**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 42, No. 3, June 2026, pp. 818~826
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i3.pp818-826 

# Statistical comparison of MLP and LSTM for mobile health sentiment analysis

**Ghanim Kanugrahan[1], Win Ce[2], Vito Hafizh Cahaya Putra[1], Yudi Ramdhani[3], Febriyanti Panjaitan[4]**
[1]Department of Computer Science, Faculty of Creative Technology, Satu University, Bandung, Indonesia
[2]Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
[3]Department of Information Systems, Faculty of Creative Technology, Satu University, Bandung, Indonesia
[4]Department of Computer Science, Faculty of Creative Technology, Satu University, Palembang, Indonesia

**Article Info**

***Article history:***

Received Jun 30, 2025
Revised Mar 7, 2026
Accepted May 26, 2026

***Keywords:***

Long short-term memory
Mobile app
Multi-layer perceptron
Sentiment analysis
Wilcoxon test

**ABSTRACT**

This study investigates user sentiment towards the Mobile JKN public health application by applying text classification models based on deep learning. Two approaches were compared: a multi-layer perceptron (MLP) with TF-IDF features and long short-term memory (LSTM) with Word2Vec embeddings. The dataset consists of 114,364 Indonesian-language user reviews collected from the Google Play Store. To address class imbalance, we applied random oversampling. Each model was evaluated using 5-fold stratified shuffle split cross-validation. The results showed that MLP models achieved higher accuracy (up to 83.90%), while LSTM models demonstrated better recall and precision on minority classes such as neutral sentiment. However, statistical validation using the Wilcoxon signed-rank test revealed that the performance differences between models were not statistically significant ($p > 0.05$). These findings suggest that both models are viable for sentiment analysis, with trade-offs depending on the evaluation metric of interest. Future work may explore hybrid architecture and larger datasets for improved performance and statistical confidence.



***Corresponding Author:***

Win Ce
Information Systems Department, School of Information Systems
Bina Nusantara University
11480 West Jakarta, DKI Jakarta, Indonesia
Email: wn.@binus.edu

## 1. INTRODUCTION

The rapid growth of mobile applications has significantly changed how people interact with digital services. With smartphones and the internet becoming more affordable, mobile apps have become essential tools in modern life [1]. This trend has grown rapidly over the past decade around the world [2], [3]. People now use these mobile applications for many activities, from social networking to managing daily tasks. Mobile applications have expanded across various sectors, including education, finance, and transportation [4]–[6]. One of the sectors that has been greatly affected by mobile app development is healthcare [7], [8]. Health-related mobile applications now provide services such as telemedicine, appointment scheduling, and medical record management [9], [10].

In Indonesia, the Mobile JKN app developed by BPJS Kesehatan serves as a good example of this healthcare transformation [11]. The app is designed to help participants of the national health insurance (JKN-KIS) program and offers various digital services, including registration, facility search, premium payment, and medical history tracking [12]. Despite being useful and widely used, Mobile JKN has mixed

user satisfaction levels. This can be seen in the many reviews posted on the Google Play Store, which provide valuable feedback about user experiences, complaints, and expectations.

However, this feedback is typically unstructured and difficult to interpret manually due to its large volume and varied nature [13]. With so many user reviews available, BPJS Kesehatan, as the owner of the Mobile JKN application, can take advantage of this data to understand what users experience and improve the quality of the application [9]. Several studies have explored the use of sentiment analysis and machine learning to understand user opinions from app reviews [14]–[16]. These reviews provide useful insights into the methods and models that can be applied to analyze unstructured text data from mobile applications [17].

Nurfikri [18] conducted a study on the Halodoc telemedicine app. The aim of the study was to analyze user sentiment towards Halodoc in Indonesia after the COVID-19 pandemic began to decrease. The dataset consisted of 1,129 user reviews collected from the Google Play Store between June and August 2022. He used quantitative analysis to categorize user ratings into five sentiment levels. After that, he also used qualitative analysis using NVIVO software to identify the most common words in positive and negative reviews. The study found that 74.8% of the reviews were positive, while 15.5% were negative. The results suggest that although user sentiment was mostly positive, there is still room for improvement, especially in system performance and service quality.

