Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

Evaluating SMOTE Performance for Imbalanced Multi-Label Sentiment Classification in MLSE Usability Testing of Mobile App Reviews

Hasan Basri*[1], Wahyu Noviani Purwanti[2], Ihsan Alparisi[3]

[1,2,3]Information Systems, Faculty of Science and Technology, Universitas Terbuka, Indonesia

Email: [1]hasan.basri@ecampus.ut.ac.id

Received : Sep 29, 2025; Revised : Nov 26, 2025; Accepted : Dec 4, 2025; Published : Apr 15, 2026

Abstract

Imbalanced data poses a significant challenge in multi-label classification tasks, especially when combining sentiment analysis with usability testing of mobile application reviews. This study investigates the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) in improving classification performance on a multi-label dataset consisting of 10,000 Indonesian language user reviews from the Google Play store. The classification labels represent a combination of usability criteria and sentiment polarity, with strong imbalance observed across several classes. Three machine learning algorithms SVM, Decision Tree, and Random Forest were evaluated on datasets of increasing sizes (1,000 to 10,000 entries), each tested under both original and SMOTE-balanced conditions using stratified 10-fold cross-validation with accuracy and F1-score as the primary metrics. Experimental results show that SMOTE significantly improves the performance of Decision Tree mainly on smaller datasets but exhibits inconsistent gains as the dataset grows, provides modest and stable improvements for Random Forest, and negatively impacts SVM, whose performance remains consistently better without SMOTE. This study concludes that SMOTE is not a universally effective solution and must be applied selectively based on model characteristics. These findings contribute to the Machine Learning for Software Engineering (ML4SE) domain and the field of informatics by highlighting the importance of aligning resampling techniques with algorithmic behaviour when dealing with highly imbalanced multi-label text classification tasks.

Keywords : *Imbalanced Data, Multi-label Classification, Sentiment Analysis, SMOTE, Usability Testing.*

1. INTRODUCTION

Mobile applications have become an integral part of everyday life in modern society. With the number of apps on Google Play Store surpassing 2.4 million by 2023, developers are under increasing pressure not only to deliver innovative features but also to ensure high usability quality in order to remain competitive and retain users [1]–[3]. One emerging approach to assessing app usability is sentiment analysis, which allows user reviews to be automatically leveraged as a valuable source of real-world evaluative data [4][5]. In previous studies, the authors conducted two stages of research. The first stage evaluated the performance of several classification algorithms namely Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Random Forest in classifying user reviews based on a combination of usability testing dimensions and sentiment polarity. This study found that SVM offered the most consistent performance [5]. The second stage focused on optimizing SVM through kernel selection and tuning of the parameter C, where the linear kernel with C = 0.01 yielded the best results, achieving an accuracy of 75.20% and an F1-score of 0.8775 [6].

However, one critical challenge that remained unaddressed in these studies is the class imbalance among labels particularly within the multi-label setup that combines usability dimensions and sentiment. This imbalance is not only prevalent but also varies drastically, ranging from mild to extremely skewed cases [7][8]. For example, the label Efficiency_Positive has 552 samples, while its counterpart Efficiency_Negative only has 30. Similarly, UserSatisfaction_Positive dominates with 813 instances

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

compared to just 61 for UserSatisfaction_Negative. An extreme disparity is seen in the ErrorRate_Positive label, which appears only once, while ErrorRate_Negative occurs 373 times. Such distributions reveal that in some cases, the imbalance ratio exceeds 1:500, creating a highly suboptimal environment for training classification models. Under these conditions, classifiers tend to ignore minority classes due to their negligible contribution to the loss function during training [9]–[11]. Consequently, the model becomes biased toward majority classes, which may lead to reduced overall accuracy and poor generalization, particularly when applied to real-world data with more diverse distributions [12][13].

Several related works have explored text classification on app reviews, but their settings differ markedly from this study. Prior research on governmental service apps applied SMOTE only in a single-label Arabic dataset with moderate imbalance (about a 4:1 ratio), while recent other evaluations of oversampling techniques also focus solely on single-label text classification [14][15]. Other studies such as MNoR-BERT address multi-label classification for NFR extraction, yet they do not consider class imbalance or oversampling. Consequently, no existing work has examined how SMOTE behaves in multi-label usability sentiment classification, especially under imbalance that ranges from moderate to highly extreme [16].

