 ISSN 2579-7069

Comparative Sentiment Analysis of Digital Wallet Applications in Indonesia Using Naïve Bayes

Soeltan Abdul Ghaffar[1,*], Wilbert Clarence Setiawan[2]

[1]*Department of Marine Information Systems, Universitas Pendidikan Indonesia, Bandung, Indonesia*

[2]*Faculty of Informatics Engineering, Universitas Taruma Negara, Jakarta, Indonesia*

(Received July 1, 2024; Revised October 5, 2024; Accepted February 10, 2025; Available online March 2, 2025)

Abstract

The rapid growth of financial technology in Indonesia has led to widespread use of digital wallet applications such as OVO, DANA, GoPay, and ShopeePay. User-generated reviews on platforms like the Google Play Store offer valuable insights into customer satisfaction and application performance. This study conducts a comparative sentiment analysis of user reviews for four major Indonesian e-wallets using the Multinomial Naïve Bayes algorithm. A total of 401 Indonesian-language reviews were collected and labeled based on user ratings, with sentiment classified as positive or negative. The TF-IDF method was applied for feature extraction, and the model was evaluated using accuracy, precision, and recall metrics. Results show that ShopeePay achieved the highest classification accuracy (89%), followed by DANA and GoPay (80%), while OVO recorded lower performance due to more informal and ambiguous language. The model demonstrated strong precision for positive sentiment but low recall for negative sentiment (28%), indicating challenges in detecting minority-class feedback. Word cloud visualizations were used to highlight common keywords in each sentiment category. This study confirms that Naïve Bayes is an effective approach for classifying user sentiment in Indonesian-language app reviews, while also emphasizing the need for improved handling of class imbalance in future research. The findings provide practical insights for developers to enhance user experience based on data-driven sentiment patterns.

Keywords: Sentiment Analysis, Digital Wallet, Naïve Bayes, User Reviews, E-Wallet, Indonesia, TF-IDF

1. Introduction

The rapid development of financial technology (fintech) in Indonesia has significantly transformed the way people conduct financial transactions. One of the most widely adopted fintech innovations is the digital wallet (e-wallet), which enables users to perform cashless transactions such as payments, transfers, and purchases directly via mobile applications. According to Bank Indonesia, the value of electronic money transactions reached IDR 19.2 trillion in February 2021, marking a 26.42% increase compared to the previous year [1]. A survey conducted in 2024 found that approximately 96% of Indonesians have used e-wallet services such as OVO, DANA, GoPay, or ShopeePay for various daily needs, including online shopping, transportation, and utility payments [2].

The widespread use of e-wallets has led to a surge in user feedback available on platforms such as the Google Play Store. These user reviews represent an important source of public opinion that can be utilized to evaluate application performance and user satisfaction [3]. However, the unstructured and informal nature of these reviews poses challenges for manual analysis, especially when written in colloquial or non-standard language.

To address this challenge, sentiment analysis is commonly used as an automated approach to detect user opinions expressed in text data [4]. By categorizing sentiments as positive, negative, or neutral, developers and service providers can gain strategic insights into user experiences and service quality [5]. Among the many machine learning algorithms used in sentiment analysis, the Naïve Bayes classifier is widely recognized for its simplicity, efficiency, and effectiveness in classifying short text formats such as app reviews [6][7].

***Corresponding author: Soeltan Abdul Ghaffar (soeltan027ghaffar@gmail.com)**
DOI: https://doi.org/10.47738/ijiis.v8i2.251

Previous studies have shown that Naïve Bayes can achieve strong classification performance when applied to digital wallet review data. Reported accuracy rates have reached up to 90% for certain applications, depending on the dataset and preprocessing techniques used [8]. Studies using variants such as Multivariate Bernoulli have reported accuracy rates of over 83% for popular platforms like OVO and GoPay [9], while other experiments applying the method to different fintech applications achieved accuracy as high as 91%, including perfect recall in detecting negative sentiment [10].

Despite these promising results, most prior research has been limited to a single application, offering little comparative insight across platforms. Moreover, limited attention has been paid to the classifier's performance in identifying minority sentiment classes—particularly negative reviews—which are crucial for user feedback and service improvement. Variability in data sources also affects consistency across studies.

To address these gaps, this study conducts a comparative sentiment analysis of four leading e-wallet applications in Indonesia—OVO, DANA, GoPay, and ShopeePay—using a unified dataset collected from the Google Play Store. The analysis uses TF-IDF for feature extraction and applies the Multinomial Naïve Bayes algorithm for classification. Model performance is evaluated using accuracy, precision, and recall metrics. The goal is to contribute theoretically to Indonesian-language sentiment analysis and practically by offering app developers insights into user satisfaction using a data-driven approach [11].

