BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No.4 May 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i4.20

Comparative Analysis of Machine Learning and Deep Learning Models for Sentiment Analysis of Mobile Game Reviews

Iyus Rusmansyah[1]

Information System[1]
Citra Buana Indonesia Institute, Sukabumi, Indonesia[1]
http://www.cbi.ac.id[1]
afecto.rusmansyah24@gmail.com[1]

Abstract. The rapid growth of mobile gaming has led to a substantial increase in user-generated reviews, providing valuable insights into user experience. However, many existing sentiment analysis studies on mobile game reviews rely on balanced datasets and pay limited attention to the challenges of multi-class classification under imbalanced conditions, particularly for minority classes such as neutral sentiment. In addition, limited studies systematically examine how class imbalance affects the comparative performance of Machine Learning and Deep Learning models within a unified experimental setting. This study evaluates the performance of Machine Learning and Deep Learning approaches for sentiment analysis using imbalanced mobile game review data. A dataset of 5000 reviews collected from the Google Play Store is categorized into three classes: positive, neutral, and negative. Light Gradient Boosting Machine (LightGBM) and Convolutional Neural Network (CNN) are used as representative models, with class weighting applied to address data imbalance. The findings show that CNN achieves slightly higher accuracy (68.20%) than LightGBM (66.40%), although both models show comparable performance in macro-average metrics. Both approaches experience difficulty in identifying the neutral class, reflecting the impact of class imbalance. These findings emphasize that class distribution plays a more critical role than model choice in real-world sentiment classification.

Key words: Sentiment Analysis; Mobile Game Reviews; Machine Learning; Deep Learning; Imbalanced Dataset; NLP

1. INTRODUCTION

In recent years, the rapid growth of mobile gaming has significantly increased the number of user-generated reviews on digital distribution platforms such as Google Play Store. These reviews contain valuable information regarding user satisfaction, gameplay experience, and overall application performance. However, due to the large volume of data, manual analysis of user feedback becomes inefficient and impractical. Therefore, automated sentiment analysis has become an important approach to extract meaningful insights from user reviews.

Sentiment analysis is a Natural Language Processing (NLP) technique used to identify and classify opinions expressed in textual data into predefined categories, such as positive, negative, and neutral [1]. This technique has been widely applied in various domains, including product reviews, social media analysis, and customer feedback evaluation. In the context of mobile games, sentiment analysis can help developers understand user preferences and improve game quality based on user feedback.

Various approaches have been proposed for sentiment analysis, particularly using Machine Learning (ML) and Deep Learning (DL) methods. Traditional ML models, such as Light Gradient Boosting Machine (LightGBM), are known for their efficiency and strong performance when combined with feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF). On the other hand, DL models, such as Convolutional Neural Networks (CNN), are capable of automatically learning feature representations from text data, making them suitable for complex language patterns [2].

Despite the advantages of both approaches, there is still a need to evaluate and compare their performance in specific contexts, such as mobile game reviews, especially when dealing with imbalanced datasets. Understanding the strengths and limitations of each model is essential to determine the most effective approach for sentiment classification tasks.

Although numerous studies have explored sentiment analysis on user reviews and game-related datasets, most of them focus on either binary classification or relatively balanced datasets. In addition, previous comparative

Vol. 1 No.4 May 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i4.20

studies often emphasize model performance without considering the impact of class imbalance in multi-class sentiment classification. This limitation is particularly important in real-world mobile game reviews, where neutral opinions are often underrepresented and difficult to classify accurately.

Furthermore, limited studies have specifically examined Indonesian-language mobile game reviews using both Machine Learning and Deep Learning approaches within the same experimental framework. Therefore, there is a need for a more comprehensive evaluation that considers not only model performance but also the challenges posed by imbalanced multi-class data.

This study explores the effectiveness of Machine Learning and Deep Learning models for multi-class sentiment classification under imbalanced data conditions. Specifically, this study compares LightGBM and CNN models to analyze their performance differences, robustness, and limitations in handling minority classes in real-world mobile game reviews. The results of this study are expected to provide insights into the effectiveness of each approach and contribute to the development of more accurate sentiment analysis models.