Class imbalance is a recurring problem in text classification, especially when minority classes contain only a small number of instances compared to dominant classes [17][18]. This imbalance often causes models to favor majority labels and overlook minority ones, which becomes even more problematic in multi-label settings where each label combination may have a different imbalance ratio [19]–[21]. A common strategy to mitigate this issue is the Synthetic Minority Over-sampling Technique (SMOTE). Instead of duplicating minority samples, SMOTE creates new synthetic instances by interpolating between a sample and its nearest neighbors, allowing the minority class to become more representative and diverse [22]–[26].

To address this issue, the present study focuses on applying the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate data imbalance and enhance classification performance. This research investigates the impact of SMOTE on the performance of three classifiers SVM, Decision Tree, and Random Forest both with and without the application of SMOTE. The experiments were conducted using user review data from a popular application on Google Play Store, with datasets ranging from 1,000 to 10,000 samples, evaluated using 10-fold stratified cross-validation. Through this experimental setup, the study aims to test the hypothesis that SMOTE can significantly improve classification performance on imbalanced multi-label datasets. The findings are expected to contribute valuable insights into the effectiveness of SMOTE in the context of multi-label text classification within the domain of Machine Learning for Software Engineering (ML4SE) particularly for automated usability evaluation based on user reviews.

2. METHOD

2.1. Research Design

This study adopts a quantitative experimental approach to examine the impact of the SMOTE oversampling technique on the performance of classification models in sentiment analysis combined with usability testing parameters. The complete sequence of research stages is illustrated in figure 1.

2.2. Dataset and Labeling

The dataset used in this study consists of 10,000 user reviews collected from a popular mobile application on the Google Play Store. Each review was annotated using a multi-label classification approach that combines usability testing criteria with sentiment polarity [5]. The annotation process was carried out in two stages.

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

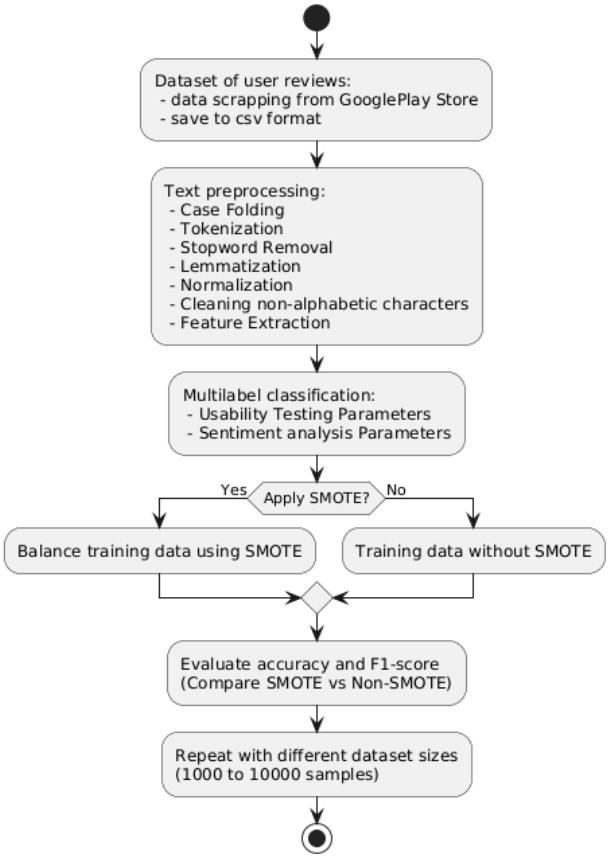

Figure 1. Research Stages

First, Annotator A labeled each review with sentiment polarity (positive or negative). After this step was completed, Annotator B assigned usability testing labels (such as User Satisfaction, Learnability, Error Rate, Effectiveness, and others) and reviewed the sentiment labels assigned by Annotator A to ensure consistency. Finally, Annotator A performed a second pass to validate the usability labels provided by Annotator B. This multi-stage annotation ensured that each review could contain more than one label, reflecting the multi-label nature of the dataset.

To handle the multi-label classification technically, this study adopted a binary relevance transformation, in which each usability-sentiment combination was treated as an independent binary classification subproblem. This approach is widely used for multi-label text classification because it allows traditional classifiers such as SVM, Decision Tree, and Random Forest to be applied directly to each label without requiring algorithmic modification [27][28] .

The distribution of data based on the combination of criteria and sentiment is presented in figure 2. figure 2 illustrates a hierarchical imbalance in data distribution, ranging from moderate to highly extreme disparities. For instance, the label UserSatisfaction_ Positive dominates with 4,446 samples, while Learnability_Negative contains only 24 samples. The most striking imbalance is observed in the ErrorRate_Positive label, which appears only 15 times out of 10,000 samples. This imbalance becomes even more problematic in the earlier stages of testing, such as at the 1,000 to 5,000 sample scale where ErrorRate_Positive drops to fewer than five instances. As a result, this label was excluded from the training process to prevent extreme imbalance that could negatively affect model accuracy and stability.

