Indonesian Applied Research Computing and Informatics
Vol. 2; No. 1, July 2026 Page 24-35;
https://jurnal.tdinus.com/index.php/iarci
e-ISSN : 3110-8806

Addressing Class Imbalance in Sentiment Analysis of Klik Indomaret Reviews Using SMOTE and Random Forest

Sutriawan[1*], M. Fitrah Nur[1], Irman[1], Wahyu Tisno Atmojo[2], Yo Ceng Giap[3]

[1] Department of Computer Science, Universitas Muhammadiyah Bima, Bima 84113, Indonesia
[2] Department of Information System,Universitas Pradita. Tangerang 15810, Indonesia
[3] Department of Informatics Engineering, Universitas Buddhi Dharma, Tangerang, Banten, 15115, Indonesia

Email : sutriawan@umbima.ac.id[1*], fitranour40@gmail.com[1], irman@umbima.ac.id[1], wahyu.tisno@pradita.ac.id[2], cenggiap@ubd.ac.id[3]

Article Info

Article history:
Received:06-16-2026
Revised: 07-21-2026
Accepted: 07-27-2026

Keywords:
sentiment analysis,
Klik Indomaret,
Google Play Store,
Random Forest,
SMOTE

ABSTRACT

User reviews provide valuable insights into customer satisfaction and application performance. This study analyzes sentiment in Klik Indomaret reviews collected from the Google Play Store using Random Forest and SMOTE to address class imbalance. A total of 3,097 reviews from March 2021 to June 2024 were scraped and labeled into positive and negative sentiment classes based on user ratings. The review texts were preprocessed through case folding, cleaning, tokenization, stopword removal, and stemming, followed by feature extraction using TF-IDF. The dataset was split into training and testing subsets with a 75:25 ratio, and SMOTE was applied to the training data to balance class distribution. The results show that the baseline Random Forest model achieved an accuracy, precision, recall, and F1-score of 0.84, while the RF-SMOTE model improved these metrics to 0.85. These findings indicate that balancing the training data enhances classification performance and reduces bias toward the majority class. The proposed TF-IDF, SMOTE, and Random Forest framework provides an effective approach for sentiment analysis of mobile application reviews and can support service improvement based on user feedback.

This is an open access article under the CC BY-SA license.

*Corresponding Author:
Sutriawan
Department of Computer Science, Universitas Muhammadiyah Bima, Bima 84113, Indonesia
Email: sutriawan@umbima.ac.id

1. Introduction

User reviews on mobile app stores have become one of the most useful sources for understanding product quality, usability, and customer satisfaction. In the Google Play Store, reviews provide direct feedback from users who have experienced the application in real use, making them especially valuable for developers who need to identify strengths, weaknesses, and recurring complaints. Several recent studies show that sentiment

https://doi.org/10.64479/iarci.v2i1.160

analysis on app reviews can turn this unstructured feedback into actionable insight, whether the target application is a public service app, an e-wallet, or a consumer platform [1], [2].

The main challenge in app-review sentiment analysis is that review data is often noisy, short, informal, and imbalanced. App users may write in mixed language, use slang, omit punctuation, or give very uneven numbers of positive and negative reviews. This is why many studies combine text preprocessing, TF-IDF feature extraction, and machine learning classifiers such as Random Forest. In one recent study, the pipeline explicitly combined preprocessing, TF-IDF, and SMOTE before Random Forest classification, and Random Forest produced the strongest performance among the compared models [2]. Google Play study reported that Random Forest achieved an accuracy of 89% on PrimaKu reviews after hyperparameter tuning, which suggests that this family of models is competitive for app-review data [3].

Class imbalance is not a minor technical issue; it can change what the model learns. When one sentiment class dominates, a classifier may appear accurate while failing to recognize the minority class well. Recent work on imbalanced text classification describes class imbalance as a common and critical challenge, and oversampling methods such as SMOTE remain among the most widely used solutions for balancing training data [4], [5]. In practice, studies on Google Play review data repeatedly report that SMOTE helps the classifier recover minority-class patterns more effectively than training on the original skewed distribution [6], [7].

This pattern appears across domains. In the PeduliLindungi study, Random Forest with SMOTE reached 71% accuracy, 70% recall, and 70% precision, showing that balancing the classes can support more stable sentiment detection on government-app review data [8], [9]. In the IKD study, the authors split the data into 80% training and 20% testing, then reported improved behavior after SMOTE, including higher precision in the balanced setup [1]. Similar results were reported in e-wallet sentiment classification, where TF-IDF, SMOTE, and Random Forest were described as effective and reliable together [10].