2. Literature Review

Sentiment analysis is a technique within natural language processing (NLP) and text mining that aims to identify and classify the sentiment polarity of a text, typically into positive, negative, or neutral categories [12]. In the context of digital wallet applications, sentiment analysis helps developers and service providers evaluate user satisfaction and application quality based on textual reviews [13].

Among the various classification algorithms available, the Naïve Bayes classifier is widely used due to its simplicity, fast computation, and relatively strong performance in handling short texts such as app reviews [14][15]. This algorithm applies Bayes' Theorem and operates under the assumption that features are conditionally independent, which allows efficient modeling even with limited training data [16].

Studies have demonstrated that Naïve Bayes can produce high accuracy in classifying user sentiments on e-wallet reviews, with some achieving accuracy levels above 90% [17]. Other research has shown that the algorithm can detect positive sentiment with high precision, although performance on negative sentiment is often lower due to class imbalance and linguistic complexity [18]. Additional results suggest that Naïve Bayes performs well when classifying structured review data from platforms like the Google Play Store, where user experiences are directly expressed [19].

However, many existing studies focus on a single digital wallet application, limiting the ability to compare model performance across different platforms. In some cases, studies utilize data from social media platforms such as Twitter, which may not accurately reflect user experiences within the application environment [20].

Recent research has explored the enhancement of Naïve Bayes performance through optimization techniques such as Particle Swarm Optimization (PSO) for feature selection, which has shown improved results in classification tasks [21]. Other studies compare Naïve Bayes with alternative algorithms such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and ensemble approaches to determine the most robust method for sentiment classification [22].

Despite these developments, there remains a lack of comprehensive studies that apply a unified sentiment analysis approach to multiple e-wallet platforms using a consistent dataset and evaluate results using detailed metrics such as accuracy, precision, and recall. This study aims to fill that gap by comparing four major e-wallet applications in Indonesia using Multinomial Naïve Bayes and standardized review data from the Google Play Store.

 ISSN 2579-7069

3. Methodology

3.1. Research Design

This study employs a quantitative experimental approach to evaluate the effectiveness of the Multinomial Naïve Bayes algorithm in classifying sentiment from user reviews of digital wallet applications. The primary objective is to perform a comparative sentiment analysis on four leading e-wallet platforms in Indonesia—OVO, DANA, GoPay, and ShopeePay—using a consistent dataset and standardized classification framework. The research procedure follows a structured pipeline that includes data collection, text preprocessing, sentiment labeling, feature extraction, model training and testing, and performance evaluation. This design ensures consistency and comparability across all application datasets.

3.2. Data Source and Collection

User reviews were collected from the Google Play Store using a scraping tool (google-play-scraper). The reviews were filtered to retain only those written in Indonesian and were cleaned by removing duplicate entries, irrelevant content, and empty reviews. The final dataset consisted of 401 reviews, distributed nearly equally among the four applications. The table 1 presents the number of reviews collected per application.

Table 1. Distribution of Reviews per Application

Application	Number of Reviews
DANA	101
ShopeePay	100
OVO	100
GoPay	100
Total	**401**

Each review included metadata such as rating, which was later used for sentiment labeling.

3.3. Text Preprocessing

Text preprocessing was carried out to normalize and clean the raw textual data for further analysis. All text was converted to lowercase using case folding to ensure uniformity in token representation. Tokenization was applied to segment the text into individual words or tokens. Stopword removal was conducted to eliminate commonly used words in Bahasa Indonesia that do not convey sentiment meaning. Stemming was used to reduce words to their base forms, allowing the model to treat different morphological variants of the same word as a single feature. These steps improved the quality and consistency of the data used in classification.

3.4. Sentiment Labeling

Sentiment labels were assigned automatically based on the numeric star ratings associated with each review. Reviews with a rating of 4 or 5 were classified as expressing positive sentiment. Those with ratings of 1 or 2 were labeled as negative sentiment. Reviews with a rating of 3 were excluded from the dataset as they are considered neutral and may introduce ambiguity. This binary classification method simplifies the modeling process and ensures a clear division between sentiment classes.

3.5. Feature Extraction using TF-IDF

The cleaned review texts were transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) method. This technique assigns a weight to each term in a document relative to its frequency in the overall corpus. Terms that are common in a specific document but rare in others are given higher weights, making them more influential in classification. The TF-IDF formula is defined as:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

 ISSN 2579-7069

In this equation, $\mathrm{TF}(t, d)$represents the frequency of term $t$in document d, Ndenotes the total number of documents, and $DF(t)$is the number of documents that contain the term t. The result is a weighted vector representation of each review, used as input to the classification model.