To investigate the effect of the SMOTE technique in addressing this imbalance, a stepwise experiment was conducted, starting from 1,000 up to 10,000 reviews, increasing in increments of 1,000 samples. At each step, classification was performed using SVM, Decision Tree, and Random Forest, both with and without SMOTE, and evaluated using stratified 10-fold cross-validation.

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

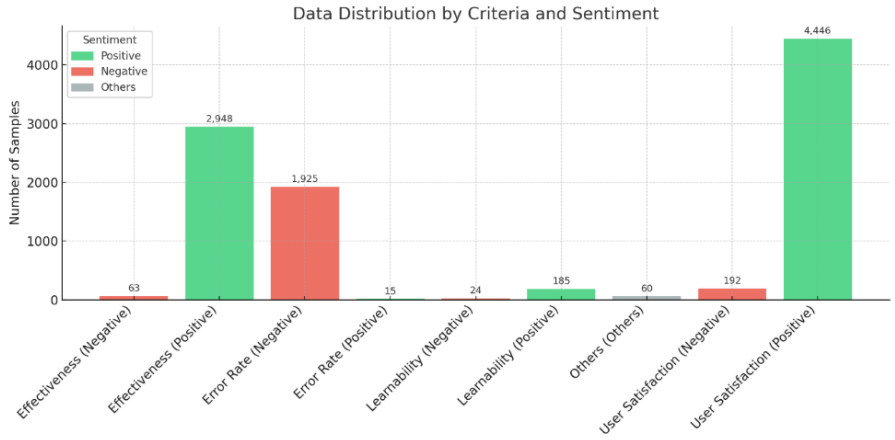

Figure 2. Data Distribution by Usability Criteria and Sentiment

2.3. Text Preprocessing

The collected user reviews were subjected to a text preprocessing pipeline to ensure that the data fed into the classification models is clean, consistent, and analytically meaningful [29] . Since all reviews were written in Indonesian, the preprocessing steps were specifically tailored to match the linguistic characteristics of the language.

The process began with case folding, followed by tokenization, which splits sentences into individual word tokens. This was then followed by stopword removal, eliminating common Indonesian words such as “dan” (and), “yang” (which), or “itu” (that), which typically carry little semantic weight in the context of sentiment and usability analysis [30]–[32]. Next, lemmatization was applied to reduce each word to its base or root form for example, “membeli” (buying) becomes “beli” (buy). This was followed by normalization of numeric and non-alphabetic characters to simplify the text format. As part of this normalization, numerical values were also converted into words (e.g., the number “1” is transformed into “satu”), allowing them to be processed linguistically on par with regular words [6]. The final step involved feature extraction using the TF-IDF (Term Frequency–Inverse Document Frequency) method. This technique assigns weights to words based on their importance within a document relative to the entire collection of reviews. The outcome of this entire preprocessing pipeline is a TF-IDF matrix a numerical representation of each user review in Indonesian that serves as the input for the classification algorithms [33][34].

2.4. Handling Imbalanced Data with SMOTE

Imbalanced data is a common issue in classification tasks, especially when one or more classes have significantly fewer samples compared to others [17][18]. This imbalance can lead to classification models becoming overly biased toward the majority class while ignoring the minority class, ultimately degrading the overall performance particularly in multi-label classification settings [19]–[21]. To address this issue, this study employs the Synthetic Minority Over-sampling Technique (SMOTE). Unlike traditional oversampling methods that replicate existing data, SMOTE generates synthetic samples based on the existing minority class instances, thereby creating more diverse and representative training data [22]–[24].

SMOTE works by randomly selecting a sample from the minority class, identifying its nearest neighbors (commonly using the k-nearest neighbors algorithm), and then generating a new sample by interpolating between the original point and one of its neighbors [25][26]. In this study, the feature vectors x_i and x_{zi} represent TF-IDF embeddings in $\mathbb{R}^d$, and the number of nearest neighbors k was set to 5, which is the standard configuration used in most SMOTE implementations. For example, if x_i is

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

the feature vector of a minority sample and x_{zi} is its nearest neighbor, the synthetic sample x_{new} is generated using the formula:

$$x_{new} = x_i + \delta \times (x_{zi} - x_i) \quad (1)$$

where:

x_i = original minority data point (vector in $\mathbb{R}^d$)
x_{zi} = nearest neighbor of x_i (vector in $\mathbb{R}^d$)
$\delta \in [0, 1]$ = a random value drawn from a uniform distribution

Through this approach, SMOTE produces new samples that lie between existing data points, thereby introducing natural variation and enriching the representation of the minority class without simply duplicating data [35][36].