Against that background, the present study focuses on Klik Indomaret, an application whose reviews reflect both functional complaints and positive shopping experiences. The goal is to test whether a standard text-classification pipeline, combined with SMOTE, can improve sentiment classification on this dataset. The study contributes a practical case of Indonesian-language app-review mining and shows how class balancing changes model performance in a real review stream. The literature suggests this approach is well grounded, but the effect still depends on the quality of preprocessing, the balance of the labels, and the expressiveness of the features used [11], [12].

Recent app-review studies consistently show that Google Play data can be mined effectively for sentiment classification. In the SIGNAL application study, the authors used preprocessing, TF-IDF, and SMOTE, and Random Forest achieved the best overall performance among several classifiers, with 91.04% accuracy and 94.80% F1-score [2]. The IKD study followed a similar design on 28,134 reviews and showed that Random Forest remained competitive even before balancing, then changed again after SMOTE was applied [1]. These studies are useful because they confirm that the same pipeline generalizes across app categories rather than only one product type.

Studies on other public or consumer apps point in the same direction. The PeduliLindungi paper reported that Random Forest and SMOTE yielded 71% accuracy, which is lower than some later studies but still shows that sentiment structure can be captured even in a difficult dataset[8], [9]. The Shopee study used TF-IDF and Random Forest on app reviews and reported a higher accuracy of 94%, suggesting that dataset composition and text cleanliness can strongly affect outcomes [11]. In the PrimaKu study, Random Forest with tuned parameters reached 89% accuracy, again supporting the idea that random-forest classifiers are robust baselines for Bahasa Indonesia app reviews [3].

Other app domains report similar results. The PLN Mobile case study used Random Forest with SMOTE to handle imbalance, and the authors explicitly concluded that the approach improved sentiment-analysis performance on unbalanced data [7]. The e-wallet study across GoPay, OVO, and DANA also found that TF-IDF, SMOTE, and Random Forest worked reliably across multiple applications, which strengthens the argument that the method is not app-specific but rather a useful general pipeline for review analytics [10]. The broader oversampling literature agrees, noting that class imbalance is a central problem in text classification and that SMOTE remains a standard response when the minority class matters [4]. [5], [6].

Taken together, these studies support three points. First, Google Play reviews are a practical dataset for application evaluation. Second, TF-IDF and Random Forest remain strong and interpretable choices for Indonesian sentiment tasks. Third, SMOTE often helps when the label distribution is uneven, although the size of the gain depends on the dataset and the feature representation. The present study on Klik Indomaret extends this line of work by applying the same logic to a different retail application, where user feedback is shaped by stock availability, service response, and shopping convenience.

2. Methodology

The proposed method follows a standard sentiment-analysis workflow that is consistent with prior app-review studies using TF-IDF, SMOTE, and Random Forest [2], [7], [10].

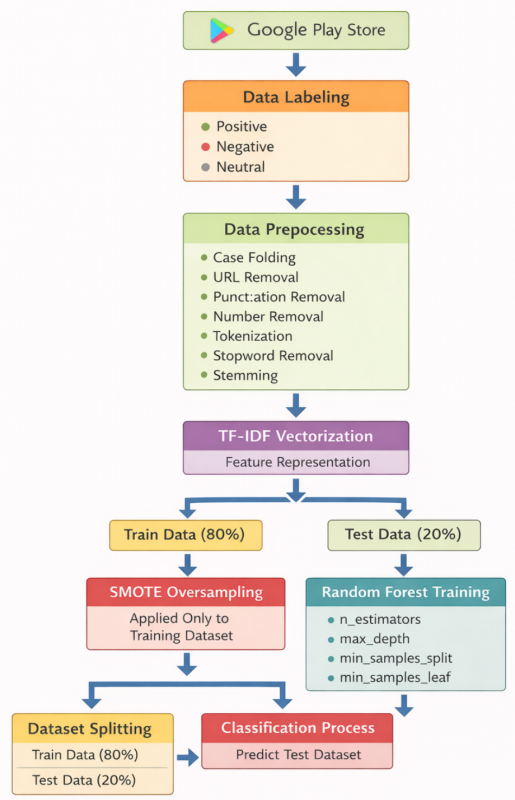


Figure 1. Detailed Proposed Method for Sentiment Classification Using TF-IDF, SMOTE, and Random Forest.

2.1. Data Scraping

The dataset was collected from user reviews on the Google Play Store using the google-play-scraper library. A total of 3,097 reviews were obtained, including review text, rating scores, and review dates from March 2021 to June 2024 [13], [14], [15]. Let the collected dataset be represented as:

$$D = \{(r_i, s_i)\}_{i=1}^{N} \tag{1}$$

where:
r_i denotes the review text,
s_i denotes the rating score,
$N = 3097$ represents the total number of reviews.