3.6. Classification Model

The classification algorithm used is the Multinomial Naïve Bayes, which is widely used in text classification tasks due to its simplicity and efficiency. This algorithm is based on Bayes' Theorem and assumes conditional independence among features given the class label. It estimates the posterior probability of a class $C_k$given an input feature vector $x$using the following formula:

$$P(C_k \mid x) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(x)}$$

Here, $P(C_k)$is the prior probability of class C_k, $P(x_i \mid C_k)$is the likelihood of feature $x_i$given class C_k, and $P(x)$is the normalizing constant. The algorithm is particularly effective for handling sparse feature vectors derived from TF-IDF representations.

3.7. Experimental Setup and Evaluation

The dataset was split into training and testing sets, with 80 percent of the data used to train the model and 20 percent used for evaluation. This approach enables the assessment of how well the model generalizes to unseen data. To evaluate the performance of the classification model, three standard metrics were used: accuracy, precision, and recall. Accuracy measures the overall proportion of correct predictions. Precision quantifies the number of true positive predictions over all positive predictions made by the model. Recall assesses the proportion of actual positives correctly identified by the model. These metrics were calculated both for the entire dataset and for each individual application, allowing a comprehensive comparison of classification performance across platforms. In addition to quantitative evaluation, word cloud visualizations were generated to highlight frequently occurring words in positive and negative reviews. These visual tools provide insight into the dominant themes and expressions used by users in both sentiment classes.

4. Results and Discussion

The sentiment classification model using the Multinomial Naïve Bayes algorithm was evaluated on 401 user reviews from four digital wallet applications in Indonesia: OVO, DANA, GoPay, and ShopeePay. Each application contributed approximately 100 reviews, collected from the Google Play Store. This section presents the results in a structured manner using several tables, followed by interpretation and analysis.

4.1. Sentiment Distribution

Understanding the overall distribution of sentiment is a crucial initial step in evaluating user perceptions toward digital wallet applications. In this study, sentiment classification was performed by assigning labels based on the numeric ratings provided in Google Play Store reviews. Reviews with ratings of 4 and 5 were categorized as positive sentiment, while ratings of 1 and 2 were classified as negative sentiment. Reviews with a rating of 3, often considered neutral or ambiguous, were excluded from the dataset to focus the classification model on clear, polarized sentiment categories.

The application of this labeling method aligns with prior studies in sentiment analysis, where star-based annotation provides a scalable alternative to manual labeling. However, it also comes with limitations, such as potential mismatches between numerical ratings and the actual tone of the text—especially in cases where users assign a low rating but write a positive comment, or vice versa.

After preprocessing and filtering, the final dataset comprised 401 user reviews, distributed across four applications: OVO, DANA, GoPay, and ShopeePay. The overall sentiment breakdown is shown in the table below.

Table 2. Overall Sentiment Distribution

Sentiment	Total Reviews	Percentage
Positive	260	64.8%
Negative	141	35.2%
Total	401	100%

The data clearly indicates a predominance of positive sentiment, with approximately 65% of all reviews expressing satisfaction or approval. This trend is consistent with general user behavior on app stores, where satisfied users tend to rate highly and provide short affirmations such as “mantap” (roughly translated as "awesome" or "great"), “bagus” ("good"), or “sangat membantu” ("very helpful"). These simple expressions, though informal, strongly signal user approval and contribute to the dominance of positive sentiment.

It also reflects the growing trust in digital wallet services in Indonesia, especially during and after the COVID-19 pandemic, when contactless transactions became the norm and were actively encouraged by both the government and private sector. On the other hand, 35% of the reviews were negative, representing a substantial minority that should not be overlooked. Negative feedback often highlights pain points such as failed top-ups, delays in fund transfers, login issues, or dissatisfaction with customer service. These critical reviews, while fewer in number, are highly informative for developers aiming to improve their apps' performance and user experience. A more granular look at sentiment distribution across the four applications provides further insights into the nature of user experience for each platform.

Table 3. Sentiment Distribution by Application

Application	**Positive Reviews**	**Negative Reviews**	**Total Reviews**
DANA	67	34	101
ShopeePay	70	30	100
GoPay	61	39	100
OVO	62	38	100

Among the four e-wallets, ShopeePay recorded the highest number of positive reviews, with 70 out of 100 reviews (70%) categorized as positive. This aligns with the application's reputation for offering integrated promotions, cashback offers, and a seamless in-app experience, especially for users who frequently shop on Shopee’s e-commerce platform. ShopeePay’s strong ecosystem integration may contribute to higher user satisfaction.

DANA also performed well, with 67 positive reviews out of 101, representing approximately 66.3%. DANA is known for its easy-to-use interface and support for QRIS (Quick Response Code Indonesian Standard), a national QR code system that enables standardized payments across merchants. User satisfaction with DANA is also reinforced by features like free interbank transfers and recurring bill payments.