Hou and Zhu [19] conducted a study to identify the usefulness of online product reviews by combining grounded theory and a multi-layer perceptron (MLP) neural network. The research used 6,215 user reviews collected from JD.com, focusing on mobile phone products. The authors first conducted semi-structured interviews with 35 consumers to extract features related to different stages of the purchasing process. The processes are demand generation, information collection, product evaluation, purchase decision, and post-purchase behavior. These features were then used to train an MLP model. The results showed that the MLP model trained with features based on consumer decision-making achieved an F1-score of 89.3%, significantly outperforming models using only TF-IDF features (F1-score of 59.2%) and traditional classifiers like SVM and Naïve Bayes. This research suggests that including behavioral context in feature design can improve the performance of machine learning models in identifying useful online reviews.

While these studies show great results, determining the true effectiveness of different models requires more than just comparing performance metrics. When evaluating machine learning models, researchers must ensure that observed performance differences are statistically significant rather than occurring by chance. Recognizing this critical need for statistical validation, Sharma and Kaur [20] conducted a benchmarking study involving 35 deep learning models for aspect-level sentiment classification (ALSC). They used eight benchmark datasets and evaluated model performance based on accuracy, macro-F1 score, and training time. To check the statistical validity of the performance differences, they applied the Friedman test, followed by post-hoc tests such as the Nemenyi and Wilcoxon signed-rank tests. The Wilcoxon test was used to perform pairwise comparisons between top-performing models. Their results showed that models like GAT-BERT and ASGCN achieved the best performance, although some trade-offs existed between accuracy and efficiency.

In sentiment analysis research, the selection of feature extraction method and model architecture plays an important role in determining classification performance. One commonly used technique is Term Frequency-Inverse Document Frequency (TF-IDF), which converts text into numerical vectors by giving weight to terms based on their frequency in a document and their rarity across the dataset [21]. This method has been proven effective in traditional machine learning tasks, as it helps capture important terms while minimizing the influence of commonly used words [22], [23].

MLP is a basic deep learning model consisting of several fully connected layers that can learn complex patterns in data using backpropagation [24]. When combined with TF-IDF features, MLP models can effectively recognize relationships between words in text classification tasks, especially because they are capable of modeling non-linear dependencies between features [25], [26].

On the other hand, Word2Vec is a feature extraction technique that transforms words into dense vector representations based on their context in sentences [27]. Unlike the sparse vectors generated by TF-IDF, Word2Vec preserve the semantic meaning of words, allowing the model to understand similarity based on usage [28]. These dense embeddings are particularly well-suited for long short-term memory (LSTM) networks, which are a type of recurrent neural network designed to handle sequence data [29]. LSTMs have special gating mechanisms that allow them to retain or forget information as needed, making them effective for understanding the flow and context of natural language text [30], [31].

Several previous studies have applied machine learning and deep learning techniques for sentiment analysis on mobile applications, including healthcare platforms such as Mobile JKN. Existing works generally report that models such as LSTM and traditional classifiers can effectively classify user sentiment, with performance evaluation primarily based on accuracy-based metrics [11], [12]. However, these studies rely on direct metric comparisons without statistical validation, making it unclear whether observed performance differences between models are significant. In addition, comparative analyses between different

deep learning architectures and feature extraction methods, particularly MLP with TF-IDF and LSTM with Word2Vec, remain limited for large-scale Indonesian mobile health review data. Therefore, this study aims to systematically compare these two modeling approaches within a unified experimental framework and to employ the Wilcoxon signed-rank test to statistically evaluate the significance of performance differences.

## 2. METHOD

This study applies a machine learning approach to analyze user sentiment from Mobile JKN reviews. The process consists of four main stages. The stages are data collection, data preprocessing, model building, and model evaluation. Figure 1 illustrates the complete experimental setup and workflow used in this study.

