J-MS
JOURNAL MOBILE TECHNOLOGIES

Received: 15 Jul 2026 | Revised: 24 Aug 2026 | Accepted: 1 Sept 2026 | Published: 10 Sept 2026 E-ISSN: 2986-478X, P-ISSN: 2986-1535

DOI: 10.59431/jms.v4i2.972

RESEARCH ARTICLE

Machine Learning-Based Sentiment Classification of Reviews from Indonesian Mobile Applications Using TF-IDF

Tuti Handayani [1] | Sri Mardiyati [2*]

[1,2*] Department of Informatics Engineering, Universitas Indraprasta PGRI, South Jakarta City, Special Capital Region of Jakarta, Indonesia.

Correspondence

[2*] Department of Informatics Engineering, Universitas Indraprasta PGRI, South Jakarta City, Special Capital Region of Jakarta, Indonesia.

Email: srimardiyati05@gmail.com.

Funding information
This research received no external funding.

Abstract
The increasing volume of mobile application reviews has created a need for automated sentiment classification capable of handling large-scale and imbalanced user-generated data. This study aimed to compare classical machine learning algorithms and evaluate the effectiveness of class-weighted learning in improving minority-class recognition in Indonesian mobile application reviews. The dataset was cleaned, deduplicated, and controlled for conflicting labels and data leakage, resulting in 357,327 unique reviews across Negative, Neutral, and Positive sentiment classes. Textual features were represented using Term Frequency-Inverse Document Frequency (TF-IDF) with 50,000 features, and the data were divided using an 80:20 stratified train-test split. Logistic Regression, Multinomial Naive Bayes, Linear Support Vector Machine (SVM), and Random Forest were evaluated alongside class-weighted Logistic Regression and Linear SVM. Logistic Regression achieved the highest accuracy of 83.47%, whereas Balanced Linear SVM achieved the highest Macro F1-score of 62.87% and improved the Neutral F1-score from 8.70% to 21.18%, while maintaining an accuracy of 79.09%. The findings demonstrate that the highest overall accuracy does not necessarily indicate the most balanced classifier under class imbalance. Balanced Linear SVM provided the most favorable trade-off between overall performance and minority-class recognition, highlighting the importance of class-sensitive evaluation and class-weighted learning for imbalanced sentiment classification.

Keywords
Class Imbalance; Class-Weighted Learning; Indonesian Mobile Application Reviews; Sentiment Analysis; TF-IDF.

1 | INTRODUCTION

The rapid growth of the mobile application ecosystem has increased the importance of user reviews as a valuable source of information for understanding user experiences, satisfaction, complaints, and perceptions of application quality. User-generated reviews provide developers with direct feedback that can support application maintenance, feature improvement, and software evolution, and can be systematically analyzed to identify user requirements, recurring issues, and feature-related concerns (Dabrowski *et al*., 2022). As the volume of user feedback continues to increase, automated review analysis has become increasingly important because manual examination is inefficient and difficult to scale (Assi *et al*., 2025; T. Liu *et al*., 2023). Sentiment analysis provides an effective approach for transforming large volumes of user reviews into meaningful information by classifying user opinions into positive, neutral, and negative sentiments, and in mobile application environments, it can reveal user satisfaction and complaints that may not be adequately represented by numerical ratings alone. Previous studies have demonstrated its usefulness across application domains, including health and wellness applications (Varghese *et al*., 2025) and educational applications (Mondal *et al*., 2022). Nevertheless, sentiment classification remains challenging because user-generated texts contain diverse linguistic expressions and contextual variations that complicate accurate polarity recognition (Birjali *et al*., 2021; Wankhade *et al*., 2022).

Machine learning has been widely applied to sentiment classification because it enables large-scale textual data to be categorized automatically. Since conventional machine learning algorithms require numerical inputs, textual reviews must first be transformed into suitable feature representations. Term Frequency-Inverse Document Frequency (TF-IDF) is widely used for this purpose because it represents terms according to their relative importance within individual documents and the overall corpus (Robertson, 2004; Tang *et al*., 2022). Supervised algorithms such as Logistic Regression (LR), Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Random Forest (RF) have been combined with TF-IDF for text classification (Hassan *et al*., 2022; Solovyeva & Abdullah, 2022), although their performance can vary across datasets, making controlled comparative evaluation necessary for identifying the most appropriate classifier for a particular review corpus (Hassan *et al*., 2022).

