the observed misclassification should not be interpreted solely as model failure, but also as a consequence of proxy-label mismatch in large-scale weakly supervised sentiment learning.

3.4 Training Process and Learning Behavior

Figure 5 illustrates the training dynamics of the proposed model across epochs. The learning curve shows that both training and validation performance improved during the early epochs and then stabilized before reaching the maximum training limit. This pattern suggests that the optimization process converged in a controlled manner and that the early stopping strategy functioned as intended to prevent unnecessary additional training.

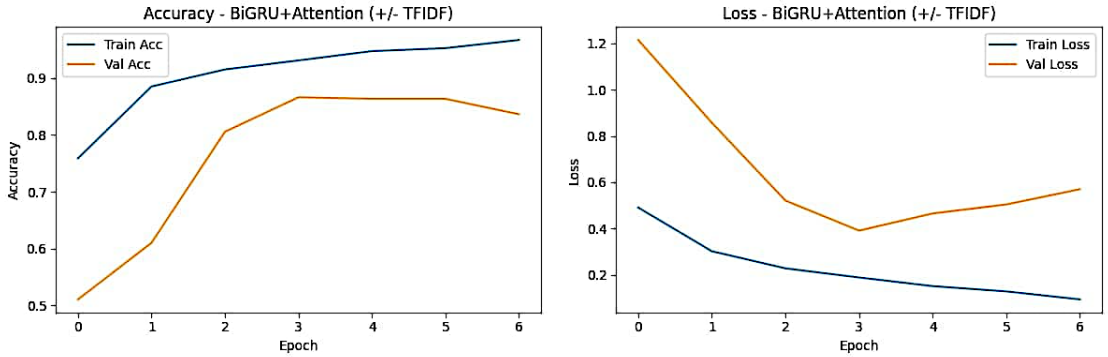

Figure 5. Learning curve (training history) of the proposed model.

The absence of extreme divergence between training and validation trajectories also indicates that the model did not exhibit severe overfitting under the adopted configuration. This is relevant because the proposed framework integrates two feature pathways with moderate dimensionality, and without appropriate monitoring the model could have become overly fitted to the training partition. The use of PR-AUC-based monitoring, early stopping, and learning-rate reduction therefore appears to have contributed to training stability.

3.5 Class Distribution after Labeling and Filtering

Figure 6 shows the final class distribution after filtering and rating-based label construction. The resulting dataset is relatively balanced between positive and negative sentiment classes, which is beneficial for evaluation because it reduces the risk that aggregate metrics are inflated by class dominance.

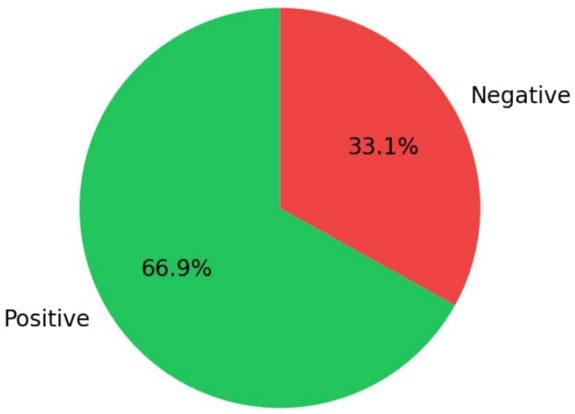

Figure 6. Final class distribution

This balance also helps explain why the macro-F1 and weighted-F1 scores are identical at 0.881. In imbalanced settings, weighted averages often mask weaker minority-class performance. In the present study, the near-balanced distribution means that the reported performance more credibly reflects model behavior across both sentiment categories.

3.6 Lexical Characteristics of Sentiment Classes

Figure 7 presents a lexical summary of dominant terms in positive and negative reviews. Although this visualization is descriptive rather than inferential, it helps contextualize the classification results by showing that the corpus contains repeated sentiment-bearing expressions associated with user satisfaction, technical problems, transaction delays, failed payments, and feature usability.