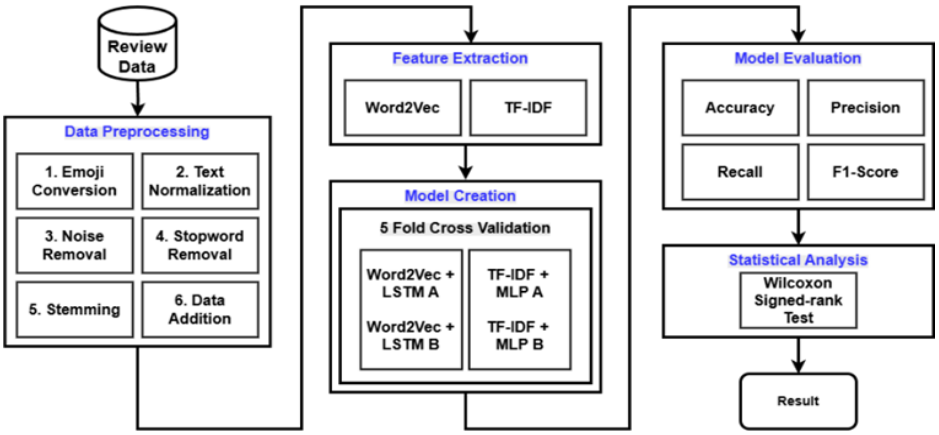


Figure 1. Research framework

### 2.1. Data collection

User review data was collected from the Google Play Store using the google_play_scraper Python library. Data scraping was conducted on April 30, 2025, and it includes reviews from version 1.18 (July 21, 2021) up to version 4.12 (April 29, 2025). We focused only on reviews written in Indonesian langage. Each review was extracted along with other data such as review content, rating score, and timestamp. From the total number of Mobile JKN reviews (more than 2 million reviews), we only selected 200,000 reviews as a sample or around 10% total reviews, for this study. The extracted data can be described in Table 1.

Table 1. Sample of collected user reviews

| Review Id | User name | Content | Score | AppVersion |
|---|---|---|---|---|
| be77e4d3… | Junaedi | *Setelah di update baru... mantaap... bisa pake…* | 5 | 3.4.0 |
| 76e0cbe7… | Ekotri | *Jaga kesehatan tetap semangat terapkan hidup…* | 5 | 3.4.0 |
| 6d6e3fed… | Andri Lukmana | *tidak bisa masuk* | 1 | 1.18 |

### 2.2. Data preprocessing

After the reviews were collected, we added a new column called sentiment. In this column, we transformed the rating scores into sentiment labels: reviews with a score of 1 or 2 were labeled as negative (label 0), scores of 3 as neutral (label 1), and scores of 4 or 5 as positive (label 2). The text preprocessing pipeline consisted of the following steps:

- Emoji conversion: emojis were converted into textual representation using the emoji library.
- Text normalization: text was converted to lowercase.
- Noise removal: punctuation, elongated characters, and URLs were removed.
- Stopword removal and stemming: we used the Indonesian stopword list from NLTK and applied stemming using the Sastrawi stemmer.

The review text was available in the content column. The cleaned text was stored in a new column called cleaned_content. During preprocessing, we found that many user reviews were extremely short, such as “ok” or “top”, which caused the cleaned_content to become empty after filtering. To address this issue, we retained these short texts by copying them directly into the cleaned_content column when the cleaned result was empty. This ensured that all reviews still reflected their original meaning and sentiment.

Next, duplicate reviews were removed to avoid biased learning from repeated samples. After that, only two columns were retained for model training and evaluation, which are cleaned_content and sentiment. The final data that will be used can be described in Table 2.

Table 2. Sample of extracted data from data preprocessing

| cleaned_content | sentiment |
|---|---|
| *Update mantaap pake nik* | 2 |
| *Jaga sehat semangat terap hidup sehat* | 2 |
| *Masuk* | 0 |

### 2.3. Model building

Before training the models, random oversampling was applied to handle the class imbalance in the dataset. After preprocessing, the negative sentiment class had the largest number of samples, while the neutral and positive classes were much smaller. To reduce this imbalance, the RandomOverSampler from the imblearn library was used to oversample the neutral and positive classes by duplicating existing samples until their numbers were equal to the negative class. This oversampling process was applied only to the training data in each fold of the cross-validation, while the test data was kept unchanged. This step was intended to help the models learn more evenly from all sentiment classes.

We use a deep learning approach to compare the performance of two types of models: MLP with TF-IDF features, and LSTM with Word2Vec embeddings. We applied the TF-IDF vectorizer with unigram (ngram_range=(1,1)) and a maximum of 5,000 features to convert the cleaned text into numerical vectors. On the other hand, we trained a Word2Vec model on the tokenized reviews using vector_size=100, window=5, and min_count=1. The embedding for each review was obtained by averaging the vectors of the words it contained. We also use the Adam optimizer with a learning rate of 0.001 and were trained using CrossEntropyLoss. The models were trained for 20 epochs with a batch size of 64. Four model architectures were built:

- MLP A: one hidden layer with 128 units and dropout.
- MLP B: two hidden layers with 128 and 64 units respectively, both with dropout.
- LSTM A: one LSTM layer with `hidden_dim=128` with dropout and a dense layer.
- LSTM B: two stacked LSTM layers with dimensions 128 and 64 with dropout and a dense layer.