In contrast, GoPay had the highest number of negative reviews (39 out of 100), comprising 39% of its total dataset. This may indicate recurring issues among GoPay users during the review period, such as delayed transaction confirmations, app bugs, or difficulties syncing with Gojek’s broader ecosystem (ride hailing, food delivery, etc.). GoPay's multifunctionality may expose it to more potential points of failure, increasing the likelihood of negative feedback.

OVO had a relatively balanced distribution, with 62 positive and 38 negative reviews. Although OVO was one of the first movers in Indonesia's digital payment space, rising competition and service overlap with newer platforms may have contributed to reduced user satisfaction. Another potential factor is the use of informal and regionally diverse language in OVO reviews, which may introduce ambiguity—both for human readers and classification models. For example, terms like “agak susah” (a bit difficult) or “kurang jelas” (not clear enough) may carry nuanced criticism that is harder to detect through automatic classification.

In summary, sentiment distribution analysis reveals not only quantitative trends in user satisfaction but also early indicators of application strengths and weaknesses. While ShopeePay and DANA enjoy generally favorable sentiment, GoPay and OVO face more challenges in meeting user expectations. Furthermore, the imbalance between positive and negative sentiment—both overall and per app—introduces a form of class imbalance that can affect model

 ISSN 2579-7069

performance, particularly in recall for minority classes. These implications are addressed in detail in the following sections on classification performance.

4.2. Model Performance

The performance of the sentiment classification model was evaluated after training the Multinomial Naïve Bayes classifier on 80 percent of the dataset and testing it on the remaining 20 percent. This division of training and testing data allowed the model to learn from a representative portion of the review corpus while assessing its ability to generalize to unseen examples. The evaluation focused on three key metrics: accuracy, precision, and recall. Accuracy refers to the overall correctness of the model's predictions. Precision measures the proportion of correctly predicted positive (or negative) instances among all predictions labeled as such. Recall, on the other hand, reflects the model's ability to capture all actual positive (or negative) instances from the dataset.

Table 4 below presents the accuracy achieved by the classifier for each application. ShopeePay demonstrated the highest accuracy at 89 percent, indicating that the model correctly classified nearly nine out of ten reviews for this platform. This strong performance is likely due to the consistency and clarity of language used in ShopeePay user reviews, where expressions of sentiment—both positive and negative—are often direct and formulaic. DANA and GoPay followed with equal accuracy scores of 80 percent. While still high, this slightly lower accuracy may indicate greater linguistic variability in user reviews, requiring the model to process a broader range of lexical patterns. OVO obtained the lowest accuracy at 76 percent. This result suggests that reviews for OVO contained more ambiguous or informal expressions, which can complicate classification, especially for frequency-based models that rely on word distribution patterns like Naïve Bayes.

Table 4. Accuracy per Application

Application	Accuracy
ShopeePay	89%
DANA	80%
GoPay	80%
OVO	76%

In addition to accuracy, the model's precision and recall were also examined to better understand its strengths and limitations. As shown in Table 5, ShopeePay again led in both metrics, achieving a precision of 89 percent and a recall of 83 percent. These values reflect the model's high confidence and correctness in identifying sentiment polarity within ShopeePay reviews. DANA yielded a precision of 80 percent and a recall of 75 percent, while GoPay recorded similar precision at 80 percent but a slightly lower recall at 72 percent. OVO, consistent with its performance in accuracy, had the lowest scores across both metrics, with a precision of 76 percent and a recall of 65 percent.

Table 5. Precision and Recall per Application

Application	Precision	Recall
ShopeePay	89%	83%
DANA	80%	75%
GoPay	80%	72%
OVO	76%	65%

High precision across all applications indicates that the model was able to correctly identify positive and negative sentiments without generating excessive false positives. However, recall values were generally lower, especially for OVO. A lower recall suggests that the model failed to recognize a substantial portion of reviews that truly belonged to a particular sentiment category, particularly negative ones. This shortfall is critical because it implies that user dissatisfaction may go undetected, potentially hindering efforts to improve app quality or respond to user concerns.

The discrepancy between precision and recall reflects the impact of class imbalance in the dataset, where positive reviews outnumber negative ones. As a result, the model becomes biased toward predicting the majority class. While this improves overall precision, it diminishes the model's ability to recall less frequent sentiment types. This issue is particularly visible in OVO, where reviews may not only be fewer in number but also linguistically inconsistent, using informal expressions or mixed sentiment in a single sentence.