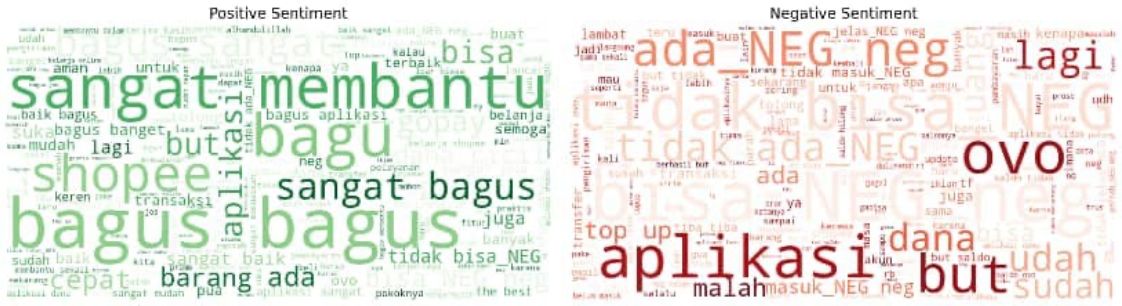

Figure 7. Word cloud / top terms comparison (Positive vs Negative).

This lexical contrast supports the relevance of combining TF-IDF-based lexical weighting with sequence-based contextual modeling. Frequent complaint-related words may be captured effectively through statistical weighting, while sentiment shifts caused by contrast markers and negation are better handled through sequential context modeling. Accordingly, the lexical profile shown in Figure 7 is consistent with the rationale of the proposed hybrid architecture.

3.7 External Temporal Inference Results and Application Comparison

The trained sentiment classification model was applied to Dataset-2 to examine sentiment dynamics under an external temporal inference setting. Dataset-2 contained 24,000 reviews from DANA, GoPay, OVO, and ShopeePay during 2024–2026, with a balanced design of 2,000 reviews per application per year. Because this dataset was not used during training, the results reported here should be interpreted as external

inference outcomes rather than supervised accuracy estimates. Accordingly, the reported yearly sentiment proportions should be interpreted as comparative sentiment indicators rather than direct measures of application quality.

Table 7. Positive sentiment proportion by application and year (2024–2026).

App	2024	2025	2026
DANA	0.7705	0.7755	0.774
GoPay	0.776	0.8395	0.8195
OVO	0.385	0.314	0.203
ShopeePay	0.828	0.784	0.7605

As shown in Table 7, DANA exhibited a relatively stable positive sentiment proportion across the observation period, with values of 0.7705 in 2024, 0.7755 in 2025, and 0.7740 in 2026. This pattern suggests a comparatively consistent user-perception profile, with no major directional shift over time. GoPay showed a different trajectory, increasing from 0.7760 in 2024 to 0.8395 in 2025, before remaining high at 0.8195 in 2026. This indicates a notable improvement in predicted user sentiment between 2024 and 2025, followed by a slight but still favorable adjustment in 2026.

In contrast, ShopeePay displayed a gradual decline from 0.8280 in 2024 to 0.7840 in 2025 and 0.7605 in 2026. Although it remained relatively strong in absolute terms, the downward direction suggests a weakening sentiment trajectory over time. The most pronounced negative pattern appeared in OVO, where the positive sentiment proportion decreased consistently from 0.3850 in 2024 to 0.3140 in 2025 and 0.2030 in 2026. Compared with the other applications, OVO's values indicate a substantially less favorable sentiment profile and the clearest negative temporal trend in the observation window.

3.8 Temporal Trend Visualization

Figure 8 visualizes the yearly sentiment trajectories across the four applications and makes the directional differences more explicit. DANA follows a nearly horizontal pattern,

GoPay shows an upward shift followed by slight normalization, ShopeePay declines moderately, and OVO declines sharply and continuously. Because the dataset was balanced by application-year cell, these comparisons are less likely to be distorted by unequal review volume and therefore provide a clearer basis for relative temporal interpretation.