All experiments were conducted in Python using PyTorch, with TF-IDF and Word2Vec implemented via Scikit-learn and Gensim. Training was accelerated using an NVIDIA 3050Ti Laptop GPU.

### 2.4. Model evaluation

The dataset was split using StratifiedShuffleSplit with 5 splits with 80% training and 20% testing set. This method ensured that the proportion of sentiment labels remained consistent across the training and test sets. In each fold of the cross-validation, both MLP models were trained using the TF-IDF feature extraction, while both LSTM models were trained using Word2Vec feature extraction. After training, each model was evaluated on the test set, and the classification accuracy along with the confusion matrix was recorded. Once all folds were completed, the average accuracy of each model was calculated and compared to assess overall performance across the dataset.

After all folds were completed, the average accuracy of each model was compared. This evaluation aimed to determine whether one model type significantly outperformed the others in terms of classification accuracy. The Wilcoxon Signed-Rank Test was then used to examine the statistical significance between:

- MLP A vs MLP B
- LSTM A vs LSTM B
- The best MLP vs the best LSTM

## 3. RESULTS AND DISCUSSION

This section presents and discusses the results of sentiment classification experiments using MLP and LSTM models. The evaluation focuses on classification performance based on accuracy, confusion matrix, and additional metrics such as precision, recall, and F1-score. In addition, we conducted a statistical test to determine whether performance differences between models are significant. The objective is to assess how well each model classifies user sentiment in the Mobile JKN app reviews.

### 3.1. Data preprocessing result

Initially, the dataset contained 200,000 user reviews. After removing duplicate entries and rows with missing values (NaN), the number of usable data points decreased to 114,364. This data reduction reflects the

importance of the cleaning process in improving data quality and ensuring that only valid reviews were used in the analysis. The data cleaning process also affected the distribution of sentiment classes in the dataset. Several reviews were removed unevenly across sentiment categories, which altered the original balance between positive, negative, and neutral reviews. As a result, the cleaned dataset exhibited a shift in class proportions compared to the original data.

As illustrated in Figure 2, positive reviews experienced the largest reduction, decreasing from 121,813 to 42,800 samples. Negative reviews were reduced from 70,902 to 65,411 samples, while neutral reviews showed the smallest change, decreasing from 7,215 to 6,153 samples. This uneven reduction across sentiment classes caused negative reviews to become the dominant class in the cleaned dataset.

The resulting class imbalance was addressed during the model training stage by applying a random oversampling technique. In this process, the neutral and positive sentiment classes were treated as minority classes and were oversampled by duplicating existing samples until their sizes matched the negative sentiment class. This strategy was adopted to reduce bias toward the majority class and to allow the models to learn sentiment patterns more evenly.

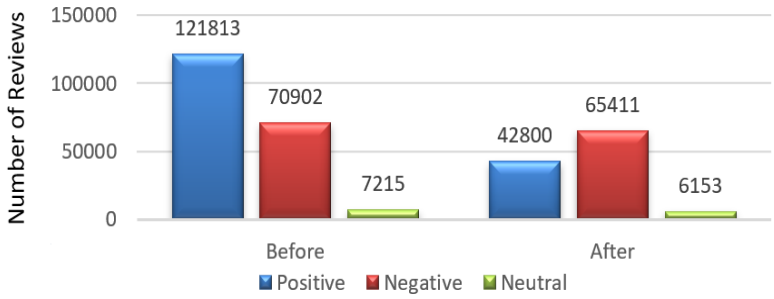


Figure 2. Data distribution before and after

### 3.2. Model training and evaluation setup

After balancing the dataset, we trained four different models: MLP A, MLP B, LSTM A, and LSTM B. Each model was evaluated using 5-fold cross-validation to ensure robust and generalizable results. The 5-fold cross-validation approach divides the data into five equal parts, where each fold serves as a test set while the remaining four folds are used for training. The average confusion matrix results across all five folds, showing how each model classified samples in each sentiment category can be described in Table 3.

Table 3 shows the classification performance of each model across the three sentiment classes: negative, neutral, and positive. When looking at negative sentiment, MLP models worked better than LSTM models. MLP B correctly identified 11,970 negative reviews, which was the best result, while MLP A identified 11,754 negative reviews. The LSTM models did not perform as well, especially LSTM A, which incorrectly labeled many negative reviews as neutral (3,296 reviews). This shows that MLP models are better at recognizing negative sentiments.