2.5. Classification and Evaluation

In the classification stage, this study employed three machine learning algorithms SVM, Decision Tree, and Random Forest. The SVM was implemented with a linear kernel, following previous research findings that demonstrated its consistent and stable performance in multi-label text classification tasks [5]. The Decision Tree was selected as a baseline approach due to its interpretability, while the Random Forest served as an extension that aggregates multiple decision trees to improve accuracy and reduce the risk of overfitting [37][38].

Each model was tested under two primary conditions without SMOTE (the model was trained using the original, imbalanced dataset) and with SMOTE (the dataset was first balanced synthetically using the Synthetic Minority Over-sampling Technique (SMOTE) before being used for model training). Model performance was evaluated using stratified 10-fold cross-validation, in which the dataset is divided into ten equal parts while preserving the label distribution in each fold. This method ensures that the training and testing process remains consistent and unbiased across class distributions [39].

As the primary evaluation metric, this study use accuracy, precision, recall and F1-Score, including macro average variants, to provide a more comprehensive assessment of performance on imbalanced multi-label data. These metrics allow the evaluation to capture both the model's ability to correctly identify minority classes and its overall classification correctness. The evaluation focused on comparing the classification results between models trained on the original data and those trained on SMOTE-balanced data, in order to observe whether the use of synthetic data had a significant impact on improving classification performance.

2.6. Data Size Variation

To examine the effect of data scale on model performance and the effectiveness of SMOTE, experiments were conducted incrementally using the following dataset sizes: 1,000; 2,000; 3,000; 4,000; 5,000; 6,000; 7,000; 8,000; 9,000; and 10,000 user reviews. Each scenario was evaluated independently to provide a comprehensive understanding of how model performance trends vary across different dataset sizes.

3. RESULT

This study aims to evaluate the effectiveness of the SMOTE technique in addressing imbalanced data problems in multi-label classification involving sentiment and usability testing. Three machine learning algorithms SVM, Decision Tree, and Random Forest were tested under both SMOTE and non-SMOTE conditions, across varying dataset sizes ranging from 1,000 to 10,000 samples. Evaluation was performed using stratified 10-fold cross-validation, with accuracy and macro-averaged metrics as the primary performance metric.

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

3.1. Overall Model Performance Across Dataset Sizes

Figure 3 shows that the impact of SMOTE varies across models. For SVM, the non-SMOTE condition consistently achieved higher accuracy at all dataset sizes. Decision Tree gained small improvements on smaller datasets, but SMOTE became less effective as the dataset grew. Random Forest showed stable results in both settings, with only marginal differences between SMOTE and non-SMOTE.

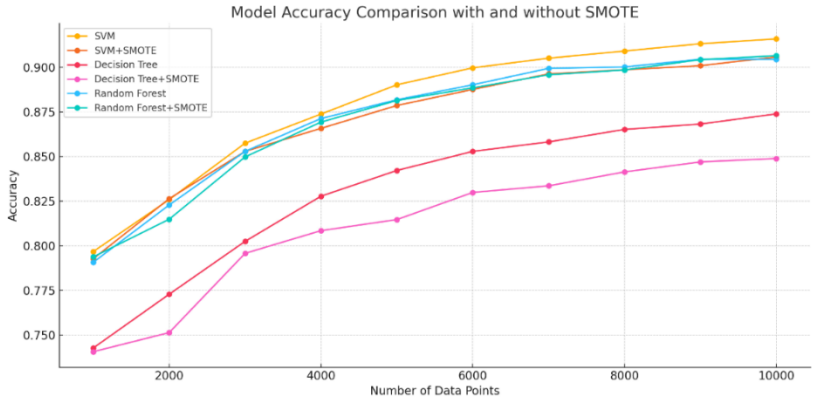

Figure 3. Model Accuracy Comparison With and Without SMOTE

3.2. Performance Trends Across Dataset Sizes and Algorithms

The results in figure 4 to 6 summarize the performance of SVM, Decision Tree, and Random Forest on datasets ranging from 1,000 to 10,000 samples, evaluated using accuracy and macro-averaged precision, recall, and F1-score. A consistent pattern emerges across all dataset sizes: the impact of SMOTE depends strongly on the underlying classification model.