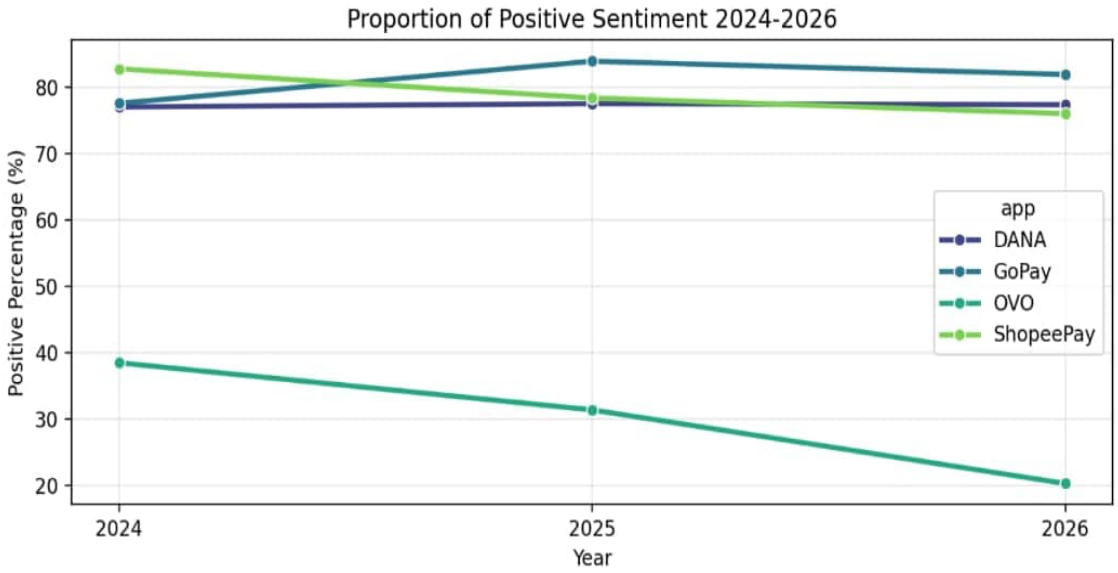

Figure 8. Time-series trend of positive sentiment proportion $P_{a,k}$(2024–2026) across applications.

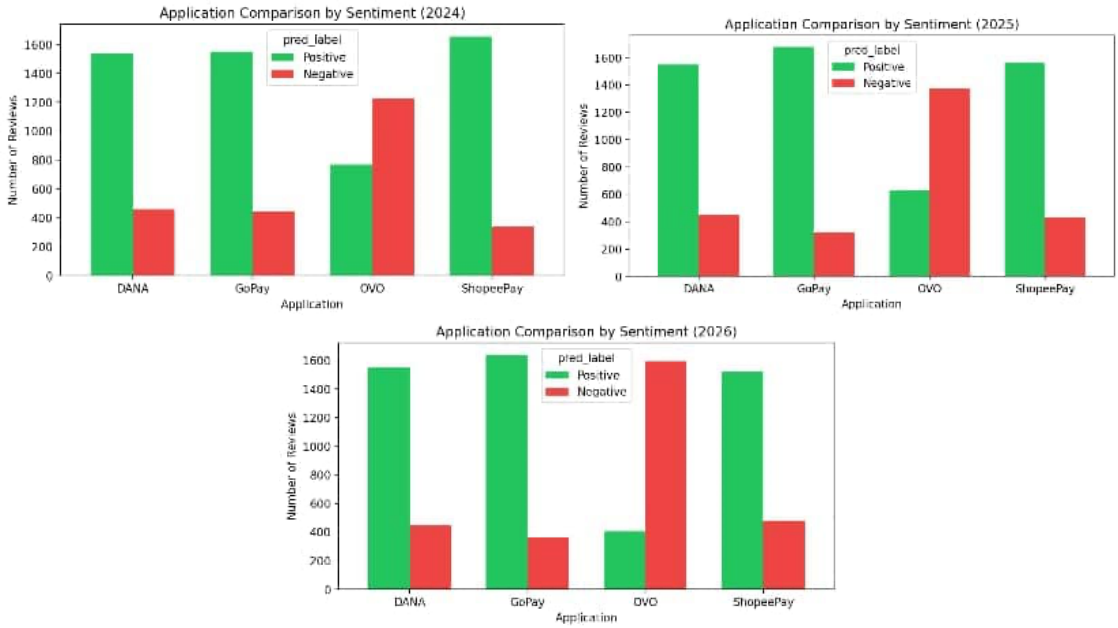

Figure 9. Cross-application comparison per year (grouped bar/line visualization).