Class imbalance remains a major challenge in sentiment classification because unequal class distributions can cause learning algorithms to favor majority classes while providing inadequate recognition of minority categories. This issue is particularly important when neutral sentiment is underrepresented, as high overall accuracy may conceal poor minority-class performance. Therefore, imbalanced classification should be evaluated using metrics beyond accuracy, including precision, recall, and F1-score, to provide a more representative assessment of class-specific performance (Kaope & Pristyanto, 2023; Vluymans, 2019). Various strategies, including undersampling, oversampling, and cost-sensitive learning, have been investigated to mitigate the effects of unequal class distributions (Desuky & Hussain, 2021; Mortaz, 2020), indicating that sentiment classification on imbalanced datasets requires both appropriate learning strategies and evaluation metrics that adequately reflect minority-class recognition.

Despite extensive research on sentiment classification and class-imbalance handling, comparative evidence remains limited regarding the performance of conventional TF-IDF-based classifiers and their class-weighted counterparts under the same experimental setting for reviews from Indonesian mobile applications. Existing studies have predominantly addressed imbalanced classification through data-level strategies such as undersampling and oversampling or through more complex learning mechanisms, which may introduce information loss, synthetic noise, or additional computational complexity (Desuky & Hussain, 2021; Y. Liu *et al*., 2024; Wang *et al*., 2021). Moreover, the effectiveness of imbalance-handling strategies can vary according to data distribution and classification characteristics (Zhao *et al*., 2025). Therefore, a controlled comparison using both overall and class-sensitive evaluation metrics is needed to clarify the trade-off between aggregate classification performance and minority-class recognition.

Accordingly, this study aims to evaluate and compare the performance of Logistic Regression (LR), Multinomial Naive Bayes (MNB), Linear Support Vector Machine (SVM), and Random Forest (RF) for sentiment classification of reviews from Indonesian mobile applications using Term Frequency-Inverse Document Frequency (TF-IDF) feature representation. In addition to the baseline classifiers, class-weighted LR and Linear SVM are evaluated to investigate their effectiveness in addressing class imbalance and improving minority-class recognition. This study provides three main contributions: a controlled comparative evaluation of conventional and class-weighted machine learning models using 357,327 cleaned and unique reviews; a leakage-controlled experimental procedure in which duplicate and conflicting texts are removed prior to stratified data splitting, while TF-IDF is fitted exclusively on the training data; and an evaluation of model performance using overall and class-sensitive metrics, complemented by five-fold stratified cross-validation and McNemar's test, providing a more robust assessment of model stability, prediction differences, and minority-class recognition.

2 | BACKGROUND THEORY

2.1 Sentiment Analysis of Mobile Application Reviews

Sentiment analysis, also known as opinion mining, is a computational approach for identifying and classifying subjective information in textual data into positive, neutral, and negative sentiments (Birjali *et al*., 2021; Wankhade *et al*., 2022). In mobile applications, user reviews provide valuable information regarding user satisfaction, complaints, service quality, and application performance (Hardiansyah *et al*., 2024; Mahmud *et al*., 2022). Sentiment classification enables large volumes of unstructured reviews to be transformed into structured information that supports application evaluation and improvement. However, user-generated reviews often contain contextual expressions, linguistic variations, mixed sentiments, and imbalanced class distributions that may reduce classification performance (Memon *et al*., 2025). Therefore, appropriate text representation and classification methods are essential for reliable sentiment analysis of mobile application reviews.

2.2 TF-IDF Feature Representation

Term Frequency-Inverse Document Frequency (TF-IDF) is a widely used feature representation technique that assigns weights to terms according to their frequency within a document and their importance across the document collection. Term Frequency (TF) measures how frequently a term occurs in a document, whereas Inverse Document Frequency (IDF) reduces the influence of terms that frequently appear across documents (Robertson, 2004). TF-IDF therefore emphasizes terms that are more discriminative for representing textual content and has been widely applied in text classification (Tang *et al*., 2022). However, TF-IDF may produce high-dimensional and sparse feature spaces, which can increase computational and memory costs in text classification tasks (Cunha *et al*., 2020). Despite this limitation, its simplicity, interpretability, and effectiveness make TF-IDF suitable for conventional machine learning-based text classification.

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

Where IDF(t) is defined as:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (2)$$

Where $t$represents a term, $d$represents a document, Ndenotes the total number of documents in the corpus, and $df(t)$represents the number of documents containing term t.

2.3 Machine Learning Algorithms for Text Classification

Several conventional machine learning algorithms are widely applied to text classification. Logistic Regression (LR) is a probabilistic linear classifier that is effective for high-dimensional text representations such as TF-IDF (Hassan *et al*., 2022; Kholwal, 2023). Multinomial Naive Bayes (MNB) estimates class probabilities based on the assumption of feature independence and is computationally efficient for large-scale text classification. Linear Support Vector Machine (SVM) identifies a separating hyperplane with a maximum margin and is particularly suitable for high-dimensional and sparse text features (Hassan *et al*., 2022). Random Forest (RF) is an ensemble classifier that combines multiple decision trees and determines predictions through majority voting, providing robustness across classification tasks (Kholwal, 2023). The performance of these algorithms may vary depending on dataset characteristics, feature representation, and class distribution (Allam *et al*., 2025).