From these results, it can be concluded that the Naïve Bayes model performs well when classifying clearly expressed, majority-class sentiments—especially positive reviews. However, its performance declines when dealing with more nuanced, imbalanced, or minority-class data, such as negative reviews with indirect or sarcastic tone. Addressing this challenge may require incorporating class balancing techniques, such as oversampling the minority class or assigning higher weight to negative samples during training. Additionally, experimenting with more robust classification models such as Support Vector Machines or ensemble methods may improve both recall and overall performance without sacrificing precision.

4.3. Confusion Matrix Results

To gain a deeper understanding of the model's classification behavior beyond general performance metrics, a confusion matrix was constructed to analyze how the Multinomial Naïve Bayes classifier performed in distinguishing between positive and negative sentiment classes. The confusion matrix provides a detailed view of the distribution of correctly and incorrectly classified instances, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This level of analysis is particularly important in datasets with imbalanced class distributions, as overall accuracy can sometimes mask performance deficiencies in the minority class.

Table 6. Confusion Matrix (Overall Model Performance)

	Predicted Positive	**Predicted Negative**
Actual Positive	128	2
Actual Negative	54	21

The confusion matrix shows that out of all actual positive reviews, 128 were correctly predicted as positive, while only 2 were incorrectly predicted as negative. This indicates that the model was highly effective in detecting positive sentiment, with a false negative rate of less than 2% for the majority class. However, the model's performance on negative sentiment was notably weaker. Of the 75 actual negative reviews, only 21 were correctly classified, while 54 were incorrectly labeled as positive. This results in a false positive rate of 72% for the negative class, meaning the model frequently failed to detect user dissatisfaction.

This misclassification of negative reviews as positive suggests that the model lacks sensitivity toward less common or more complex expressions of sentiment. Since the dataset is imbalanced—with positive reviews comprising approximately 65% of the total—the classifier tends to favor the majority class, leading to biased predictions that reduce recall for the minority class.

To further quantify this imbalance in performance, both macro average and weighted average scores were calculated for the main evaluation metrics: precision, recall, and F1-score. These scores offer a more nuanced assessment than overall accuracy by accounting for the distribution and importance of each class.

Table 7. Macro and Weighted Average Scores

Metric	**Macro Average**	**Weighted Average**
Precision	83%	82%
Recall	63%	82%
F1-score	66%	78%

The macro average represents the unweighted mean of each metric across both sentiment classes, treating positive and negative reviews as equally important. In this case, macro average recall is relatively low at 63%, which reinforces the conclusion that the model struggles with detecting negative sentiment. The F1-score, which harmonizes precision and recall, is also low at 66% under macro averaging, suggesting an imbalance between correctly predicted classes.

In contrast, the weighted average takes class imbalance into account by giving more weight to metrics from the dominant class—in this case, positive sentiment. The weighted precision and recall are both higher at 82%, and the F1-score reaches 78%. These results highlight how the presence of more positive reviews in the dataset positively influences the overall performance metrics, potentially creating a misleading impression of model robustness.

 ISSN 2579-7069

The gap between macro and weighted averages, particularly in recall, underscores the need for caution when interpreting high performance scores in imbalanced datasets. While the model appears accurate overall, its effectiveness in identifying minority sentiment—especially negative feedback that could inform application improvements—is limited.

To mitigate this issue, future iterations of the model could apply techniques such as resampling (e.g., SMOTE), adjusted class weights, or the use of ensemble learning algorithms that are better suited to handling class imbalance. Additionally, incorporating syntactic or semantic features beyond TF-IDF, such as dependency parsing or contextual embeddings, may help the classifier better understand informal or nuanced negative expressions common in user-generated reviews.

4.4. Word Cloud Insights

To gain a qualitative understanding of the lexical patterns associated with positive and negative sentiments, a word frequency analysis was conducted. The results were visualized using word clouds and summarized in tabular form to highlight the most common terms found in each sentiment category. This analysis provides insight into how users linguistically express their satisfaction or dissatisfaction with digital wallet applications, and also supports the interpretability of the machine learning classification process.

The words were extracted from preprocessed review texts after tokenization and stopword removal. High-frequency terms reflect dominant expressions that users rely on to convey their experiences, making them critical lexical features for the TF-IDF-based sentiment classifier. The presence of such terms in both training and test datasets contributes to the model's ability to recognize sentiment-bearing patterns. The most frequent terms found in positively labeled reviews are listed in the table 8.

Table 8. Frequent Words in Positive Reviews

Word	Frequency
mantap	45
bagus	38
ok	35
sangat	30
membantu	27

The word "mantap" is the most dominant expression, appearing 45 times. In Indonesian, "mantap" is a highly colloquial term equivalent to "awesome" or "excellent," often used in short positive reviews to express approval. Similarly, "bagus" translates to "good," and is another common term that indicates user satisfaction. The term "ok", though borrowed from English, is widely used in Indonesian digital communication to signal general approval or functionality. The words "sangat" ("very") and "membantu" ("helpful") further reinforce the positive tone, often appearing in phrases such as "sangat bagus" ("very good") or "aplikasinya sangat membantu" ("the app is very helpful").