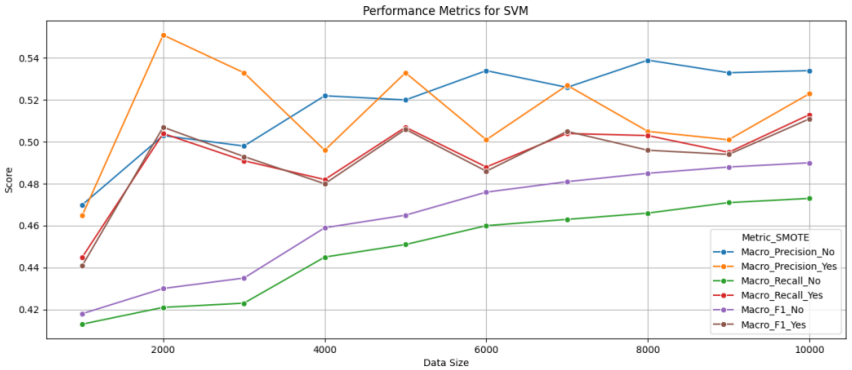

Figure 4. SVM Performance Summary

Figure 4 shows the performance trends for SVM across dataset sizes from 1,000 to 10,000 samples show a clear and consistent pattern. Macro precision for both SMOTE and non-SMOTE settings increases gradually as the dataset size grows, but the non-SMOTE version maintains slightly higher stability, especially from 4,000 samples onward. Macro recall shows that SMOTE offers a mild benefit at smaller dataset sizes, with higher recall at 1,000 to 3,000 samples. However, as the dataset becomes larger, the gap narrows and both conditions converge with only minor differences. Macro F1-score follows a similar pattern, the SMOTE version improves performance on very small datasets but gradually loses its advantage once the dataset exceeds 4,000 samples. Across nearly all metrics, the SMOTE curves fluctuate more sharply, while the non-SMOTE curves show smoother and more stable growth.

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

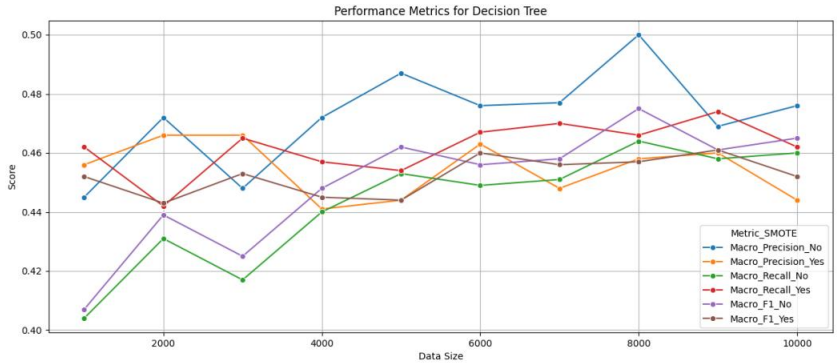

Figure 5. Decision Tree Performance Summary

In figure 5, the Decision Tree model shows a different behavioral pattern compared to SVM. Based on the performance curves, macro precision, recall, and F1-score fluctuate across dataset sizes for both SMOTE and non-SMOTE conditions, indicating that Decision Tree is more sensitive to variations in data distribution. SMOTE provides small gains at certain points, particularly in recall and F1-score for mid-sized datasets between 4,000 and 8,000 samples. However, these improvements are inconsistent and do not persist as the dataset grows. In several cases, the non-SMOTE model achieves higher macro precision, especially at 6,000 and 8,000 samples, suggesting that synthetic oversampling does not consistently enhance the model's ability to generalize.

Figure 6. Random Forest Performance Summary

Figure 6 illustrates that Random Forest demonstrates the most stable performance across dataset sizes, with both precision and F1-score increasing gradually as the number of samples grows. Compared to SVM and Decision Tree, the effect of SMOTE on Random Forest is modest but consistently positive. The SMOTE-enhanced model shows slightly higher macro precision and macro F1-scores, particularly at 3,000 to 8,000 samples, indicating improved recognition of minority labels without degrading performance on majority classes. Macro recall follows a similar upward trend in both conditions, with SMOTE offering small but steady improvements.

4. DISCUSSION

4.1. SMOTE Performance Analysis

Based on the model accuracy comparison chart (Figure 3) covering datasets from 1,000 to 10,000 samples, it is evident that the effectiveness of SMOTE varies across different classification algorithms. The results in this study show that the effect of SMOTE on Decision Tree is actually inconsistent across dataset sizes. While SMOTE helps improve accuracy on smaller datasets (1,000 to 3,000 samples), its

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

benefit does not persist and even fluctuates as the dataset grows. This indicates that Decision Trees are indeed sensitive to data imbalance, but they do not always gain stable improvement from SMOTE, especially at larger data scales where oversampling can introduce noise that affects the split criteria.