2.2. Sentiment Labeling

Each review was assigned a sentiment label based on its rating score. Ratings from 1 to 3 were categorized as negative, while ratings from 4 to 5 were categorized as positive [16], [17], [18]. The labeling function is defined as:

$$y_i = \begin{cases} Negative, & s_i \leq 3 \\ Positive, & s_i \geq 4 \end{cases} \tag{2}$$

where y_i represents the sentiment class of review i. The resulting dataset can be expressed as:

$$D = \{(r_i, y_i)\}_{i=1}^{N} \tag{3}$$

2.3. Text Preprocessing

Text preprocessing was performed to normalize the review text before feature extraction. The preprocessing stage consisted of:
Case Folding, Cleaning (removal of URLs, punctuation, symbols, and numbers), Tokenization,

Stopword Removal, Stemming [13], [19]. Given a review document r_i, preprocessing transforms it into:

$$r_i \rightarrow T_i = \{w_1, w_2, \ldots, w_n\} \quad (4)$$

where T_i represents the cleaned sequence of tokens. This normalization process reduces noise and improves the quality of the extracted textual features.

2.4. TF-IDF Vectorization

After preprocessing, the textual data were transformed into numerical vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme [8], [20], [21]. The Term Frequency (TF) is computed as:

$$TF(t,d) = \frac{f(t,d)}{\sum_k f(k,d)} \quad (5)$$

where:
$f(t,d)$is the frequency of term $t$in document d. The Inverse Document Frequency (IDF) is calculated as:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (6)$$

where:
$N$is the total number of documents,
$df(t)$is the number of documents containing term t.
The final TF-IDF weight is defined as:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \quad (7)$$

The output of this stage is a sparse feature matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nm} \end{bmatrix} \quad (8)$$

where $m$denotes the number of unique terms.

2.5. Dataset Splitting

The TF-IDF feature matrix was divided into training and testing subsets using a 75:25 ratio.

$$| D_{train} | = 0.75N \quad (9)$$
$$| D_{test} | = 0.25N \quad (10)$$

The training set was used to build the classification model, while the testing set was reserved for model evaluation.

2.6. SMOTE Oversampling

Class imbalance is a common issue in sentiment classification datasets. To address this problem, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training dataset [22], [23], [24]. SMOTE generates synthetic minority samples according to:

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad (11)$$

where:
$x_i$is a minority-class instance,
$x_{nn}$is one of its nearest neighbors,
$\lambda \in [0,1]$is a random number.
The purpose of SMOTE is to balance the class distribution:

$$N_{minority} \approx N_{majority} \quad (12)$$

thereby reducing classification bias toward the majority class.

2.7. Random Forest Classification

Random Forest was employed as the classification algorithm due to its robustness, ability to handle high-dimensional feature spaces, and strong performance in text classification tasks. A Random Forest consists of multiple decision trees trained using bootstrap samples. The entropy of a dataset S is calculated as:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (13)$$

where $p_i$represents the probability of class i. Information Gain is computed as:

$$IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (14)$$

where $A$is a candidate feature used for splitting. The final prediction is determined through majority voting:

$$\hat{y} = mode(h_1(x), h_2(x), \ldots, h_k(x)) \quad (15)$$

where:
$h_j(x)$denotes the prediction of the j-th tree,
$k$is the number of trees in the forest.

2.8. Performance Evaluation

Model performance was evaluated using a confusion matrix and four standard classification metrics: accuracy, precision, recall, and F1-score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (19)$$

where:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The confusion matrix was also analyzed to provide a detailed understanding of the model's classification performance across sentiment classes.

3. Results

3.1. Data Collection Result

The first stage of this study involved collecting user reviews of the *Klik Indomaret* mobile application from the Google Play Store. Data acquisition was performed using a web-scraping technique, resulting in a total of 3,097 reviews collected between March 2021 and June 2024. Each review record contained textual comments, rating scores, and review dates, providing a comprehensive dataset for sentiment analysis.

Table 1 presents examples of the collected reviews. The dataset contains diverse opinions regarding various aspects of the application, including product availability, promotional offers, shopping convenience, delivery services, and customer support quality. Reviews with low ratings generally

expressed dissatisfaction related to stock availability and customer service responsiveness, whereas reviews with high ratings frequently highlighted the convenience of online shopping, promotional discounts, and overall user satisfaction.

The collected reviews represent authentic user experiences and provide valuable insights into customer perceptions of the Klik Indomaret application. Therefore, the dataset serves as a reliable source for evaluating user sentiment and identifying strengths and weaknesses of the application from the customer perspective.