2.4 Class Imbalance and Class-Weighted Learning

Class imbalance occurs when the distribution of classes in a dataset is unequal, causing classification models to favor the majority class while reducing their ability to recognize minority classes (Niaz *et al*., 2022; Rekha & Tyagi, 2020). Consequently, accuracy alone may provide a misleading evaluation, particularly when minority-class recall and F1-score are low (Kaope & Pristyanto, 2023; Vluymans, 2019). Class-weighted and cost-sensitive learning address this problem by assigning greater importance or misclassification costs to underrepresented classes, thereby encouraging the model to give greater attention to minority samples without directly modifying the original class distribution (Das *et al*., 2022). A commonly used balanced class-weighting scheme is expressed as:

$$w_c = \frac{N}{K \times n_c} \quad (3)$$

Where $w_c$represents the weight assigned to class c, Ndenotes the total number of samples, Krepresents the number

of classes, and $n_c$denotes the number of samples belonging to class c. Based on this theoretical framework, the experimental design integrates TF-IDF as the common feature representation and LR, MNB, Linear SVM, and RF as baseline classifiers with different learning characteristics. Because the review corpus exhibits substantial class imbalance, class-weighted variants of LR and Linear SVM are additionally evaluated to examine whether algorithm-level weighting improves minority-class recognition without modifying the original training distribution. Accordingly, the comparison between baseline and class-weighted models is evaluated using both overall and class-sensitive metrics, particularly Macro F1-score and class-specific recall and F1-score. This design directly connects the theoretical concepts of feature representation, classifier characteristics, and imbalanced learning with the empirical evaluation conducted in this study.

2.5 Previous Studies and Theoretical Positioning

Previous studies have applied various machine learning and deep learning approaches to sentiment analysis of mobile application reviews. Algorithms such as Support Vector Machine (SVM), Naive Bayes, and Convolutional Neural Network (CNN) have demonstrated their effectiveness in classifying user sentiments (Varghese *et al*., 2025b). TF-IDF has also been widely combined with conventional machine learning classifiers, including Logistic Regression and Random Forest, to provide effective text representations for sentiment classification (Abbas *et al*., 2024). To address class imbalance, previous research has explored resampling and optimization techniques such as SMOTE, ADASYN, and Particle Swarm Optimization; however, these approaches may modify the original data distribution and increase methodological complexity. Therefore, this study adopts a controlled comparison of conventional and class-weighted classifiers using a consistent TF-IDF representation, with particular attention to minority-class performance and data leakage prevention.

3 | METHOD

3.1 Research Design and Dataset

The sentiment labels used as the target variable were derived deterministically from the numerical star ratings associated with the original user reviews. Reviews with ratings of 1–2 stars were categorized as Negative, reviews with a rating of 3 stars as Neutral, and reviews with ratings of 4–5 stars as Positive. This rating-to-sentiment mapping was consistently applied across the entire initial dataset. Therefore, the sentiment labels were not generated through manual human annotation, and inter-rater agreement measures such as Cohen's kappa were not applicable to the labeling procedure. The initial corpus consisted of reviews from six Indonesian mobile applications representing different service domains and collection periods, as summarized in Table 1. The datasets were pooled to construct a large and heterogeneous corpus for evaluating classifier performance under a common preprocessing, feature-representation, and evaluation framework. The purpose of pooling was not to compare sentiment patterns among individual applications or to estimate application-specific effects, but to assess the behavior of the investigated classifiers across a heterogeneous review corpus. Accordingly, differences among application domains and collection periods should be considered when interpreting the results, and application-specific evaluation is identified as an important direction for future research.

Table 1. Composition of the Initial Review Dataset

Application	Initial Reviews	Review Period
SATUSEHAT	190,696	Mar 2020 – Oct 2021
JMO	200,000	Sep 2021 – Mar 2023
Mobile JKN	122,103	Jun 2016 – Nov 2021
Pertamina	96,291	Aug 2017 – Jul 2022
KAI	7,311	Jul 2014 – Dec 2017
BMKG	1,321	May 2012 – Dec 2015
Total	617,722	May 2012 – Mar 2023

Source: Research data processed by the authors (2026).

Following dataset integration, data cleaning was performed by removing missing and empty review texts, duplicate records, conflicting labels, URLs, mentions, numbers, and non-alphabetic characters, followed by text normalization. Duplicate review texts associated with conflicting sentiment labels were removed before data splitting to prevent ambiguous ground-truth assignments and potential data leakage. After preprocessing and duplicate verification, 357,327 unique reviews remained for analysis. The final dataset exhibited substantial class imbalance, as presented in Table 2.