For neutral sentiment, all models had difficulty, which makes sense because neutral feelings are harder to detect. However, LSTM models performed slightly better here. LSTM A correctly identified 622 neutral reviews, which was much better than MLP A (81) and MLP B (71). This suggests that LSTM models might be better at understanding subtle neutral sentiments because they can process text in sequence.

Table 3. Average confussion matrix

| | | Predicted label | | |
|---|---|---|---|---|
| | | Negative | Neutral | Positive |
| True label: negative | MLP A | 11754 | 403 | 923 |
| | MLP B | 11970 | 359 | 752 |
| | LSTM A | 9193 | 3296 | 592 |
| | LSTM B | 9360 | 2958 | 763 |
| True label: neutral | MLP A | 877 | 81 | 272 |
| | MLP B | 916 | 71 | 242 |
| | LSTM A | 451 | 622 | 157 |
| | LSTM B | 564 | 452 | 214 |
| True label: positive | MLP A | 1117 | 150 | 7292 |
| | MLP B | 1270 | 140 | 7148 |
| | LSTM A | 568 | 1051 | 6940 |
| | LSTM B | 786 | 779 | 6994 |

When classifying positive sentiment, MLP models again performed better than LSTM models. MLP A had the highest number of correct predictions (7,292), followed by MLP B (7,148). LSTM A often confused positive reviews with neutral ones (1,051 times), which made its overall performance lower for positive sentiment classification.

To better understand the performance of each model, four evaluation metrics were calculated, namely accuracy, precision, recall, and F1-score using macro averaging. Macro averaging assigns equal importance to each class and is therefore suitable for imbalanced datasets. The average performance results across five-fold cross-validation, along with the standard deviation of accuracy to reflect performance stability across folds, are presented in Table 4.

Based on the accuracy results, the MLP-based models achieved higher overall accuracy than both LSTM models, with MLP B slightly outperforming MLP A. The accuracy difference between the best-performing MLP model (MLP B) and the best-performing LSTM model (LSTM B) was approximately 10%. However, a different pattern can be observed when examining precision, recall, and F1-score.

LSTM A achieved the highest recall value of 67.31% and the highest precision of 64.27%, indicating that this model was more effective at correctly identifying relevant samples across sentiment classes. Despite having lower overall accuracy, this result suggests that the LSTM-based model may better capture patterns from minority classes, particularly the neutral sentiment category.

The observed performance differences between the MLP and LSTM models can be further explained by the nature of the review texts in the dataset. Most user reviews in the Mobile JKN application are relatively short and contain limited contextual information. In such cases, TF-IDF representations are effective in highlighting discriminative keywords associated with sentiment, which benefits MLP-based classifiers despite their simpler architecture. On the other hand, LSTM models rely on sequential information and distributed word representations. Although this approach may not maximize overall accuracy on short texts, it enables the model to better capture subtle semantic patterns, particularly for minority classes such as neutral sentiment. This explains why LSTM A achieved higher recall and precision values, even though its accuracy was lower than that of the MLP models.

These findings highlight a trade-off between achieving high overall accuracy and effectively identifying underrepresented sentiment classes. Therefore, the choice of model should be aligned with the application objective. For large-scale sentiment monitoring where accuracy is prioritized, MLP models are more suitable. However, for applications that require deeper analysis of neutral or ambiguous user feedback, LSTM-based models may offer additional advantages.

Table 4. Average score

| | MLP A | MLP B | LSTM A | LSTM B |
|---|---|---|---|---|
| Accuracy | 83.63 ± 0.28% | 83.90 ± 0.18% | 73.26 ± 0.69% | 73.48 ± 0.90% |
| Precision | 61.40% | 61.60% | 64.27% | 61.98% |
| Recall | 60.55% | 60.27% | 67.31% | 63.34% |
| F1 | 60.62% | 60.46% | 61.48% | 60.00% |

### 3.3. Statistical significance testing

To determine whether the observed performance differences were statistically meaningful, we conducted Wilcoxon signed-rank tests comparing the models. This test is suitable for comparing paired samples, such as evaluation scores from cross-validation results. The first set of tests was conducted using the accuracy scores obtained from each of the five folds. The evaluation was conducted in three stages. First, we compared the two MLP models (MLP A vs MLP B) to assess whether the additional hidden layer in MLP_B led to a significant performance gain. Second, we compared the two LSTM models (LSTM A vs LSTM B) to evaluate whether the deeper LSTM B outperformed the simpler LSTM A. Finally, the best-performing model from each group, MLP B and LSTM B, was compared to determining which architecture offered better overall performance in sentiment classification. The results are presented in Table 5.