The main contributions of this study are as follows:

1. A comparative evaluation of Machine Learning and Deep Learning models for multi-class sentiment classification using real-world mobile game reviews.
2. An empirical analysis of the impact of class imbalance, with particular emphasis on the underrepresented neutral class.
3. Identification of the limitations of current approaches in handling imbalanced multi-class sentiment data.
4. Practical insights into model selection by considering performance trade-offs and computational efficiency.

This study differentiates itself from prior work by systematically analyzing how class imbalance affects model behavior at both overall and class-specific levels, particularly for the underrepresented neutral class, within a unified experimental framework using real-world mobile game reviews.

2. RELATED WORK

2.1. Sentiment Analysis on User Reviews

Sentiment analysis has been widely applied to extract opinions and emotions from textual data, particularly in user-generated content such as online reviews and social media [1]. Several studies have demonstrated that sentiment analysis can effectively classify user opinions into multiple categories, including positive, negative, and neutral, providing valuable insights for decision-making processes [3], [4]. In the context of mobile games, user reviews serve as an important source of feedback that reflects player experience, satisfaction, and expectations [5].

Previous research has also explored sentiment analysis specifically on game reviews, showing that user feedback can be systematically analyzed to evaluate game quality and user satisfaction [6]. In addition, previous studies have applied Machine Learning techniques such as Random Forest to analyze sentiment in game reviews from platforms like Steam, demonstrating that user-generated feedback can be effectively utilized to assess player satisfaction and game performance [7]. Furthermore, multi-class sentiment classification has been widely adopted to provide a more comprehensive understanding of user opinions compared to binary classification approaches [4], [8]. These studies highlight the importance of sentiment analysis as a tool for understanding user behavior and improving digital products, particularly in the gaming industry.

2.2. Machine Learning and Deep Learning Approaches.

Various approaches have been proposed for sentiment analysis, particularly using Machine Learning (ML) and Deep Learning (DL) techniques. Traditional ML models rely on feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF), which have proven effective in representing textual data and improving classification performance [6], [9]. In particular, ML-based classification methods have shown strong performance in handling structured textual features and relatively smaller datasets [6].

On the other hand, Deep Learning models such as Convolutional Neural Networks (CNN) have been widely used due to their ability to automatically learn feature representations from raw text data. These models are capable of capturing complex patterns and contextual information, making them suitable for sentiment classification tasks involving unstructured data [2], [6]. Several comparative studies have evaluated the performance of ML and DL approaches, indicating that each method has its own strengths depending on the dataset characteristics and preprocessing techniques [6].

In addition, class imbalance remains a significant challenge in sentiment analysis, particularly in multi-class classification scenarios where certain classes are underrepresented. Previous studies have addressed this issue using techniques such as class weighting and data resampling to improve model performance, especially for minority classes [10]. Therefore, handling class imbalance is an important consideration in developing reliable sentiment analysis models.

Based on these previous studies, it can be concluded that both ML and DL approaches offer distinct advantages and limitations. Therefore, a comparative analysis is necessary to evaluate their effectiveness in sentiment analysis of mobile game reviews, particularly in the presence of imbalanced data.

3. METHODS

3.1. Data Collection

The dataset used in this study consists of 5000 user reviews collected from the Google Play Store, which

Vol. 1 No.4 May 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i4.20

represent user feedback on mobile games. These reviews contain opinions regarding gameplay experience, performance, and overall satisfaction. Previous studies have shown that user-generated reviews are a valuable source for sentiment analysis in evaluating digital products [5], [6].

3.2. Sentiment Labeling

To perform sentiment classification, the rating scores are transformed into three sentiment categories: negative, neutral, and positive. Reviews with scores less than or equal to 2 are categorized as negative, a score of 3 is categorized as neutral, and scores greater than or equal to 4 are categorized as positive. This multi-class labeling approach has been widely used in sentiment analysis to provide a more comprehensive understanding of user opinions [4].