Table 1. Sample of Collected Reviews

No	Review (Original ID)	Review (Translated into English)
1	*Beli air galon Indomaret via aplikasi kosong stock terus, padahal di toko stock melimpah. Menghubungi CS via email tidak ada balasan atau tanggapan.*	I ordered a gallon of water through the Indomaret application, but it was constantly marked as out of stock, even though the product was abundantly available in the store. I contacted customer service via email, but received neither a response nor any follow-up.
2	*Baru tahu ternyata belanja di Klik Indomaret bikin ketagihan. Bisa cek stok dan harga dari rumah, apalagi saat promo ditambah diskon sehingga lebih hemat.*	I recently discovered that shopping through Klik Indomaret is quite addictive. I can check product availability and prices from home, and during promotional periods, additional discounts make shopping even more economical.
3	*Aplikasi sering error saat checkout dan pembayaran sehingga transaksi tidak dapat diselesaikan dengan baik.*	The application frequently experiences errors during checkout and payment processing, causing transactions to fail and preventing purchases from being completed successfully.
4	*Sangat membantu untuk belanja kebutuhan sehari-hari tanpa harus datang langsung ke toko.*	The application is very helpful for purchasing daily necessities without having to visit the physical store.
5	*Pelayanan cukup baik, tetapi proses pengiriman terkadang lebih lama dari yang dijanjikan.*	The service is generally good; however, the delivery process occasionally takes longer than the promised delivery time.

```
scrape_ps, token=reviews(
    'com.indomaret.klikindomaret',
    lang='id',
    country='ID',
    sort=Sort.MOST_RELEVANT,
    count=3500,
)
df_ps = pd.DataFrame(scrape_ps)
```

Figure 2 illustrates the process used to collect review data from the Google Play Store.

3.2. Sentiment Labeling Results

Following data collection, the reviews were assigned sentiment labels based on the rating scores provided by users. Reviews with ratings ranging from 1 to 3 were categorized as negative sentiment, while reviews with ratings ranging from 4 to 5 were categorized as positive sentiment. This labeling strategy transformed the raw review data into a supervised learning dataset suitable for sentiment classification.

Table 2 presents examples of the labeled reviews. The assigned sentiment labels generally correspond to the opinions expressed in the review texts. Reviews describing application errors, poor customer service, or product availability issues were classified as negative sentiment, whereas reviews expressing satisfaction with convenience, promotions, and overall shopping experience were classified as positive sentiment.

Table 2. Examples of Sentiment Labeling

No	Original Text (Indonesian)	Translated Text (English)	Rating	Sentiment Label

1	Beli air galon Indomaret via aplikasi kosong stock terus, padahal di toko stock melimpah. Menghubungi CS via email tidak ada balasan atau tanggapan.	I ordered a gallon of water through the Indomaret application, but it was constantly marked as out of stock even though it was available in the store. Customer service did not respond to my email inquiries.	2	Negative
2	Baru tahu ternyata belanja di Klik Indomaret bikin ketagihan. Bisa cek stok dan harga dari rumah, apalagi saat promo ditambah diskon sehingga lebih hemat.	Shopping through Klik Indomaret is very convenient. Users can check product availability and prices from home, and promotional discounts make purchases more economical.	5	Positive
3	Aplikasi sering error saat checkout dan pembayaran sehingga transaksi tidak dapat diselesaikan dengan baik.	The application frequently experiences errors during checkout and payment processing, causing transactions to fail.	1	Negative
4	Sangat membantu untuk belanja kebutuhan sehari-hari tanpa harus datang langsung ke toko.	The application is very helpful for purchasing daily necessities without visiting the physical store.	5	Positive
5	Pelayanan cukup baik, tetapi proses pengiriman terkadang lebih lama dari yang dijanjikan.	The service is generally good; however, delivery occasionally takes longer than promised.	4	Positive

The sentiment labeling process resulted in two sentiment categories that represent users' perceptions of the application. Table 3 and Figure 3 illustrates the overall sentiment distribution in the dataset.

Table 3. Distribution *of Sentiment Labels*

Sentiment	Number of Reviews	Percentage
Positive	1,808	58.38%
Negative	1,289	41.62%
Total	3,097	100%