In contrast, Random Forest more stability. The accuracy curve of Random Forest with SMOTE remains slightly higher than its non-SMOTE counterpart across most dataset sizes, particularly from 3,000 samples and above. This suggests that Random Forest's ensemble nature reduces the negative impact of synthetic noise and allows SMOTE to provide modest yet consistently positive contributions. Although the improvement is not large, the pattern is stable and predictable, unlike the fluctuations observed in Decision Tree.

For the Support Vector Machine (SVM) model, SMOTE did not improve performance at any dataset size. The non-SMOTE condition consistently outperformed the SMOTE-enhanced version, especially in the 8,000 to 10,000 sample range. This indicates that SVM is already structurally robust against imbalanced data due to its margin maximization mechanism. Introducing synthetic samples through SMOTE likely distorts the minority class distribution and disrupts the optimal hyperplane, resulting in reduced accuracy.

These findings reaffirm a key insight, SMOTE is not universally beneficial, and its effectiveness depends heavily on the characteristics of the model. Decision Tree benefits from SMOTE only in smaller datasets and shows unstable improvements at larger scales, Random Forest receives a small but steady enhancement, and SVM experiences performance degradation due to synthetic data interfering with its decision boundary. These results highlight the importance of aligning imbalance-handling strategies with model behaviour. Oversampling techniques such as SMOTE should be used selectively, with consideration of both dataset size and algorithmic sensitivity.

5. CONCLUSION

This study evaluated the impact of the SMOTE technique in addressing data imbalance issues within multi-label classification that combines sentiment analysis and usability testing of mobile app user reviews. By testing three classification algorithms SVM, Decision Tree, and Random Forest across various dataset sizes and two experimental conditions (with and without SMOTE), the findings reveal that the effectiveness of SMOTE is strongly influenced by the characteristics of each model rather than by the imbalance ratio alone. The experimental results show that Decision Tree benefits from SMOTE primarily on smaller datasets, but the improvements are inconsistent and tend to fluctuate as the dataset size increases. This reflects the model's high sensitivity to distributional changes introduced by synthetic oversampling. Random Forest demonstrates more stable performance, with SMOTE providing modest yet consistently positive improvements, supported by the ensemble's robustness against noise. For SVM, SMOTE consistently reduces performance, as the introduction of synthetic samples interferes with the optimal separating margin, especially at larger data scales. These findings confirm that SMOTE is not a universal solution for imbalanced multi-label text classification. Its effectiveness depends greatly on the inherent learning mechanism of each algorithm. Accordingly, the application of SMOTE or any oversampling method should be carefully aligned with the structural properties of the chosen model. This study contributes meaningful insights to the field of Machine Learning for Software Engineering (ML4SE), particularly in developing strategies for handling imbalanced multi-label datasets derived from user reviews.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Terbuka for the support and resources provided throughout the course of this research. This work would not have been possible without the encouragement and facilities made available by the institution.

REFERENCES

[1] A. R. Akbar, M. G. H. Fikri, N. D. Putra, Sunardi, and A. Hairuman, "Comparing the User Experience of Mobile Banking Applications Using System Usability Scale and Usability Testing," in *2024 9th International Conference on Business and Industrial Research (ICBIR)*, IEEE, May 2024, pp. 0106–0111. https://doi.org/10.1109/ICBIR61386.2024.10875913.

[2] S. Gottschalk, F. Rittmeier, and G. Engels, "Intertwined Development of Business Model and Product Functions for Mobile Applications: A Twin Peak Feature Modeling Approach," 2019, pp. 192–207. https://doi.org/10.1007/978-3-030-33742-1_16.

[3] M. K. Uddin, H. Qiang, H. Jun, and C. Caslon, "Feature Recommendation by Mining Updates and User Feedback from Competitor Apps," in *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, New York, NY, USA: ACM, Dec. 2020, pp. 18–28. https://doi.org/10.1145/3448891.3448953.

[4] G. T. Roy and D. Biswas, "Exploring Transformer and Recurrent Neural Models for Sentiment Analysis of Mobile App Reviews," in *2024 2nd International Conference on Information and Communication Technology (ICICT)*, IEEE, Oct. 2024, pp. 209–213. https://doi.org/10.1109/ICICT64387.2024.10839678.