3.3. Data Preprocessing

The textual data undergoes several preprocessing steps to improve data quality. These steps include removing punctuation, special characters, and irrelevant symbols, as well as converting all text into lowercase. These preprocessing techniques are commonly applied in sentiment analysis tasks to enhance model performance [3], [9], [11].

3.4. Data Splitting

The dataset is divided into training and testing sets using a stratified sampling approach to preserve the original class distribution. A ratio of 80% is used for training data and 20% for testing data. This approach ensures that each sentiment class is proportionally represented in both datasets, resulting in more reliable model evaluation.

3.5. Feature Extraction

For the Machine Learning model, the Term Frequency–Inverse Document Frequency (TF-IDF) method is used to convert textual data into numerical features. TF-IDF is a widely used term-weighting scheme in information retrieval that reflects the importance of a term in a document relative to a document collection and is calculated as follows [9]:

$$TF - IDF(d,t) = TF(d,t) \cdot \log\left(\frac{N}{df(t)}\right) \quad (1)$$

Where: N is total number of documents, df(t) is the number of documents containing term t. This formulation follows the standard approach in text mining for sentiment analysis [12], [13].

For the Deep Learning model, text data is processed using tokenization and sequence padding. The Tokenizer converts text into sequences of integers, while padding ensures uniform input length for the neural network.

3.6. Handling Imbalanced Data

The dataset exhibits an imbalanced distribution, where the neutral class contains significantly fewer samples compared to the positive and negative classes. The distribution of sentiment classes is illustrated in Figure 1, where the neutral class is significantly underrepresented compared to the positive and negative classes. This imbalance can negatively impact the model's ability to learn and accurately classify minority class instances [14], [15].

To address this issue, a class weighting technique is applied during model training. This method assigns higher importance to minority classes, allowing the model to learn more effectively without modifying the original data distribution. Previous studies have highlighted the importance of handling imbalanced datasets in improving classification performance [12], [13].

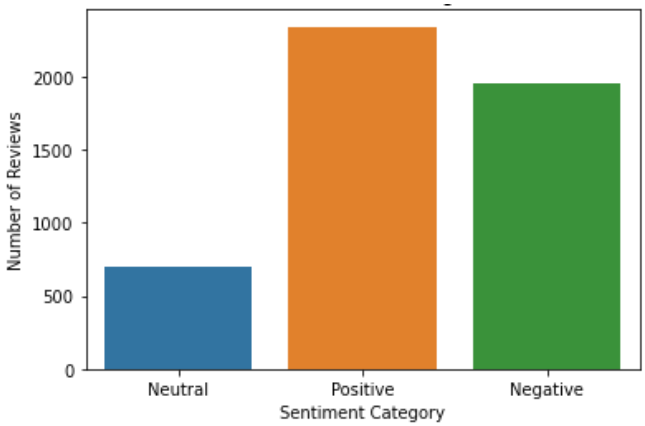

Fig.1. Distribution of Sentiment Classes in the Dataset

3.7. Model Development

This study employs two different approaches for sentiment classification: a Machine Learning model and a Deep Learning model.

The Machine Learning model used is Light Gradient Boosting Machine (LightGBM), which is known for its efficiency and strong performance in classification tasks involving structured features [8]. LightGBM is trained using TF-IDF features extracted from the textual data.

The Deep Learning model used is Convolutional Neural Network (CNN). CNN has been widely used for sentence classification tasks and has shown strong performance in capturing local semantic features in text data [2]. The CNN architecture consists of an embedding layer, a convolutional layer, a pooling layer, and fully connected layers. This model is capable of capturing contextual and semantic relationships within text data, making it suitable for sentiment analysis tasks [8].