Table 2. Final Dataset Distribution

Sentiment	Percentage
Negative	59.0%
Neutral	7.1%
Positive	33.9%
Total	100.0%

Source: Research data processed by the authors (2026).

The dataset was divided using an 80:20 stratified split, resulting in 285,861 training and 71,466 testing samples. Duplicate and conflicting texts were removed before splitting, and the final verification showed zero text overlap between the training and testing sets.

3.2 TF-IDF and Classification Models

Textual reviews were represented using TF-IDF with unigram and bigram features, min_df = 2, max_df = 0.95, and a maximum of 50,000 features. The TF-IDF vectorizer was fitted exclusively on the training set to prevent data leakage. Four baseline classifiers—Logistic Regression (LR), Multinomial Naive Bayes (MNB), Linear Support Vector Machine (SVM), and Random Forest (RF)—were evaluated. Class-weighted LR and Linear SVM were additionally examined to address class imbalance. The main model configurations are summarized in Table 3.

Table 3. Model Configuration

Model	Main Configuration
LR	SAGA, max_iter = 1000
MNB	Default
Linear SVM	C = 1.0
RF	100 trees, max_depth = 30
Balanced LR	class_weight = balanced
Balanced Linear SVM	class_weight = balanced

Source: Research data processed by the authors (2026).

3.3 Evaluation

Model performance was evaluated using Accuracy, Macro Precision, Macro Recall, Macro F1-score, Weighted F1-score, class-specific F1-score, confusion matrix, and training time. Because the dataset was imbalanced, Macro F1-score was used as the primary criterion for comparing model performance, while class-specific metrics were examined to assess recognition of the minority Neutral class. All experiments were conducted using Python and Scikit-learn in Google Colab with a fixed random state of 42 where applicable. To assess the stability of the principal model comparison beyond a single train-test split, five-fold stratified cross-validation was additionally performed on the training set for Logistic Regression and Balanced Linear SVM. The held-out 20% test set was retained exclusively for final evaluation. In addition, McNemar's test with continuity correction was applied to the paired predictions of Logistic Regression and Balanced Linear SVM on the same held-out test set to determine whether their prediction differences were statistically significant. Statistical significance was assessed at the 0.05 level.

4 | RESULTS AND DISCUSSION

4.1 Results

After data cleaning and preprocessing, 357,327 unique reviews were retained from the initial 617,722 records. As summarized in Table 4, the final dataset exhibited substantial class imbalance, with Negative reviews accounting for 59.0%, Positive reviews for 33.9%, and Neutral reviews for only 7.1%. The stratified 80:20 split resulted in 285,861 training samples and 71,466 testing samples while preserving the class distribution, and no overlapping cleaned texts were identified between the training and testing sets, confirming that model evaluation was conducted under a leakage-controlled setting. Given the relatively small proportion of Neutral reviews, Macro F1-score and class-specific F1-score were considered alongside overall accuracy as particularly important criteria for comparing classifier performance.

Table 4. Dataset Characteristics

Characteristic	Value
Final reviews	357,327
Training samples	285,861
Testing samples	71,466
Negative	59.0%
Neutral	7.1%
Positive	33.9%
TF-IDF features	50,000

Source: Research data processed by the authors (2026).

The performance of the four baseline classifiers and two class-weighted configurations is presented in Table 5. Logistic Regression achieved the highest accuracy of 0.8347 and the highest Weighted F1-score of 0.8081; however, its Neutral F1-score was only 0.0636, indicating limited recognition of the minority class. Linear SVM produced a slightly lower accuracy of 0.8272 but achieved a higher Macro F1-score of 0.5967 and a Neutral F1-score of 0.0870. Random Forest exhibited the weakest overall performance, with a Macro F1-score of 0.4616 and a Neutral F1-score of 0.0000, whereas Balanced Linear SVM achieved the highest Macro F1-score of 0.6287 and the highest Neutral F1-score of 0.2118, indicating a more favorable balance between overall classification performance and minority-class recognition.

Table 5. Performance Comparison of Classification Models

Model	Accuracy	Macro F1	Weighted F1	Neutral F1
Logistic Regression	0.8347	0.5931	0.8081	0.0636
Multinomial Naive Bayes	0.8166	0.5642	0.7858	0.0240
Linear SVM	0.8272	0.5967	0.8043	0.0870
Random Forest	0.7253	0.4616	0.6733	0.0000
Balanced Logistic Regression	0.6081	0.5227	0.6359	0.2041
Balanced Linear SVM	0.7909	0.6287	0.7979	0.2118

Source: Research data processed by the authors (2026).

To provide a clearer comparison across evaluation metrics, the performance of all six model configurations is visualized in Figure 1.