Figure 9 complements this view by comparing applications within each year. In 2024, ShopeePay recorded the highest positive sentiment proportion, followed by GoPay and DANA, while OVO lagged substantially behind. In 2025, GoPay became the strongest performer, whereas OVO continued to remain lowest. By 2026, GoPay and DANA still occupied the upper range, ShopeePay remained positive but weaker than in earlier years, and OVO declined further. Taken together, these figures demonstrate that cross-application sentiment differences are not static and can evolve meaningfully over time.

3.9 Discussion

The empirical results indicate that the proposed hybrid framework is useful not only for supervised sentiment classification but also for structured temporal monitoring across e-wallet applications. On Dataset-1, the balanced macro-F1 and weighted-F1 values suggest that the model maintained relatively stable predictive behavior across both sentiment classes. Rather than demonstrating superiority solely in terms of raw accuracy, the hybrid architecture should be understood as a framework that integrates complementary lexical and contextual representations within a two-stage evaluation design. This is particularly relevant for app-review data, which are often short, noisy, and linguistically variable, and for research settings where the analytical objective extends beyond benchmark classification toward temporally structured comparative monitoring [6], [22].

The class-level results also reveal an important nuance. The model achieved slightly higher precision for the negative class but higher recall for the positive class. This suggests that the classifier was somewhat stricter in assigning negative sentiment while being more inclusive in assigning positive sentiment. Such behavior is plausible in app-review data because many user reviews contain short praise expressions, whereas negative reviews often include more diverse and context-dependent complaint forms [23]. The error pattern may therefore reflect both linguistic variability and label noise inherited from the rating-based supervision strategy.

Compared with previous app-review sentiment studies, the contribution of the present study lies more in evaluation design than in proposing a completely new text-classification architecture. Prior studies in e-wallet and related digital-service contexts have generally emphasized platform-specific sentiment classification, feature-level

opinion detection, or algorithm comparison under static train-test settings [4]–[6], [24]. In contrast, the present study combines supervised holdout evaluation with balanced external temporal inference across multiple applications and years. This design makes it possible to move beyond one-time benchmark performance and instead examine how relative sentiment trajectories differ across platforms under a controlled application-year structure.

The external temporal results strengthen the contribution of this study beyond static classification. Unlike prior app-review sentiment studies that mainly focus on single-platform classification accuracy or random train-test evaluation, the present study introduces a balanced application-year inference design to track relative sentiment movement over time. In this setting, DANA appears stable, GoPay improves markedly in 2025 and remains strong in 2026, ShopeePay shows a gradual weakening trend, and OVO experiences the sharpest decline. These differences suggest that sentiment monitoring can reveal platform-specific trajectories that would be obscured if all reviews were treated as one undifferentiated static corpus [24].

From a practical standpoint, the proposed framework may serve as an early monitoring instrument for digital financial service providers, particularly for detecting relatively stable, improving, or declining sentiment tendencies across platforms. However, these outputs should be interpreted cautiously. Because Dataset-2 was used only for inference, and because sentiment labels in Dataset-1 were derived from star ratings, the resulting yearly sentiment proportions should be understood as comparative sentiment indicators rather than direct measures of service quality. This caution is important to ensure that the practical implications remain aligned with the current data structure, proxy-labeling strategy, and temporal aggregation design [9], [24], [25].

These findings are generally consistent with previous studies showing that app-review sentiment analysis can provide useful feedback for service evaluation and user-opinion monitoring in digital financial applications [24], while also supporting the argument that more temporally aware evaluation settings are needed when review distributions shift over time [9]. Accordingly, the present study contributes by extending e-wallet review sentiment analysis from static benchmark-style classification toward comparative, time-aware monitoring under a structured external inference design.