3.8. Model Evaluation

To evaluate the performance of the models, several evaluation metrics are used, including accuracy, precision, recall, and F1-score. Accuracy is calculated as follows[4], [15], [16]:

Vol. 1 No.4 May 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i4.20

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where (TP) represents true positives, (TN) represents true negatives, (FP) represents false positives, and (FN) represents false negatives. In addition, precision, recall, and F1-score are used to provide a more comprehensive evaluation, particularly in handling imbalanced datasets. Confusion matrices are also used to analyze classification performance across different sentiment classes, allowing a detailed comparison between the Machine Learning and Deep Learning models.

4. RESULTS AND DISCUSSIONS

4.1. Experimental Results.

This section presents the performance evaluation of the Machine Learning and Deep Learning models used in this study, namely Light Gradient Boosting Machine (LightGBM) and Convolutional Neural Network (CNN). The evaluation is conducted using accuracy, precision, recall, and F1-score metrics. The results of the classification performance are summarized in Table 1.

TABLE 1. Results of the Classification Performance

Model	Accuracy	Precision	Recall	F1-score
LightGBM	66.40%	57.00%	57.00%	57.00%
CNN	68.20%	59.00%	57.00%	57.00%

Based on Table 1, the CNN model achieved a slightly higher accuracy of 68.20% compared to LightGBM, which achieved 66.40%. However, both models show similar macro-average F1-scores, indicating comparable performance across all classes. The comparison of model accuracy is illustrated in Figure 2.

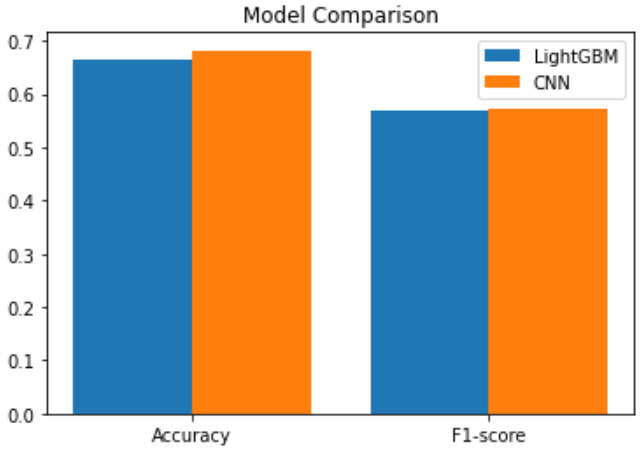

Fig.2. Model Accuracy Comparison

4.2. Class-Level Performance Analysis.

A detailed analysis of class-level performance reveals differences in how each model handles sentiment categories. The confusion matrix of the LightGBM model is shown in Figure 3.

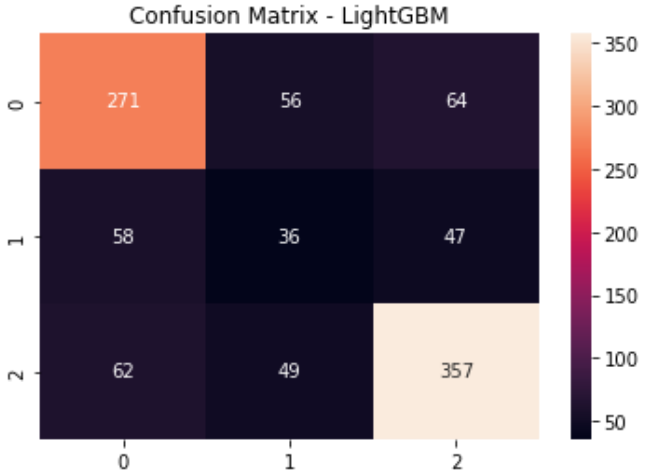

Fig.3. Confusion Matrix of LightGBM

For the LightGBM model, the positive class achieved the highest F1-score of 0.76, followed by the negative class with 0.69. However, the neutral class showed significantly lower performance, with an F1-score of only 0.26.
Similarly, the CNN model also performed well on the positive class with an F1-score of 0.75 and showed improved performance on the negative class with an F1-score of 0.72. However, the neutral class remained the most challenging, with an F1-score of 0.25. The confusion matrix of the CNN model is shown in Figure 4.