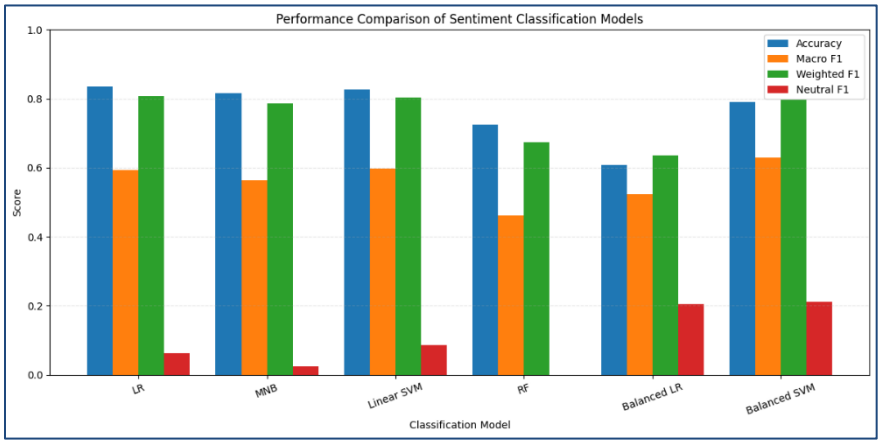

Figure 1. Performance Comparison of Sentiment Classification Models

As illustrated in Figure 1, Logistic Regression achieved the highest accuracy and Weighted F1-score, whereas Balanced Linear SVM obtained the highest Macro F1-score and Neutral F1-score. The results indicate a clear trade-off between overall predictive accuracy and balanced performance across sentiment classes. Although class weighting reduced the overall accuracy of Linear SVM, it substantially improved the recognition of the minority Neutral class, and Balanced Linear SVM therefore provided the most favorable performance when class balance was considered. To examine whether the observed performance was stable beyond the single train-test split, five-fold stratified cross-validation was conducted on the training set for Logistic Regression and Balanced Linear SVM. Logistic Regression achieved a mean accuracy of 0.8349 ± 0.0015, Macro F1-score of 0.5918 ± 0.0007, and Weighted F1-score of 0.8079 ± 0.0014, while Balanced Linear SVM achieved a mean accuracy of 0.7912 ± 0.0013, Macro F1-score of 0.6265 ± 0.0016, and Weighted F1-score of 0.7970 ± 0.0012. These results were consistent with the

corresponding held-out test-set performance, indicating that the relative performance patterns of both classifiers were stable across the training folds.

Table 6. Five-Fold Stratified Cross-Validation Results

Model	Accuracy	Macro F1	Weighted F1
Logistic Regression	0.8349 ± 0.0015	0.5918 ± 0.0007	0.8079 ± 0.0014
Balanced Linear SVM	0.7912 ± 0.0013	0.6265 ± 0.0016	0.7970 ± 0.0012

Source: Research data processed by the authors (2026).

McNemar's test was subsequently conducted on the paired predictions of Logistic Regression and Balanced Linear SVM using the same 71,466 held-out test observations. Both classifiers correctly predicted 54,900 observations, whereas both incorrectly predicted 10,188 observations. Among the discordant predictions, Logistic Regression was correct while Balanced Linear SVM was incorrect for 4,755 observations, whereas Balanced Linear SVM was correct while Logistic Regression was incorrect for 1,623 observations. This difference in prediction outcomes was statistically significant (McNemar's $\chi^2 = 1537.0274$, $p < 0.001$), confirming a statistically significant difference in the paired prediction outcomes of the two classifiers. The application of class weighting produced different effects on Logistic Regression and Linear SVM. For Logistic Regression, the Neutral F1-score increased from 0.0636 to 0.2041 after class weighting, but this improvement was accompanied by a substantial decrease in accuracy from 0.8347 to 0.6081 and Macro F1-score from 0.5931 to 0.5227, indicating that class weighting increased the model's sensitivity toward the minority Neutral class while considerably reducing its overall classification performance. A more favorable trade-off was observed for Linear SVM, whose Neutral F1-score increased from 0.0870 to 0.2118 and Macro F1-score from 0.5967 to 0.6287, although accuracy decreased from 0.8272 to 0.7909; unlike Balanced Logistic Regression, Balanced Linear SVM therefore improved both minority-class recognition and balanced overall performance. These findings indicate that the effect of class weighting varies across classifiers, and in the present dataset, Balanced Linear SVM provided the most favorable trade-off between overall predictive performance and minority-class recognition. Based on Macro F1-score, Balanced Linear SVM provided the most balanced class-level performance among the evaluated configurations. Its class-specific performance is presented in Table 7.

Table 7. Class-Specific Performance of Balanced Linear SVM

Sentiment	Precision	Recall	F1-score	Support
Negative	0.8643	0.8514	0.8578	42,161
Neutral	0.1862	0.2454	0.2118	5,077
Positive	0.8339	0.8000	0.8166	24,228

Source: Research data processed by the authors (2026).