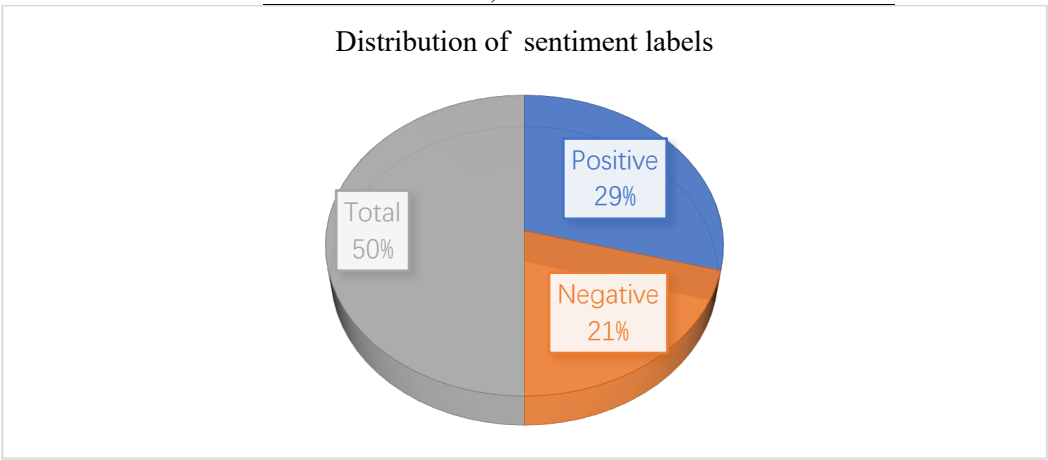


Figure 3. Distribution of Sentiment Labels

Table 3 presents the distribution of sentiment labels obtained after the labeling process. The reviews were categorized into two sentiment classes, namely positive and negative, based on the rating scores provided by users. Reviews with ratings between one and three stars were assigned to the negative class, whereas reviews with ratings between four and five stars were assigned to the positive class.

The results indicate that the dataset is dominated by positive reviews, suggesting that most users expressed favorable opinions toward the Klik Indomaret application. Positive reviews generally highlighted the convenience of online shopping, attractive promotional offers, and ease of accessing product information. In contrast, negative reviews primarily reflected user dissatisfaction related to application errors, stock availability issues, delivery delays, and customer service responsiveness.

The imbalance between positive and negative reviews can be observed in Table 3 and is further illustrated in Figure 3. Such class imbalance is a common characteristic of user-generated review datasets and may negatively affect the performance of machine learning models by causing a bias toward the majority class. Consequently, handling class imbalance becomes an important consideration in the subsequent classification process.

Figure 3 visually depicts the distribution of sentiment labels in the dataset. The figure clearly shows that the positive class contains a larger number of reviews than the negative class, indicating an unequal class distribution. This observation motivates the application of the Synthetic Minority Oversampling Technique (SMOTE) in the training phase to generate additional samples for the minority class and improve the classifier's ability to learn representative sentiment patterns from both classes.

3.3. TF IDF

Figure 4 visualizes the most influential terms based on their TF-IDF weights. Several positive-related terms, such as *promo*, *diskon*, and *cepat*, frequently appeared in favorable reviews, indicating that users appreciated promotional offers and service convenience. Conversely, terms such as *error*, *checkout*, and *stok* were commonly associated with negative reviews, reflecting issues related to application performance and product availability.

```
from sklearn.feature_extraction.text import TfidfVectorizer
X = df_f['stem_string']
tfid = TfidfVectorizer()
tf_X = tfid.fit_transform(X)
```

Figure 4. Top Terms with the Highest TF-IDF Weights

The results demonstrate that TF-IDF successfully identified informative terms that characterize user opinions toward the Klik Indomaret application. These weighted features subsequently served as input for the Random Forest classifier in the sentiment classification stage.

3.4. Class Distribution and SMOTE Results

Following feature extraction, the dataset was divided into training and testing subsets using a 75:25 ratio. The training dataset initially consisted of 966 negative reviews and 1,356 positive reviews. This distribution indicates the presence of class imbalance, where the positive class significantly outnumbered the negative class.

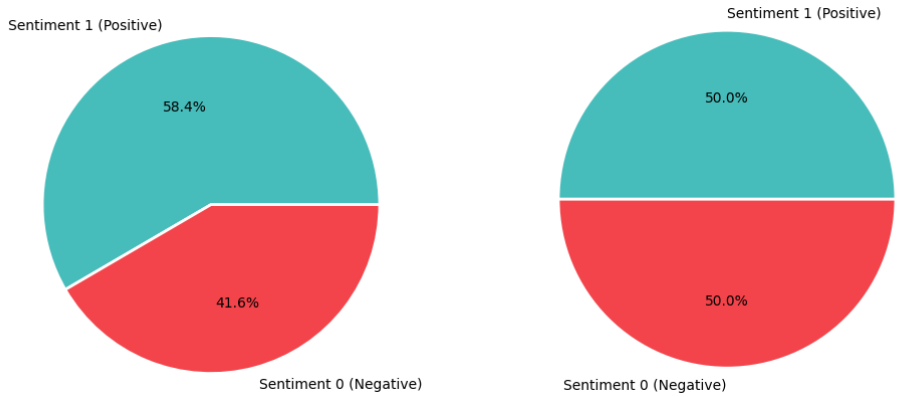


Figure (a) Before SMOTE Figure (b) After SMOTE

Figure 5. Before and After SMOTE for imbalance class dataset

Following feature extraction, the dataset was divided into training and testing subsets using a 75:25 ratio. The training dataset initially consisted of 966 negative reviews and 1,356 positive reviews, indicating an imbalanced class distribution. As illustrated in Figures 5, the positive class substantially outnumbered the negative class prior to oversampling, which may lead to biased classification performance and reduced recognition of minority-class instances. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training dataset. SMOTE generates synthetic samples for the minority class by interpolating between existing minority instances