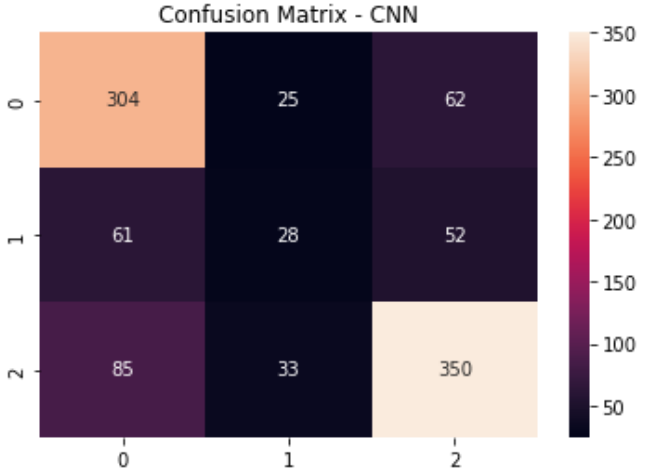

Fig.4. Confusion Matrix of CNN

These results indicate that both models struggle to accurately classify the neutral class.

4.3. Comparative Analysis.

The comparative analysis shows that the CNN model slightly outperforms LightGBM in terms of overall accuracy. This improvement may be attributed to CNN's ability to capture contextual information from text through embedding and convolutional operations.
On the other hand, LightGBM demonstrates competitive performance despite relying on TF-IDF features, which represent text in a sparse and structured form. This

Vol. 1 No.4 May 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i4.20

indicates that traditional Machine Learning approaches remain effective for sentiment analysis tasks, especially when combined with appropriate feature extraction techniques

4.4. Impact of Class Imbalance

One of the key factors influencing the performance of both models is class imbalance. The neutral class has significantly fewer samples compared to the positive and negative classes, which affects the model's ability to learn its characteristics.

Although class weighting techniques were applied during training, the results show that both models still struggle to classify the neutral class effectively. This finding is consistent with previous studies, which highlight that imbalanced datasets can lead to lower performance for minority classes [14], [15].

4.5. Discussion

Overall, both models demonstrate satisfactory performance for sentiment classification of mobile game reviews. The CNN model shows slightly better performance in terms of accuracy, while LightGBM remains competitive due to its simpler architecture and faster training process.

The findings indicate that the performance gap between Machine Learning and Deep Learning models is relatively small compared to the influence of data characteristics. This suggests that factors such as data quality, class distribution, and feature representation play a more critical role than model complexity in multi-class sentiment classification tasks. In this context, improvements in data preprocessing and imbalance handling strategies may contribute more significantly to performance gains than relying solely on more complex models.

Furthermore, while Deep Learning models such as CNN are effective in capturing contextual relationships in text data, Machine Learning models like LightGBM remain strong baselines due to their efficiency and robustness. Therefore, the selection of an appropriate model should consider factors such as dataset size, computational resources, and the presence of class imbalance.

5. CONCLUSIONS

This study conducted a comparative analysis of Machine Learning and Deep Learning approaches for sentiment analysis of mobile game reviews. Specifically, the performance of the Light Gradient Boosting Machine (LightGBM) and Convolutional Neural Network (CNN) models was evaluated using a multi-class sentiment classification framework consisting of positive, neutral, and negative categories.

The experimental results show that the CNN model achieved slightly higher accuracy (68.20%) compared to LightGBM (66.40%). However, both models demonstrated similar performance in terms of macro-average precision, recall, and F1-score, indicating comparable effectiveness across all sentiment classes.

Further analysis revealed that both models performed well in classifying positive and negative sentiments but struggled to accurately classify the neutral class. This limitation is primarily influenced by class imbalance in the dataset, where the neutral class contains significantly fewer samples. Although class weighting techniques were applied, the performance for the minority class remains relatively low.