The model performed strongly for Negative and Positive sentiments, achieving F1-scores of 0.8578 and 0.8166, respectively, while Neutral remained the most difficult class, with an F1-score of 0.2118. Nevertheless, the substantial improvement compared with the baseline Linear SVM indicates that class weighting increased the model's sensitivity to minority-class observations. The remaining performance gap may reflect greater linguistic ambiguity and overlapping characteristics between Neutral reviews and the Positive and Negative classes. To further evaluate the classification behavior of Balanced Linear SVM, a confusion matrix was analyzed to identify correct predictions and misclassification patterns across the three sentiment classes, as presented in Figure 2.

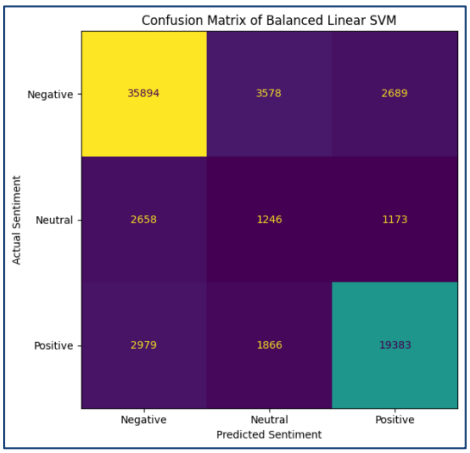

Figure 2. Confusion Matrix of Balanced Linear SVM

As shown in Figure 2, Balanced Linear SVM correctly classified 35,894 Negative, 1,246 Neutral, and 19,383 Positive reviews. The model demonstrated strong classification performance for the Negative and Positive classes, while Neutral remained the most difficult class to identify: of the 5,077 Neutral reviews, 2,658 were incorrectly classified as Negative and 1,173 as Positive. These results indicate that class weighting improved minority-class recognition compared with the baseline Linear SVM; however, the relatively high misclassification rate of Neutral reviews suggests that class imbalance and overlapping linguistic characteristics remain important challenges. This pattern indicates that, from a practical application-monitoring perspective, the classifier should not be used as the sole mechanism for automatically prioritizing or filtering user feedback, since genuinely Neutral feedback may be incorrectly interpreted as expressing satisfaction or dissatisfaction, potentially distorting aggregate sentiment summaries and the prioritization of user concerns. Therefore, for operational app-review monitoring, predictions for the Neutral class should be interpreted cautiously and may benefit from supplementary manual review or confidence-based screening, particularly when sentiment outputs are used to support service evaluation or application-improvement decisions.

4.2 Discussion

The experimental results demonstrate that high overall accuracy does not necessarily indicate balanced sentiment classification, particularly when the dataset exhibits substantial class imbalance. Logistic Regression achieved the highest accuracy of 83.47%, but its F1-score for the Neutral class was only 6.36%. In contrast, Balanced Linear SVM achieved a lower accuracy of 79.09%, while producing the highest Macro F1-score of 62.87% and the highest Neutral F1-score of 21.18%. This finding supports previous studies emphasizing that accuracy alone may be insufficient for evaluating classifiers on imbalanced datasets and that class-sensitive measures such as precision, recall, and F1-score should also be considered (Imran *et al*., 2014; Mortaz, 2020). The results also show that class weighting did not affect all classifiers equally. Applying class weights to Logistic Regression improved Neutral-class recognition but reduced its accuracy and Macro F1-score considerably, whereas class weighting improved the Macro F1-score of Linear SVM from 59.67% to 62.87% and the Neutral F1-score from 8.70% to 21.18%. This result indicates that imbalance-handling strategies should be evaluated in relation to the characteristics of individual classifiers rather than assumed to provide uniform improvements, an observation consistent with previous research showing that the effectiveness of imbalance-handling techniques varies according to data characteristics and classification methods (Desuky & Hussain, 2021).

The strong performance of Linear SVM is also relevant to the use of TF-IDF representation in this study. TF-IDF generates a sparse and high-dimensional representation of textual data, for which linear classifiers are generally effective, and previous comparative studies have similarly reported competitive performance of SVM and Logistic Regression in text classification tasks (Hassan *et al*., 2022; Solovyeva & Abdullah, 2022). In the present study, however, the contribution extends beyond identifying the classifier with the highest accuracy by examining how class weighting changes model behavior under substantial sentiment imbalance. Despite these improvements, the confusion matrix reveals that Neutral sentiment remains difficult to distinguish. Balanced Linear SVM correctly classified only 1,246 of 5,077 Neutral reviews, while 2,658 were classified as Negative and 1,173 as Positive. This suggests that the challenge cannot be attributed solely to numerical class imbalance; Neutral reviews may also contain lexical patterns that overlap with positive and negative expressions, limiting the discriminative capability of TF-IDF-based representations. Consequently, improving minority-class recognition may require methods that capture richer contextual and semantic information in addition to class-balancing strategies.