These expressions indicate that satisfied users tend to use short, positive adjectives and intensifiers, often with minimal elaboration. This pattern contributes to the model's high precision in identifying positive sentiment, as these words carry strong and unambiguous emotional polarity. Conversely, the most frequent words in reviews labeled as negative are presented below.

Table 9. Frequent Words in Negative Reviews

Word	Frequency
uang	41
masuk	36
susah	33
error	28
kendala	24

The word "uang", which means "money," appears most frequently in negative reviews, often in contexts such as "uang tidak masuk" ("money did not go through") or "kehilangan uang" ("lost money"). The term "masuk" ("entered" or "received") is often associated with failed transactions or delayed fund transfers, highlighting core functionality issues

with wallet applications. The word “susah” means “difficult” and is commonly used to describe problems with access, verification, or usage of the app. “Error”, although borrowed directly from English, is universally understood by Indonesian users and typically refers to app crashes, bugs, or system failures. Finally, “kendala”, which means “obstacle” or “problem,” is a generic complaint term that users often include when expressing frustration with technical or customer service issues.

These frequent negative terms reflect recurring complaints related to financial security, transaction reliability, and user experience barriers. Unlike positive reviews, which often consist of brief praise, negative reviews tend to be more descriptive, indicating frustration and unmet expectations. These linguistic features are harder to detect and classify accurately, especially when they are context-dependent or embedded in compound sentences.

Overall, the word frequency analysis provides a valuable layer of interpretability for the classifier. It confirms that sentiment-laden keywords align closely with model predictions and helps identify potential focus areas for future app improvement. Developers may use these insights to track user concerns over time, prioritize feature enhancements, and improve user satisfaction based on lexical feedback patterns.

4.5. Discussion

The experimental results demonstrate that the Multinomial Naïve Bayes classifier performs well in detecting positive sentiment, particularly in contexts where user reviews exhibit high lexical consistency and clarity. This is most evident in the case of ShopeePay, where the majority of reviews use simple and direct language that aligns with the assumptions of the Naïve Bayes model. The classifier's success in this domain underscores its strength in handling sentiment detection tasks where high-frequency terms and clear polarity are present.

However, the model shows significant limitations in accurately identifying negative sentiment. Despite achieving high overall accuracy and precision, the recall for the negative class is considerably low—only 28 percent in overall performance metrics. This indicates that the model frequently misclassifies negative reviews as positive, which may result in missed opportunities for corrective action or customer support interventions in practical deployments. In real-world applications, such as automated feedback monitoring or alert systems for app developers, the inability to detect negative sentiment reliably may hinder response effectiveness and degrade user trust.

One of the primary causes of this issue is the imbalance in the sentiment distribution of the dataset. Positive reviews dominate the dataset, comprising nearly two-thirds of the total, which biases the classifier toward the majority class during training. As a result, the model becomes more adept at recognizing positive patterns while neglecting the comparatively infrequent and often more linguistically complex negative expressions.

Furthermore, linguistic complexity in negative reviews exacerbates the problem. Users may express dissatisfaction using sarcasm, regional slang, idioms, or mixed-sentiment constructions, all of which are difficult for basic frequency-based models like Naïve Bayes to interpret accurately. For instance, a review that assigns a low star rating but uses polite or ambiguous language may be incorrectly interpreted as neutral or even positive, especially when common negative keywords are absent.

To address these challenges and improve sentiment classification in future research, several strategies are recommended. First, the use of data balancing techniques such as Synthetic Minority Oversampling Technique (SMOTE), undersampling of the majority class, or adjusting class weights during model training can help reduce bias toward the dominant class and enhance recall for minority sentiment. Second, the incorporation of more advanced classification algorithms, including Support Vector Machines (SVM), Random Forests, or ensemble methods, could offer improved performance by capturing more nuanced decision boundaries and complex feature interactions.

In addition, exploring hybrid approaches that combine machine learning models with lexicon-based sentiment analysis could help address the shortcomings of statistical models when dealing with linguistically subtle reviews. Lexicon-based techniques offer interpretability and can be tailored to recognize domain-specific sentiment cues or intensifiers. Moreover, incorporating manual labeling of a representative subset of the dataset—especially for ambiguous or mixed-sentiment cases—would improve ground truth reliability and serve as a better benchmark for model evaluation.

In conclusion, while Multinomial Naïve Bayes provides a solid baseline for sentiment classification in Indonesian-language app reviews, especially for positive sentiment detection, its effectiveness is limited when facing class imbalance and complex negative expressions. Addressing these limitations will be essential for building more robust and practical sentiment analysis systems that can support real-time monitoring and quality assurance in fintech applications.