and their nearest neighbors, thereby increasing the number of minority samples without simply duplicating existing observations. As a result, the number of negative reviews increased from 966 to 1,356, producing a balanced training dataset with an equal number of positive and negative instances (1,356 reviews per class). The comparison between Figures 5 demonstrates the effectiveness of SMOTE in balancing the class distribution. This balanced dataset is expected to improve the learning capability of the Random Forest classifier by providing equal representation of both sentiment classes, enabling the model to construct more robust decision boundaries and reducing prediction bias toward the majority class.

3.5. Random Forest Classification Results

To evaluate the effectiveness of SMOTE, two classification scenarios were investigated Random Forest without SMOTE (RF) and Random Forest with SMOTE (RF-SMOTE). The performance of both models was evaluated using accuracy, precision, recall, and F1-score. The results are summarized in Table 4.

Table 4. Performance Comparison of RF and RF-SMOTE

Model	Accuracy	Precision	Recall	F1-Score
RF	0.84	0.84	0.84	0.84
RF-SMOTE	0.85	0.85	0.85	0.85

Table 4, demonstrate that the RF-SMOTE model achieved superior performance across all evaluation metrics. Compared with the baseline Random Forest model, the implementation of SMOTE improved accuracy from 84% to 85%. Similar improvements were observed for precision, recall, and F1-score.

Although the improvement appears relatively modest, it indicates that balancing the class distribution contributes positively to the learning process. The synthetic samples generated by SMOTE provide additional information about minority-class characteristics, enabling the classifier to better recognize negative sentiment patterns.

To further analyze model performance, confusion matrices were generated for both classification scenarios. Figure 6(a) shows the confusion matrix of the Random Forest model trained on the original imbalanced dataset, while Figure 6(b) presents the confusion matrix of the RF-SMOTE model.

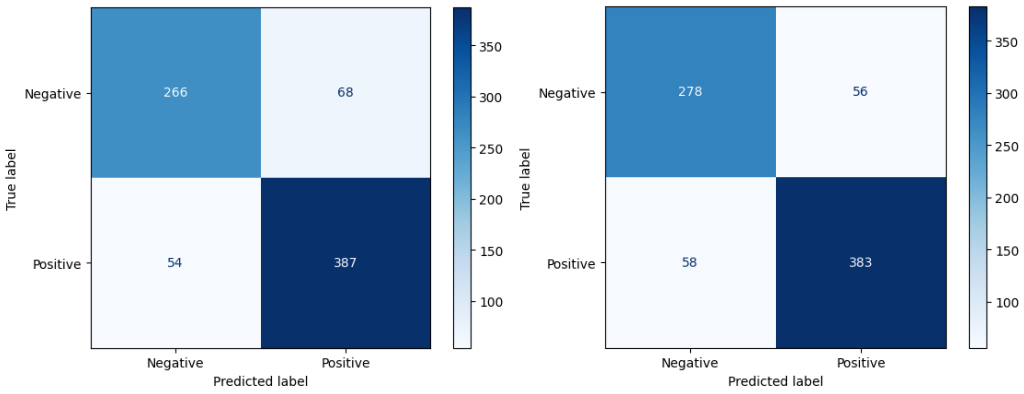


Figure (a) RF Figure (b) RF-SMOTE

Figure 6. Confusion Matrix of (a) RF and (b) RF-SMOTE

Figure 6, showing the confusion matrices reveal that the RF-SMOTE model produced fewer classification errors than the baseline RF model. In particular, the balanced training dataset improved the model's ability to correctly classify minority-class instances. This result confirms that SMOTE effectively mitigated the negative impact of class imbalance and enhanced the overall predictive performance of the Random Forest classifier.

Overall, the experimental findings demonstrate that integrating TF-IDF feature extraction, SMOTE oversampling, and Random Forest classification provides an effective framework for sentiment analysis of Google Play Store reviews. The proposed RF-SMOTE model achieved the best overall performance and exhibited greater robustness in handling imbalanced sentiment data.