Overall, the results show that the proposed hybrid framework achieved stable supervised performance on the holdout test set and generated interpretable temporal sentiment indicators under the external monitoring setting. The holdout results support the effectiveness of the hybrid TF-IDF-SVD and Bi-LSTM architecture for sentiment classification on Indonesian e-wallet reviews, while the external temporal results reveal distinct yearly sentiment trajectories across DANA, GoPay, OVO, and ShopeePay. Taken together, these findings support the main objective of the study, namely to move from static review classification toward structured temporal sentiment monitoring across digital payment applications.

4. CONCLUSION

Conclusion. This study proposed a hybrid sentiment analysis framework for Indonesian e-wallet reviews by combining TF-IDF-SVD and Bi-LSTM within a two-stage evaluation design consisting of holdout classification and external temporal inference. On the holdout test set, the model achieved competitive performance with an accuracy of 0.881 and balanced macro-F1 and weighted-F1 scores of 0.881. Under the external temporal setting, DANA remained relatively stable, GoPay improved markedly in 2025 and remained strong in 2026, ShopeePay showed a gradual decline, and OVO exhibited the strongest negative trend. These findings indicate that the proposed framework is useful not only for supervised sentiment classification but also for structured comparative temporal monitoring under the current data and proxy-labeling setting. However, the study remains limited by rating-based label construction and yearly aggregation. Future studies may strengthen the framework through partial manual annotation, finer temporal granularity, and comparison with transformer-based Indonesian language models.

REFERENCES

[1] Tikno, Y. S. Dharmawan, and Ngatini, "Investigating Consumer Acceptance of Mobile Payment Services in Indonesia," *Procedia Comput. Sci.*, vol. 234, pp. 1095–1102, 2024, doi: 10.1016/j.procs.2024.03.104.

[2] A. Yasin, R. Fatima, A. Nauman, and Z. Wei, "Python data odyssey: Mining user feedback from google play store," *Data Br.*, vol. 54, p. 110499, 2024, doi: 10.1016/j.dib.2024.110499.

[3] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[4] Y. Bau, T. E. Leong, and C. Goh, "Sentiment Analysis of E-Wallet Companies : Exploring Customer Ratings and Perceptions," *J. Logist. Informatics Serv. Sci.*, vol. 10, no. 4, pp. 189–205, 2023, doi: 10.33168/JLISS.2023.0413.

[5] S. Masturoh, R. L. Pratiwi, M. R. R. Saelan, and U. Radiyah, "Application Of The K-Nearest Neighbor (Knn) Algorithm In Sentiment Analysis of The Ovo E-Wallet Application," *JITK(JURNAL ILMU Pengetah. DAN Teknol. KOMPUTER)*, vol. 8, no. 2, pp. 8–13, 2023, doi: 10.33480/jitk.v8i2.3997.

[6] H. Juandri, Hasmawati, and Bunyamin, "Aspect-level sentiment analysis on GoPay app reviews using multilayer perceptron and word embeddings," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 9, no. 4, pp. 397–408, 2024, doi: 10.22219/kinetik.v9i4.2041.

[7] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst. Appl.*, vol. 223, p. 119862, 2023, doi: 10.1016/j.eswa.2023.119862.

[8] X. Wang, T. Zhang, Y. Tan, W. Shang, and Y. Li, "How to effectively mine app reviews concerning software ecosystem? A survey of review characteristics," *J. Syst. Softw.*, vol. 213, p. 112040, 2024, doi: https://doi.org/10.1016/j.jss.2024.112040.

[9] S. Maldonado, J. López, and A. Iturriaga, "Out-of-time cross-validation strategies for classification in the presence of dataset shift," *Appl. Intell.*, vol. 52, no. 5, pp. 5770–5783, 2022, doi: 10.1007/s10489-021-02735-2.

[10] H. Adiningtyas and A. S. Auliani, "Sentiment analysis for mobile banking service quality measurement," *Procedia Comput. Sci.*, vol. 234, pp. 40–50, 2024, doi: 10.1016/j.procs.2024.02.150.