[5] H. Basri, M. B. S. Junianto, and I. Kusyadi, "Enhancing Usability Testing Through Sentiment Analysis: A Comparative Study Using SVM, Naive Bayes, Decision Trees and Random Forest," *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 4, pp. 1603–1610, Oct. 2024, https://doi.org/10.32493/jtsi.v7i4.45117.

[6] H. Basri, "OPTIMIZING SENTIMENT ANALYSIS FOR USABILITY TESTING: ENHANCING SVM ACCURACY THROUGH KERNEL SELECTION AND TUNING METHODS," *MULTITEK Indones.*, vol. 18, no. 2, pp. 105–113, Jan. 2025, https://doi.org/10.24269/mtkind.v18i2.10615.

[7] Y. Huang, B. Giledereli, A. Köksal, A. Özgür, and E. Ozkirimli, "Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 8153–8161. https://doi.org/10.18653/v1/2021.emnlp-main.643.

[8] M. Bhattacharjee, K. Ghosh, A. Banerjee, and S. Chatterjee, "Multilabel Sentiment Prediction by Addressing Imbalanced Class Problem Using Oversampling," 2021, pp. 239–249. https://doi.org/10.1007/978-981-15-9433-5_23.

[9] K. R. M. Fernando and C. P. Tsokos, "Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 7, pp. 2940–2951, Jul. 2022, https://doi.org/10.1109/TNNLS.2020.3047335.

[10] Y. Xu, H. Ye, N. Zhang, and G. Du, "Leveraging Autoencoder and Focal Loss for Imbalanced Data Classification," in *2022 12th International Conference on Information Technology in Medicine and Education (ITME)v*, IEEE, Nov. 2022, pp. 502–506. https://doi.org/10.1109/ITME56794.2022.00110.

[11] M. Abdelhamid and A. Desai, "Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification," Sep. 2024, https://doi.org/10.48550/arXiv.2409.19751.

[12] K. S. Raslan, A. S. Alsharkawy, and K. R. Raslan, "iHHO-SMOTe: A Cleansed Approach for Handling Outliers and Reducing Noise to Improve Imbalanced Data Classification," Apr. 2025, https://doi.org/10.48550/arXiv.2504.12850.

[13] S. A. Alex and J. J. V. Nayahi, "Classification of Imbalanced Data Using SMOTE and AutoEncoder Based Deep Convolutional Neural Network," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 31, no. 03, pp. 437–469, Jun. 2023,

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

https://doi.org/10.1142/S0218488523500228.

[14] M. Hadwan, M. Al-Sarem, F. Saeed, and M. A. Al-Hagery, "An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique," *Appl. Sci.*, vol. 12, no. 11, p. 5547, May 2022, https://doi.org/10.3390/app12115547.

[15] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, no. 1, p. 21631, Jul. 2025, https://doi.org/10.1038/s41598-025-05791-7.

[16] K. Kaur and P. Kaur, "MNoR-BERT: multi-label classification of non-functional requirements using BERT," *Neural Comput. Appl.*, vol. 35, no. 30, pp. 22487–22509, Oct. 2023, https://doi.org/10.1007/s00521-023-08833-1.

[17] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study1," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Nov. 2002, https://doi.org/10.3233/IDA-2002-6504.

[18] S. S. Rawat and A. K. Mishra, "Review of Methods for Handling Class-Imbalanced in Classification Problems," Nov. 2022, https://doi.org/10.48550/arXiv.2211.05456.

[19] J. Chen and S. Li, "Class-aware Learning for Imbalanced Multi-Label Classification," in *2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, IEEE, Oct. 2023, pp. 903–907. https://doi.org/10.1109/ICCASIT58768.2023.10351721.

[20] R. Rastogi and S. Mortaza, "Imbalance multi-label data learning with label specific features," *Neurocomputing*, vol. 513, pp. 395–408, Nov. 2022, https://doi.org/10.1016/j.neucom.2022.09.085.

[21] G. Du *et al.*, "Graph-Based Class-Imbalance Learning With Label Enhancement," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 9, pp. 6081–6095, Sep. 2023, https://doi.org/10.1109/TNNLS.2021.3133262.

[22] X. Li and Q. Liu, "DDSC-SMOTE: an imbalanced data oversampling algorithm based on data distribution and spectral clustering," *J. Supercomput.*, vol. 80, no. 12, pp. 17760–17789, Aug. 2024, https://doi.org/10.1007/s11227-024-06132-7.