As shown in Table 5, the differences between model A and model B within the same architecture family were not statistically significant, as both p-values (0.4375 and 0.8125) were above the standard threshold of 0.05. This indicates that adding additional layers or complexity in the B variants did not consistently improve model performance across the folds.

The comparison between the best MLP model (MLP B) and the best LSTM model (LSTM B) resulted in a p-value of 0.0625, which is slightly above the 0.05 threshold. Although this result is not statistically significant, it is close to the boundary, suggesting that MLP B may have better performance than LSTM B. With a larger dataset or more validation folds, this difference might reach statistical significance in future research.

Table 5. The result using wilcoxon signed-rank test

| Comparison | p-value | Significant |
|---|---|---|
| MLP A vs MLP B | 0.4375 | No |
| LSTM A vs LSTM B | 0.8125 | No |
| MLP B vs LSTM B | 0.0625 | No |

## 4. CONCLUSION

This study compared the performance of MLP and LSTM models for sentiment classification on user reviews of the Mobile JKN app. The results showed that MLP models using TF-IDF features achieved higher overall classification accuracy, while LSTM models with Word2Vec embeddings demonstrated better precision and recall, particularly for neutral sentiment. However, statistical testing using the Wilcoxon signed-rank test indicated that the observed performance differences were not statistically significant. These findings suggest that both MLP and LSTM approaches are comparably viable for sentiment analysis of short user-generated text in this application context.

For future work, more advanced feature representations such as contextual word embeddings (e.g., IndoBERT) can be explored to better capture semantic information in short and ambiguous reviews. In addition, alternative techniques for handling class imbalance, such as synthetic oversampling methods, may be investigated to improve minority class recognition. Finally, expanding the dataset or incorporating model interpretability analysis could provide deeper insights and potentially lead to more statistically significant performance differences.

## FUNDING INFORMATION

This research was supported by Satu University and Binus University, both of which are part of Binus Higher Education. The authors acknowledge the institutional support provided during this study.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ghanim Kanugrahan | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| Win Ce | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Vito Hafizh Cahaya Putra | | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | |
| Yudi Ramdhani | | ✓ | | ✓ | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| Febriyanti Panjaitan | | ✓ | | ✓ | ✓ | ✓ | | ✓ | | ✓ | | ✓ | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, Win Ce, upon reasonable request.

## REFERENCES

[1] S. Gunathilaka and N. De Silva, "Aspect-based sentiment analysis on mobile application reviews," in *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, IEEE, Nov. 2022, pp. 183-188. doi: 10.1109/ICTer58063.2022.10024070.

[2] A. Subbarao, A. Siddika, M. A. Fathullah, and M. A. Bin Sanwani, "The role of mobile applications in shaping digital transformation in higher education among generation i: a bibliographic study," *Information*, vol. 16, no. 12, p. 1026, Nov. 2025, doi: 10.3390/info16121026.

[3] J. Joseph, S. Vineetha, and N. V. Sobhana, "A survey on deep learning based sentiment analysis," *Materials Today: Proceedings*, vol. 58, pp. 456–460, 2022, doi: 10.1016/j.matpr.2022.02.483.

[4] K. T. Shandana, A. Aminuddin, E. H. Saputra, F. F. Abdulloh, M. Rahardi, and B. P. Asadduloh, "Sentiment analysis of google classroom application using machine learning techniques," in *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, IEEE, Nov. 2023, pp. 954-959. doi: 10.1109/ICAMIMIA60881.2023.10427706.

[5] N. I. Lestari, S. M. Taib, W. Wibowo, I. A. Aziz, and M. R. Habibi, "Aspect-based sentiment analysis for mobile app review using convolutional neural network (CNN) and Word2Vec," in *2024 IEEE 7th International Conference on Electrical, Electronics and System Engineering (ICEESE)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICEESE62315.2024.10828541.

[6] A. Amalia, D. Gunawan, and K. Nasution, "Sentiment analysis of GO-JEK services quality using multi-label classification," *Journal of Physics: Conference Series*, vol. 1830, no. 1, p. 012003, Apr. 2021, doi: 10.1088/1742-6596/1830/1/012003.