The principal contribution of this study lies in demonstrating, on a large dataset of 357,327 Indonesian mobile application reviews, that the model with the highest accuracy is not necessarily the most appropriate model when minority-class recognition is important. Balanced Linear SVM provided the most favorable trade-off between overall predictive performance and balanced sentiment recognition, highlighting the importance of combining data-leakage control, class-sensitive evaluation metrics, and class-weighted learning when developing sentiment classification models for large-scale and imbalanced user-review datasets.

5 | CONCLUSIONS AND FUTURE WORK

This study evaluated the performance of classical machine learning classifiers for sentiment classification on a large-scale and imbalanced dataset of reviews from Indonesian mobile applications. After data cleaning, conflict removal, and leakage control, 357,327 unique reviews were retained and represented using TF-IDF with 50,000 features. The results showed that Logistic Regression achieved the highest accuracy of 83.47% and Weighted F1-score of 80.81%. However, its ability to recognize the minority Neutral class was limited, with an F1-score of only 6.36%. In comparison, Balanced Linear SVM achieved the highest Macro F1-score of 62.87% and improved the Neutral F1-score from 8.70% to 21.18%, while maintaining an accuracy of 79.09%. Five-fold stratified cross-

validation further confirmed the stability of this performance pattern, with Logistic Regression achieving a mean accuracy of 83.49% ± 0.15% and Balanced Linear SVM achieving a higher mean Macro F1-score of 62.65% ± 0.16%. McNemar's test also revealed a statistically significant difference between their paired prediction outcomes (χ^2 = 1537.0274, $p < 0.001$), providing further evidence that the observed performance differences were not merely descriptive. These findings demonstrate that the classifier with the highest overall accuracy is not necessarily the most effective when balanced performance across sentiment classes is required.

The main contribution of this study is the empirical demonstration that class-weighted learning combined with class-sensitive evaluation provides a more informative assessment of sentiment classification under substantial class imbalance. Balanced Linear SVM offered the most favorable trade-off between overall predictive performance and minority-class recognition. From a practical perspective, this approach can support large-scale analysis of mobile application reviews by enabling developers and service providers to identify user sentiment more systematically, including feedback from less-represented sentiment categories.

Despite these findings, Neutral sentiment remained difficult to classify, indicating that class weighting alone was insufficient to achieve strong minority-class recognition in the investigated dataset. Consequently, Neutral predictions should be interpreted cautiously in operational app-review monitoring, particularly when automated sentiment outputs are used to support service evaluation or application-improvement decisions. The use of TF-IDF also limits the representation of contextual and semantic relationships between words. Future research should therefore investigate contextual language representations such as IndoBERT or multilingual transformer-based models, alternative imbalance-handling techniques, and hybrid approaches combining class weighting with resampling or cost-sensitive learning. Further studies could also examine sentiment performance separately across different application domains and incorporate aspect-based sentiment analysis to identify not only sentiment polarity but also the specific application features associated with user satisfaction and dissatisfaction.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the academic and technical support provided during the completion of this research, as well as the constructive feedback and suggestions that helped improve the quality of the study and manuscript.

REFERENCES

Abbas, S., Boulila, W., Driss, M., Sampedro, G. A., Abisado, M. B., & Almadhor, A. S. (2024). Active learning empowered sentiment analysis: An approach for optimizing smartphone customer's review sentiment classification. *IEEE Transactions on Consumer Electronics*. https://doi.org/10.1109/tce.2023.3328878

Allam, H., Makubvure, L., Gyamfi, B., Graham, K. N., & Akinwolere, K. (2025). Text classification: How machine learning is revolutionizing text categorization. *Information*, *16*(2), 130. https://doi.org/10.3390/info16020130

Assi, M., Hassan, S., & Zou, Y. (2025). LLM-Cure: LLM-based competitor user review analysis for feature enhancement. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3744644

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M. A., & Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*, *57*(4), 102263.

Dabrowski, J., Letier, E., Perini, A., & Susi, A. (2022). Analysing app reviews for software engineering: A systematic literature review. *Empirical Software Engineering*, *27*(2), 1–63. https://doi.org/10.1007/s10664-021-10065-7

Das, S., Mullick, S. S., & Zelinka, I. (2022). On supervised class-imbalanced learning: An updated perspective and some key challenges. *IEEE Transactions on Artificial Intelligence*, *3*(6), 973–993. https://doi.org/10.1109/TAI.2022.3160658

Desuky, A. S., & Hussain, S. (2021). An improved hybrid approach for handling class imbalance problem. *Arabian Journal for Science and Engineering*, *46*(4), 1–12. https://doi.org/10.1007/S13369-021-05347-7