3.6. Discussion

The results show that the proposed pipeline is effective for sentiment classification of Klik Indomaret reviews, especially after the class imbalance problem is addressed. The dataset contained 3,097 reviews, and the label distribution in the training set was uneven before SMOTE was applied. This matters because imbalanced data can bias a classifier toward the majority class and reduce its ability to detect minority-class reviews correctly. Recent studies on app-review sentiment analysis have reached the same conclusion: preprocessing, TF-IDF, and Random Forest work well together, but class balancing often determines whether the model can learn minority patterns reliably [1], [2], [4].

In this study, Random Forest without SMOTE produced an accuracy of 0.84, while Random Forest with SMOTE improved to 0.85. The gain is small, but it is meaningful because the balanced model achieved the same improvement across accuracy, precision, recall, and F1-score. That pattern suggests that SMOTE helped the classifier learn a more representative decision boundary rather than simply inflating one metric. This is consistent with earlier Google Play Store studies. For example, the PeduliLindungi study reported 71% accuracy with Random Forest and SMOTE, while the IKD study found that Random Forest remained competitive and that SMOTE changed performance on the balanced training set [1], [8], [9].

The findings also align with broader app-review research showing that Random Forest is a strong baseline for short-text sentiment tasks in Indonesian. The PrimaKu study reported 89% accuracy after hyperparameter tuning, the Shopee study reported 94% accuracy, and the SIGNAL study found Random Forest to be the best-performing classifier among several alternatives when TF-IDF and SMOTE were used [2], [3], [11]. Although these results are not directly comparable because the datasets, labels, and preprocessing choices differ, they point in the same direction: TF-IDF plus Random Forest is a reliable approach for app-review sentiment analysis, and SMOTE is useful when the class distribution is skewed.

The content of the reviews also makes the result practically relevant. Negative reviews in the sample focused on stock availability problems and weak customer service, while positive reviews emphasized convenience, price checking, and promotional benefits. This means the sentiment model is not only classifying text, but also helping surface operational issues that matter to users. That is exactly the kind of insight app developers need from review mining: a compact way to detect service failure points and user-valued features in large, unstructured review streams [7], [12].

The modest improvement after SMOTE also suggests that the dataset is already reasonably learnable with Random Forest, but that balancing still provides a clearer view of the minority class. In that sense, SMOTE does not radically change the task; it stabilizes it. Similar work on imbalanced text classification has shown that oversampling methods are most valuable when the minority class is important and the original distribution is skewed enough to distort the classifier's learning process [5], [6].

4. Conclusions

This study showed that Klik Indomaret app reviews can be classified effectively using TF-IDF, Random Forest, and SMOTE. The best model in this work was Random Forest with SMOTE, which achieved 0.85 on accuracy, precision, recall, and F1-score. Although the improvement over the baseline model was small, it confirms that balancing the data helps the classifier handle the minority class more fairly.

The study also demonstrates that Google Play Store reviews are a useful source for understanding user sentiment and identifying recurring issues in mobile applications. For Klik Indomaret, the reviews point to both service problems and positive user experiences, making sentiment analysis a practical tool for application evaluation. Future research may improve the results by testing alternative classifiers, tuning parameters more systematically, or expanding the dataset with newer reviews.

Author Contributions

Sutriawan: Conceptualization, formal analysis, writing original draft preparation, project administration, and funding acquisition. M. Fitrah Nur: Data curation, investigation, software, experimentation, and visualization. Wahyu Tisno Atmojo: Methodology, validation, writing review and editing, and supervision.

Yo Ceng Giap: Validation, writing review and editing, and supervision. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest

References

[1] Y. Desnelita, D. Oktarina, J. Nasional, T. Elektro, and dan T. Informasi, "Sentiment Analysis of IKD Application Reviews on Play Store Using Random Forest." 2025. doi: 10.22146/jnteti.v14i3.20473.

[2] Q. W. Anadia and A. Meiriza, "Performance Comparison of Sentiment Classification Algorithms on SIGNAL Reviews Using SMOTE." 2025. doi: 10.51519/journalisi.v7i3.1196.

[3] T. Misriati and R. Aryanti, "Optimalisasi Random Forest dan Support Vector Machine dengan Hyperparameter GridSearchCV untuk Analisis Sentimen Ulasan PrimaKu." 2024. doi: 10.47065/josh.v5i4.5347.

[4] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering." 2024. doi: 10.1186/s40537-024-00943-4.

[5] A. Khurana and O. Verma, "Optimal Feature Selection for Imbalanced Text Classification." 2023. doi: 10.1109/TAI.2022.3144651.

[6] N. Santoso, W. Wibowo, and H. Hikmawati, "Integration of synthetic minority oversampling technique for imbalanced class." 2019. doi: 10.11591/IJEECS.V13.I1.PP102-108.