[23] W.-C. Cheng, T.-H. Mai, and H.-T. Lin, "From SMOTE to Mixup for Deep Imbalanced Classification," Nov. 2023, https://doi.org/10.48550/arXiv.2308.15457

[24] A. Li, T. Ma, S. Ye, and X. Liu, "SMOTE-IF: A Novel Resampling Method Based on SMOTE Using Isolation Forest Variants for Multi-Class Imbalanced Data," in *2023 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cyberma*, IEEE, Dec. 2023, pp. 570–577. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics60724.2023.00107.

[25] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, "SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling," *Appl. Intell.*, vol. 51, no. 3, pp. 1394–1409, Mar. 2021, https://doi.org/10.1007/s10489-020-01852-8.

[26] C. Srinilta and S. Kanharattanachai, "Application of Natural Neighbor-based Algorithm on Oversampling SMOTE Algorithms," in *2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*, IEEE, Apr. 2021, pp. 217–220. https://doi.org/10.1109/ICEAST52143.2021.9426310.

[27] D. Ruiz Alonso, C. Zepeda Cortés, H. Castillo Zacatelco, J. L. Carballido Carranza, and J. L. García Cué, "Multi-label classification of feedbacks," *J. Intell. Fuzzy Syst.*, vol. 42, no. 5, pp. 4337–4343, Mar. 2022, https://doi.org/10.3233/JIFS-219224.

[28] J. A. Ferreira Costa, E. D. de S. A. Silva, S. de O. Silva, and N. C. D. Dantas, "Multi-Label Classification of Legal Cases According to the Sustainable Development Goals Using Machine Learning Algorithms," in *2024 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, IEEE, Nov. 2024, pp. 1–6. https://doi.org/10.1109/LA-CCI62337.2024.10814740.

[29] S. E. Latha V, Chandre S, "Sentiment Analysis for User Reviews Based on Improved Binarization Aquila Optimization with Self-Attention Bi-LSTM Model," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, pp. 813–824, Oct. 2024, https://doi.org/10.22266/ijies2024.1031.61.

Vol. 7, No. 2, April 2026, Page. 937-947
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5351

[30] H. Candra, E. D. Madyatmadja, J. Nathaniel, and M. R. Jonathan, "Sentiment Analysis on Indonesian Telegram Reviews Using Naïve Bayes, SVM, Random Forest, and Boosting Models," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, Aug. 2024, pp. 493–498. https://doi.org/10.1109/ICITISEE63424.2024.10730718.

[31] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review," *RSF Conf. Ser. Eng. Technol.*, vol. 1, no. 1, pp. 363–367, Dec. 2021, https://doi.org/10.31098/cset.v1i1.397.

[32] S. Khomsah and Agus Sasmito Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 648–654, Aug. 2020, doi: 10.29207/resti.v4i4.2035.

[33] V. D. Antonio, S. Efendi, and H. Mawengkang, "Sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent," *Int. J. Nonlinear Anal. Appl.*, vol. 13, no. 1, pp. 1367–1373, 2022, https://doi.org/10.22075/ijnaa.2021.5735.

[34] A. S. Safitri, I. Wijayanto, and S. Hadiyoso, "Improving Classification Accuracy With Preprocessing Techniques For Sentiment Analysis," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Jul. 2024, pp. 487–490. https://doi.org/10.1109/ICoDSA62899.2024.10651657.

[35] F. Kamalov, S. E. Choutri, and A. F. Atiya, "Analytical formulation of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Gulf J. Math.*, vol. 19, no. 1, pp. 400–415, Jan. 2025, https://doi.org/10.56947/gjom.v19i1.2639.

[36] O. Kachan, A. Savchenko, and G. Gusev, "Simplicial SMOTE: Oversampling Solution to the Imbalanced Learning Problem," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, New York, NY, USA: ACM, Jul. 2025, pp. 625–635. https://doi.org/10.1145/3690624.3709268.

[37] B. Talekar, "A Detailed Review on Decision Tree and Random Forest," *Biosci. Biotechnol. Res. Commun.*, vol. 13, no. 14, pp. 245–248, Dec. 2020, https://doi.org/10.21786/bbrc/13.14/57.

[38] R. G. McClarren, "Decision Trees and Random Forests for Regression and Classification," in *Machine Learning for Engineers*, Cham: Springer International Publishing, 2021, pp. 55–82. https://doi.org/10.1007/978-3-030-70388-2_3.

[39] V. Lumumba, D. Kiprotich, M. Mpaine, N. Makena, and M. Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," *Am. J. Theor. Appl. Stat.*, vol. 13, no. 5, pp. 127–137, Oct. 2024, https://doi.org/10.11648/j.ajtas.20241305.13.