[7] T. Mescher, R. L. Hacker, L. A. Martinez, C. D. Morris, M. C. Mishkind, and C. E. Garver-Apgar, "Mobile health apps: guidance for evaluation and implementation by healthcare workers," *Journal of Technology in Behavioral Science*, vol. 10, no. 2, pp. 224-235, Sep. 2025, doi: 10.1007/s41347-024-00441-7.

[8] Kartini, A. K. Darmawan, R. I. Syah, and M. Makruf, "Sentiment analysis of social media for indonesian m-health PeduliLindungi mobile-apps(PLMA) with lexicon-based and support vector machine approach," in *2023 IEEE 9th Information Technology International Seminar (ITIS)*, IEEE, Oct. 2023, pp. 1-7. doi: 10.1109/ITIS59651.2023.10419994.

[9] M. Haoues, R. Mokni, and A. Sellami, "Machine learning for mHealth apps quality evaluation: an approach based on user feedback analysis," *Software Quality Journal*, vol. 31, no. 4, pp. 1179–1209, Dec. 2023, doi: 10.1007/s11219-023-09630-8.

[10] F. Alkhuzaimi, D. Rainey, C. Brown Wilson, and J. Bloomfield, "The impact of mobile health interventions on service users' health outcomes and the role of health professions: a systematic review of systematic reviews," *BMC Digital Health*, vol. 3, no. 1, p. 3, Feb. 2025, doi: 10.1186/s44247-024-00143-3.

[11] G. Tamami, W. A. Triyanto, and S. Muzid, "Sentiment analysis mobile JKN reviews using SMOTE based LSTM," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 19, no. 1, p. 13, Jan. 2025, doi: 10.22146/ijccs.101910.

[12] E. Septiani, T. M. Akhriza, and M. Husni, "Comparison of the accuracy between naive bayes classifier and support vector machine algorithms for sentiment analysis in mobile JKN application reviews," *Transactions on Informatics and Data Science*, vol. 1, no. 1, pp. 21–32, Apr. 2024, doi: 10.24090/tids.v1i1.12232.

[13] A. A. Qureshi, M. Ahmad, S. Ullah, M. N. Yasir, F. Rustam, and I. Ashraf, "Performance evaluation of machine learning models on large dataset of android applications reviews," *Multimedia Tools and Applications*, vol. 82, no. 24, pp. 37197–37219, Oct. 2023, doi: 10.1007/s11042-023-14713-6.

[14] M. Irsad and A. Khare, "Sentiment classification on mobile review using extraction of sentiment conveying sentences," *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, vol. 5, no. 2, pp. 1–8, 2024, doi: 10.54060/a2zjournals.jieee.116.

[15] S. Ahammad, S. A. Sinthia, M. Ahmed, M. Hossain, N. A. A. Asif, and N. A. Ikram, "Deep learning-based sentiment analysis of user generated reviews of various AI powered mobile applications," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, Apr. 2024, pp. 505–512. doi: 10.1109/ICICT60155.2024.10544699.

[16] P. R. Henao, J. Fischbach, D. Spies, J. Frattini, and A. Vogelsang, "Transfer learning for mining feature requests and bug reports from tweets and app store reviews," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, IEEE, Sep. 2021, pp. 80–86. doi: 10.1109/REW53955.2021.00019.

[17] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken Dw, F. A. Bachtiar, and N. Yudistira, "BERT fine-tuning for sentiment analysis on Indonesian mobile apps reviews," in *Proceedings of the 6th international conference on sustainable information engineering and technology*, New York, NY, USA: ACM, Sep. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.

[18] A. Nurfikri, "Sentiment analysis telemedicine apps reviews using NVIVO," in *The 5th International Conference on Vocational Education Applied Science and Technology 2022*, Basel Switzerland: MDPI, Dec. 2022, p. 4. doi: 10.3390/proceedings2022083004.

[19] J. Hou and A. Zhu, "Identification of usefulness for online reviews based on grounded theory and multilayer perceptron neural network," *Applied Sciences*, vol. 13, no. 9, p. 5321, Apr. 2023, doi: 10.3390/app13095321.

[20] T. Sharma and K. Kaur, "Benchmarking deep learning methods for aspect level sentiment classification," *Applied Sciences*, vol. 11, no. 22, p. 10542, Nov. 2021, doi: 10.3390/app112210542.