[7] M. R. F. Rahmatullah, P. Andono, Affandy, and M. A. Soeleman, "Improving Random Forest Performance for Sentiment Analysis on Unbalanced Data Using SMOTE and BoW Integration: PLN Mobile Application Case Study." 2025. doi: 10.15294/sji.v12i1.19295.

[8] A. G. Sinaga, R. Robet, and O. Pribadi, "Comprehensive Comparison of TF-IDF and Word2Vec in Product Sentiment Classification Using Machine Learning Models." 2026. doi: 10.30871/jaic.v10i1.11582.

[9] M. R. Pribadi, D. Manongga, H. Purnomo, I. Setyawan, and Hendry, "Sentiment Analysis of the PeduliLindungi on Google Play using the Random Forest Algorithm with SMOTE." 2022. doi: 10.1109/ISITIA56226.2022.9855372.

[10] M. R. Syaputa, M. Arifin, and D. L. Fithri, "Klasifikasi Sentimen Ulasan E-Wallet menggunakan TF-IDF dan Random Forest dengan Penyeimbangan Data SMOTE." 2025. doi: 10.29408/edumatic.v9i2.31084.

[11] M. R. Rahman, A. F. Diansyah, and H. Hanafi, "Sentiment Analysis on the Shopee Application on Playstore Using the Random Forest Classification Method." 2023. doi: 10.25139/inform.v9i1.5465.

[12] S. A. Memon, M. A. Mahar, A. Kehar, S. Oad, and I. Ahmed, "User Experience Enhancement Through Sentiment Analysis: A Machine Learning Approach to App Reviews." 2025. doi: 10.58325/ijisct.004.02.00131.

[13] A. Karim, A. Azhari, S. B. Belhaouri, and A. A. Qureshi, "Machine Learning Algorithm's Measurement and Analytical Visualization of User's Reviews for Google Play Store." 2020. doi: 10.20944/preprints202003.0249.v1.

[14] A. Nugroho, Wiyanto, and D. Maulana, "COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS IN HANDLING IMBALANCED DATA WITH SMOTE OVERSAMPLING APPROACH." 2025. doi: 10.33480/jitk.v11i2.6956.

[15] M. D. Rizkiyanto, M. D. Purbolaksono, and W. Astuti, "Sentiment Analysis Classification on PLN Mobile Application Reviews using Random Forest Method and TF-IDF Feature Extraction." 2024. doi: 10.31963/intek.v11i1.4774.

[16] R. Asyrofi and R. Fauzan, "Synthetic-MixUp: A Simple Framework for Imbalanced Text classification." 2023. doi: 10.1109/GCCE59613.2023.10315313.

[17] U. Kavirayani, K. Mylavarapu, J. Pilli, P. Jannu, J. A. Mohammad, and S. Panchali, "Comparative Performance of Machine Learning Techniques for Sentiment and Insight Analysis in Amazon Product Reviews." 2022. doi: 10.63282/3117-5481/aijcst-v4i1p106.

[18] S. Pandey, D. Chhabara, A.Vasuki, and S. Indhumathi, "Machine Learning Based Google Play Store Reviews Classification." 2021. [Online]. Available: https://www.semanticscholar.org/paper/071c7dbb508e36b679dfb6fcb07695da4d4b9e1c

[19] L. S. R and A. P, "Google Playstore Reviews Prediction Using ML And NLP." [Online]. Available: https://www.semanticscholar.org/paper/8fab6c06f5653d8502f8ba8f0be70964891687ed

[20] Munaldi, "Penerapan Algoritma Random Forest Berbasis TF – IDF untuk Meningkatkan untuk Akurasi Deteksi Sentimen Ulasan Produk E-Commerce." 2025. doi: 10.30606/rjti.v4i3.4045.

[21] K. P. Harmandini and K. M. L, “Analysis of TF-IDF and TF-RF Feature Extraction on Product Review Sentiment.” 2024. doi: 10.33395/sinkron.v8i2.13376.
[22] H. Barus, I. N. Fajri, and Y. Pristyanto, “Sentiment Classification Analysis of Tokopedia Reviews Using TF-IDF, SMOTE, and Traditional Machine Learning Models.” 2025. doi: 10.30871/jaic.v9i5.10524.
[23] D. S. Assyifa and A. Luthfiarta, “SMOTE-Tomek Re-sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-class Classification.” 2024. doi: 10.25139/inform.v9i2.8410.
[24] N. Curukoglu and A. Ozpinar, “SMOTE-Text: A Modified SMOTE for Turkish Text Classification.” 2020. doi: 10.1007/978-3-030-79357-9_9.