Hardiansyah, D., Aziz, R. Z. A., & Hasibuan, M. A. (2024). The classification method is used for sentiment analysis in My Telkomsel. *International Journal of Artificial Intelligence Research*, *8*(2), 169. https://doi.org/10.29099/ijair.v8i2.1229

Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, *3*, 238–248. https://doi.org/10.1016/j.susoc.2022.03.001

Imran, M., Mahmood, A. M., & Qyser, A. A. M. (2014). An empirical experimental evaluation on imbalanced data sets with varied imbalance ratio. *International Conference on Computer Communications*, 1–7. https://doi.org/10.1109/ICCCT2.2014.7066742

Kaope, C., & Pristyanto, Y. (2023). The effect of class imbalance handling on datasets toward classification algorithm performance. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *22*(2), 227–238. https://doi.org/10.30812/matrik.v22i2.2515

Kholwal, R. (2023). Text-classify: A comprehensive comparative study of logistic regression, random forest, and KNN models for enhanced text classification performance. Zenodo. https://doi.org/10.5281/zenodo.10148008

Liu, T., Wang, C., Huang, K., Liang, P., Zhang, B., Daneva, M., & van Sinderen, M. (2023). RoseMatcher: Identifying the impact of user reviews on app updates. *Information and Software Technology*, *161*, 107261. https://doi.org/10.1016/j.infsof.2023.107261

Liu, Y., Du, G., Yin, C., Zhang, H., & Wang, J. (2024). Clustering-based incremental learning for imbalanced data classification. *Knowledge-Based Systems*, *292*, 111612. https://doi.org/10.1016/j.knosys.2024.111612

Mahmud, Md. S., Bonny, A. J., Saha, U., Jahan, M., Tuna, Z. F., & Marouf, A. Al. (2022). Sentiment analysis from user-generated reviews of ride-sharing mobile applications. *International Conference Computing Methodologies and Communication*, 738–744. https://doi.org/10.1109/ICCMC53470.2022.9753947

Memon, S. A., Mahar, M. A., Kehar, A., Oad, S., & Ahmed, I. (2025). User experience enhancement through sentiment analysis: A machine learning approach to app reviews. *International Journal of Information Systems and Computer Technologies*. https://doi.org/10.58325/ijisct.004.02.00131

Mondal, A. S., Zhu, Y., Bhagat, K. K., & Giacaman, N. (2022). Analysing user reviews of interactive educational apps: A sentiment analysis approach. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2022.2086578

Mortaz, E. (2020). Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems*, *210*, 106490. https://doi.org/10.1016/j.knosys.2020.106490

Niaz, N. U., Shahariar, K. M. N., & Patwary, M. J. A. (2022). Class imbalance problems in machine learning: A review of methods and future challenges. *Proceedings of the 2nd International Conference on Computing Advancements*, 485–490.

Rekha, G., & Tyagi, A. K. (2020). Necessary information to know to solve class imbalance problem: From a user's perspective (pp. 645–658). https://doi.org/10.1007/978-3-030-29407-6_46

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. https://doi.org/10.1108/00220410410560582

Solovyeva, E. B., & Abdullah, A. S. (2022). Comparison of different machine learning approaches to text classification. 1427–1430. https://doi.org/10.1109/ElConRus54750.2022.9755806

Tang, Z., Li, W., & Li, Y. (2022). An improved supervised term weighting scheme for text representation and classification. *Expert Systems with Applications*, *189*, 115985. https://doi.org/10.1016/j.eswa.2021.115985

Varghese, L., Pai, R. R., Savitha, G., Girisha, S., & Chetty, N. (2025). Manual annotation based sentiment analysis of user feedback in health and wellness app reviews. *Scientific Reports*, *15*(1), 44766. https://doi.org/10.1038/s41598-025-28799-5

Vluymans, S. (2019). Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods (Vol. 807). Springer International Publishing. https://doi.org/10.1007/978-3-030-04663-7

Wang, L., Wang, H., & Fu, G. (2021). Multiple kernel learning with minority oversampling for classifying imbalanced data. *IEEE Access*, *9*, 565–580. https://doi.org/10.1109/ACCESS.2020.3046604

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Zhao, L., Han, F., Ling, Q., Han, H., Yao, Z., Liu, W., & Zhou, Z. (2025). A survey on class imbalance learning algorithms in complex scenarios. *IEEE Access*, *13*, 180799–180833. https://doi.org/10.1109/access.2025.3618909.

How to cite this article: Handayani, T., & Mardiyati, S. (2026). Machine Learning-Based Sentiment Classification of Reviews from Indonesian Mobile Applications Using TF-IDF. *Journal Mobile Technologies (JMS)*, *4*(2), 95–105. https://doi.org/10.59431/jms.v4i2.972.