5. Conclusion

This study has demonstrated the applicability of the Multinomial Naïve Bayes algorithm for sentiment classification of Indonesian-language user reviews from four major digital wallet applications: ShopeePay, DANA, GoPay, and OVO. By using a consistent dataset sourced from the Google Play Store and applying a standardized preprocessing and classification pipeline, the research provides comparative insights into user sentiment and model performance across different platforms. The findings show that the Naïve Bayes classifier performs well in detecting positive sentiment, particularly in applications with lexically consistent reviews, such as ShopeePay. However, its ability to identify negative sentiment remains limited due to class imbalance and the linguistic complexity often present in user complaints. These issues are reflected in the low recall for the negative class, which poses challenges for real-world applications such as customer feedback monitoring and service improvement initiatives. Despite its limitations, the model offers a reproducible and computationally efficient baseline for Indonesian-language sentiment analysis, highlighting both opportunities and challenges in the field. To enhance future work, the study recommends addressing class imbalance through resampling techniques, incorporating more advanced or ensemble classification models, and integrating lexicon-based approaches or manual annotation to better handle nuanced expressions. In sum, this research contributes a practical case study on applied sentiment analysis in the Indonesian fintech context and lays a foundation for future efforts aiming to develop more robust, sensitive, and context-aware sentiment classification systems.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, S.A.G. and W.C.S.; Methodology, S.A.G. and W.C.S.; Software, W.C.S.; Validation, S.A.G. and W.C.S.; Formal Analysis, S.A.G.; Investigation, S.A.G. and W.C.S.; Resources, S.A.G. and W.C.S.; Data Curation, W.C.S.; Writing—Original Draft Preparation, S.A.G.; Writing—Review and Editing, W.C.S.; Visualization, W.C.S. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Rizkya, R. Rianto, and A. I. Gufroni, "Implementation of the Naive Bayes Classifier for Sentiment Analysis of Shopee E-Commerce Application Review Data on the Google Play Store," *Jurnal Aplikasi Sistem Informasi dan Informatika (JAISI)*, vol. 1, no. 1, pp. 14–22, 2023, https://doi.org/10.37058/jaisi.v1i1.8993

[2] A. Basir, "Analysis of Electronic Wallet User Sentiment on Twitter (X) Social Media Using the Naïve Bayes Classifier Algorithm," *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, vol. 10, no. 1, pp. 14–22, 2025, https://doi.org/10.30591/jpit.v10i1.8180

[3] B. Harnadi and A. D. Widiantoro, "Evaluating the Performance and Accuracy of Supervised Learning Models on Sentiment Analysis of E-Wallet," in *Proc. 7th Int. Conf. Inf. Technol. (InCIT)*, Phuket, Thailand, pp. 175–180, 2023, https://doi.org/10.1109/InCIT60207.2023.10413111

[4] S. Zahra and A. Alamsyah, "Digital Wallet Service Quality Analysis using Multiclass Classification and Sentiment Analysis," in *Proc. 5th Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Yogyakarta, Indonesia, pp. 256–261, 2022, https://doi.org/10.1109/ICOIACT55506.2022.9971936

[5] A. D. Sanjaya, A. K. Ningsih, and F. Renaldi, "Sentiment Analysis of E-Wallets on Twitter Social Media with Naïve Bayes and Lexicon-Based Methods," *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, Rome, Italy, 2023, https://doi.org/10.46254/ap03.20220200

[6] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment Analysis on Customer Satisfaction of Digital Payment in Indonesia: A Comparative Study Using KNN and Naïve Bayes," *J. Phys. Conf. Ser.*, vol. 1444, no. 1, pp. 1–6, 2020, https://doi.org/10.1088/1742-6596/1444/1/012034

[7] D. A. Kristiyanti, D. A. Putri, E. Indrayuni, A. Nurhadi, and A. Umam, "E-Wallet Sentiment Analysis Using Naïve Bayes and Support Vector Machine Algorithm," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 1–6, 2020, https://doi.org/10.1088/1742-6596/1641/1/012079

[8] A. Helmayanti, F. Hamami, and R. Y. Fa'rifah, "Penerapan Algoritma TF-IDF dan Naïve Bayes untuk Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Flip pada Google Play Store," *Jurnal Indonesian Management, Informatics and Communication (JIMIK)*, vol. 4, no. 3, pp. 192–202, 2023, https://doi.org/10.35870/jimik.v4i3.415