Overall, the findings indicate that Deep Learning models such as CNN can provide improved performance due to their ability to capture contextual information in text data. However, Machine Learning models like LightGBM remain competitive, offering advantages in terms of simplicity and computational efficiency.

For future work, it is recommended to explore more advanced techniques for handling class imbalance, such as data augmentation or hybrid resampling methods. In addition, the use of larger datasets and more complex Deep Learning architectures may further improve classification performance, particularly for underrepresented classes.

REFERENCES

[1] B. Liu, “Sentiment Analysis and Opinion Mining”.
[2] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” Sep. 03, 2014, *arXiv*: arXiv:1408.5882. doi: 10.48550/arXiv.1408.5882.
[3] Z. Yejian and S. Takada, “Review Classification Based on Machine Learning: Classifying Game User Reviews,” *IEEE Access*, vol. 11, pp. 142447–142463, 2023, doi: 10.1109/ACCESS.2023.3342294.
[4] G. Mutanov, V. Karyukin, and Z. Mamykova, “Multi-Class Sentiment Analysis of Social Media Data with Machine Learning Algorithms,” *Computers, Materials & Continua*, vol. 69, no. 1, pp. 913–930, 2021, doi: 10.32604/cmc.2021.017827.
[5] Y. Yu, T. Dinh, F. Yu, and V.-N. Huynh, “Understanding Mobile Game Reviews Through Sentiment Analysis: A Case Study of PUBGm,” in *Model and Data Engineering*, vol. 14396, M. Mosbah, T. Kechadi, L. Bellatreche, and F. Gargouri, Eds., in Lecture Notes in Computer Science, vol. 14396. , Cham: Springer Nature Switzerland, 2024, pp. 102–115. doi: 10.1007/978-3-031-49333-1_8.
[6] J. Y. Tan and A. S. K. Chow, “Sentiment Analysis on Game Reviews: A Comparative Study of Machine Learning Approaches,” in *Conference Proceedings: International Conference on Digital Transformation and Applications (ICDXA 2021)*, Tunku Abdul Rahman University College, 2021, pp. 209–216. doi: 10.56453/icdxa.2021.1023.
[7] M. Dwifebri Purbolaksono, “Sentiment Analysis of Game Review in Steam Platform using Random Forest,” *ijoict*, vol. 10, no. 2, pp. 161–169, Dec. 2024, doi: 10.21108/ijoict.v10i2.1007.
[8] K. T. Wong, “The Data-Driven Myth and the Deceptive Futurity of ‘the World’s Fastest Growing Games Region’: Selling the Southeast Asian Games Market via Game Analytics,” *Games and Culture*, vol. 18, no. 1, pp. 42–61, Jan. 2023, doi: 10.1177/15554120221077731.
[9] C. Manning, P. Raghavan, and H. Schuetze, “Introduction to Information Retrieval,” 2009.
[10] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” *Prog Artif Intell*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.
[11] N. VasfiSisi and M. R. F. Derakhshi, “Text Classification with Machine Learning Algorithms,” 2013.
[12] M. Das, “A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset”.

Vol. 1 No.4 May 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i4.20

[13] J. Y. Tan, A. S. K. Chow, and C. W. Tan, "A Comparative Study of Machine Learning Algorithms for Sentiment Analysis of Game Reviews," *TJIEM*, vol. 82, no. 3, Nov. 2022, doi: 10.54552/v82i3.101.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *jair*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[15] D. I. Sumantiawan, J. E. Suseno, and W. A. Syafei, "Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction," *J. Sistem Info. Bisnis*, vol. 13, no. 1, pp. 1–9, Jun. 2023, doi: 10.21456/vol13iss1pp1-9.

[16] M. I. U. Sarkar, A. Srinivasulu, A. Lal, R. Naidu, and M. M. Rahman, "Classification of toxic element accumulation in rice grains using optimized machine learning models: A comparative study," *Journal of Food Composition and Analysis*, vol. 148, p. 108504, Dec. 2025, doi: 10.1016/j.jfca.2025.108504.