[11] R. Alawaji and A. Aloraini, "Sentiment Analysis of Digital Banking Reviews Using Machine Learning and Large Language Models," *Electronics*, vol. 14, no. 11, 2025, doi: 10.3390/electronics14112125.

[12] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: a systematic literature review," *Empir. Softw. Eng.*, vol. 27, no. 2, p. 43, 2022, doi: 10.1007/s10664-021-10065-7.

[13] D. I. Af'idah, P. D. Anggraeni, M. Rizki, A. B. Setiawan, and S. F. Handayani, "Aspect-Based Sentiment Analysis for Indonesian Tourist Attraction Reviews Using Bidirectional Long Short-Term Memory," *JUITA J. Inform.*, vol. 11, no. 1 SE-Articles, pp. 27–36, May 2023, doi: 10.30595/juita.v11i1.15341.

[14] W. M. Baihaqi and A. Munandar, "Sentiment Analysis of Student Comment on the College Performance Evaluation Questionnaire Using Naïve Bayes and IndoBERT," *JUITA J. Inform.*, vol. 11, no. 2, pp. 213–220, 2023.

[15] R. B. Adinata, S. Supriyono, and D. L. Fithri, "Sentiment Classification of MyTelkomsel Reviews Using SVM and Logistic Regression," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 20, no. 1, 2026, doi: 10.22146/ijccs.110409.

[16] Y. N. Kunang and W. P. Mentari, "Analysis of the Impact of Vectorization Methods on Machine Learning-Based Sentiment Analysis of Tweets Regarding Readiness for Offline Learning," *JUITA J. Inform.*, vol. 11, no. 2, pp. 271–280, 2023.

[17] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 2, pp. 113–124, 2022, doi: 10.22146/ijccs.68903.

[18] I. Gambo, R. Massenon, R. Oluwaseun, S. Agarwal, and W. Pak, "Identifying and resolving conflict in mobile application features through contradictory feedback analysis," *Heliyon*, vol. 10, no. 17, p. e36729, 2024, doi: 10.1016/j.heliyon.2024.e36729.

[19] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," *JUITA J. Inform.*, vol. 10, no. 2 SE-Articles, pp. 225–232, Nov. 2022, doi: 10.30595/juita.v10i2.13262.

[20] G. Tamami, W. A. Triyanto, and S. Muzid, "Sentiment Analysis Mobile JKN Reviews Using SMOTE Based LSTM," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 19, no. 1, pp. 13–24, 2025, doi: 10.22146/ijccs.101910.

[21] M. Parhusip, S. Sudianto, and T. G. Laksana, "Sentiment Analysis of the Public Towards the Kanjuruhan Tragedy with the Support Vector Machine Method," *JUITA J. Inform.*, vol. 11, no. 2, pp. 241–251, 2023.

[22] N. K. A, D. Lestari, and G. T. Pranoto, "Sentiment Analysis Review Threads Google Play Store with RoBERTa Model," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 14, no. 4, pp. 272–280, 2025, doi: 10.22146/jnteti.v14i4.22038.

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

[23] R. D. Kurniawan, A. Yohannis, and W. T. Atmojo, "Sentiment Analysis of Getcontact Application Reviews on Google Play Store Using Naive Bayes Algorithm," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2848–2858, 2025, doi: 10.52436/1.jutif.2025.6.4.5248.

[24] N. Laili, I. Fuji, T. Vani, D. Naraya, and H. Niken, "NLP-Based Sentiment Analysis of Alfagift and Klik Indomaret Application Reviews : A Comparative Study," *J. Inf. Syst. Informatics*, vol. 7, no. 3, pp. 2458–2475, 2025, doi: 10.51519/journalisi.v7i3.1178.

[25] H. Kelvin, Y. Desnelita, and D. Oktarina, "Sentiment Analysis of IKD Application Reviews on Play Store Using Random Forest," *J. Nas. Tek. Elektro dan Teknol.*, vol. 14, no. 3, pp. 171–180, 2025, doi: 10.22146/jnteti.v14i3.20473.