[9] N. Nainggolan, B. Sarumaha, J. R. Lumbanbatu, and S. Aisyah, "Analisis Sentimen Pengguna Dompet Digital Menggunakan Algoritma Multivariat Bernoulli (Studi Kasus: OVO dan GoPay)," *Jurnal Teknik Informatika dan Komputer (Tekinkom)*, vol. 7, no. 1, pp. 14–20, 2024, https://doi.org/10.37600/tekinkom.v7i1.1223

[10] F. Kurniawati, A. Wibawa, and A. B. P. Utama, "Sentiment Analysis of Wayang Climen Using Naïve Bayes Method," *Sci. Inf. Technol. Lett. (SITECH)*, vol. 3, no. 2, pp. 45–54, 2022, https://doi.org/10.31763/sitech.v3i2.1220

[11] S. A. Aaputra, D. Rosiyadi, W. Gata, and S. M. Husain, "Sentiment Analysis of E-Wallet Sentiments on Google Play Using the Naïve Bayes Algorithm Based on Particle Swarm Optimization," *J. RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 377–382, 2019, https://doi.org/10.29207/RESTI.V3I3.1118

[12] S. Zahra and A. Alamsyah, "Digital Wallet Service Quality Analysis using Multiclass Classification and Sentiment Analysis," in *Proc. 5th Int. Conf. on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, pp. 256–261, 2022, https://doi.org/10.1109/ICOIACT55506.2022.9971936

[13] A. D. Cahyani and T. Mardiana, "Sentiment Analysis of Digital Wallet Service Users using Naïve Bayes Classifier and Particle Swarm Optimization," *Jurnal Riset Informatika*, vol. 2, no. 4, pp. 241–250, 2020, https://doi.org/10.34288/jri.v2i4.160

[14] M. Luthfi, F. Martanto, and W. Istiono, "Sentiment Analysis of M-Paspor App Reviews Using Multinomial Naive Bayes," *Journal of Logistics, Informatics and Service Science*, vol. 11, no. 1, pp. 80–90, 2024, https://doi.org/10.33168/jliss.2024.1017

[15] Y. Christian, T. Wibowo, and M. Lyawati, "Sentiment Analysis by Using Naïve Bayes Classification and Support Vector Machine, Study Case Sea Bank," *Sinkron*, vol. 9, no. 1, pp. 115–122, 2024, https://doi.org/10.33395/sinkron.v9i1.13141

[16] Y. Astuti, Y. Ruldeviyani, F. Salbari, and A. Prayogi, "Sentiment Analysis of Electricity Company Service Quality Using Naïve Bayes," *Jurnal RESTI*, vol. 7, no. 2, pp. 320–327, 2023, https://doi.org/10.29207/resti.v7i2.4627

[17] J. Setiawan, A. Milenia, and A. Faza, "An Integrated Approach for Sentiment Analysis and Topic Modeling of a Digital Bank in Indonesia using Naïve Bayes and Latent Dirichlet Allocation Algorithms on Social Media Data," in *Proc. 4th Int. Conf. on Big Data Analytics and Practices (IBDAP)*, Jakarta, Indonesia, pp. 1–7, 2023, https://doi.org/10.1109/IBDAP58581.2023.10271956

[18] N. Nainggolan, B. Sarumaha, J. R. Lumbanbatu, and S. Aisyah, "Analisis Sentimen Pengguna Dompet Digital Menggunakan Algoritma Multivariat Bernoulli (Studi Kasus: OVO dan GoPay)," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 1, pp. 14–20, 2024, https://doi.org/10.37600/tekinkom.v7i1.1223

[19] D. A. Putri, D. A. Kristiyanti, E. Indrayuni, A. Nurhadi, and A. Umam, "Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review," *Journal of Physics: Conference Series*, vol. 1641, no. 1, pp. 012085, 2020, https://doi.org/10.1088/1742-6596/1641/1/012085

[20] R. Maulana, M. Raihan, and I. Santoso, "Komparasi Algoritma Naive Bayes dan K-Nearest Neighbor pada Analisis Sentimen terhadap Ulasan Pengguna Aplikasi Tokopedia," *Jurnal Teknologi Informasi*, vol. 7, no. 2, pp. 55–62, 2023, https://doi.org/10.47111/jti.v7i2.10071

[21] E. D. Madyatmadja, M. S. Putri, F. N. Wijaya, and A. A. S. Harahap, "Harmonizing Sentiments: Analyzing User Reviews of Spotify through Sentiment Analysis," *Journal of Infrastructure, Policy and Development*, vol. 8, no. 9, pp. 1–10, 2024, https://doi.org/10.24294/jipd.v8i9.7101

[22] A. Basir, "Analysis of Electronic Wallet User Sentiment on Twitter (X) Social Media Using the Naïve Bayes Classifier Algorithm," *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, vol. 10, no. 1, pp. 14–22, 2025, https://doi.org/10.30591/jpit.v10i1.8180