[21] M. Q. H. Octava, D. G. Prasetyo Putri, F. M. Hilmy, U. Farooq, R. A. Nurhaliza, and A. Ganjar, "Web-based sentiment analysis system Using SVM and TF-IDF with statistical feature," in *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2023*, IEEE, Nov. 2023, pp. 9–14. doi: 10.1109/3ICT60104.2023.10391734.

[22] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[23] S. Singh, K. Kumar, and B. Kumar, "Sentiment analysis of Twitter data using TF-IDF and machine learning techniques," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, IEEE, May 2022, pp. 252–255. doi: 10.1109/COM-IT-CON54601.2022.9850477.

[24] A. Al Ryan, H. D. Arpita, A. Tabassum, M. S. Ahammad, and M. A. Akram, "A comparative analysis between deep learning and machine learning algorithms based on user review sentiment analysis from various OTT applications," in *2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, IEEE, Feb. 2024, pp. 1–7. doi: 10.1109/ICCECE58645.2024.10497229.

[25] G. M. Raza, Z. S. Butt, S. Latif, and A. Wahid, "Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models," in *2021 International conference on digital futures and transformative technologies (ICoDT2)*, IEEE, May 2021, pp. 1-6. doi: 10.1109/ICoDT252288.2021.9441508.

[26] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: approaches, datasets, and future research," *Applied Sciences*, vol. 13, no. 7, p. 4550, Apr. 2023, doi: 10.3390/app13074550.

[27] R. A. Mangngalle, M. D. Purbolaksono, and W. Astuti, "Sentiment analysis of Lazada app review using Word2Vec and support vector machine," in *2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, IEEE, Dec. 2023, pp. 182–187. doi: 10.1109/ICICyTA60173.2023.10428771.

[28] H. Jiang, C. Hu, and F. Jiang, "Text sentiment analysis of movie reviews based on Word2Vec-LSTM," in *2022 14th International conference on advanced computational intelligence (ICACI)*, IEEE, Jul. 2022, pp. 129–134. doi: 10.1109/ICACI55529.2022.9837505.

[29] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian Hotel reviews," *Procedia Computer Science*, vol. 179, pp. 728–735, 2021, doi: 10.1016/j.procs.2021.01.061.

[30] J. Shin *et al.*, "Exploring the effectiveness of machine learning and deep learning algorithms for sentiment analysis: a systematic literature review," *Computers, Materials and Continua*, vol. 84, no. 3, pp. 4105–4153, 2025, doi: 10.32604/cmc.2025.066910.

[31] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 4997–5053, Oct. 2021, doi: 10.1007/s10462-021-09973-3.

## BIOGRAPHIES OF AUTHORS

**Ghanim Kanugrahan** Ghanim Kanugrahan received his bachelor's degree in computer science from Universitas Dian Nuswantoro in 2019, and his master's degree in information technology from Universitas Indonesia in 2021. His main areas of expertise include data science, sentiment analysis, and machine learning. He is currently a lecturer in the informatics program at Satu University Bandung, where he is actively involved in teaching, research, and community service. He can be contacted at email: ghanim.kanugrahan@univ.satu.ac.id.

**Win Ce** is an academic and professional that have been active as a lecturer at BINUS University and SATU University, specifically within the Information Systems study program. His experience is not limited to academia; he frequently serves as a speaker and expert for various companies, fulfilling an expert role in the industry. Win Ce, S. Kom., M.M.'s research interests span several key areas in information technology, including software development, machine learning, decision support systems (DSS), and information systems. He can be contacted at email: wn@binus.edu and win.ce@univ.satu.ac.id.

**Vito Hafizh Cahaya Putra** developed an interest in Informatics through the Computer and Robotics extracurricular activity in 2011. He then pursued a bachelor's degree in informatics engineering at Universitas Widyatama and later completed his postgraduate studies at Universitas Langlangbuana. He is currently teaching at Universitas Satu in the Informatics program, covering subjects such as object-oriented programming, algorithms, IoT, and web programming. His research interests are focused on artificial intelligence (AI), machine learning, natural language processing, and the IoT. He can be contacted at email: vito.putra@univ.satu.ac.id.

**Yudi Ramdhani** is a lecturer and researcher at Satu University in the Information Systems Study Program. He holds a master's degree in computer science and has focused his research on data mining, data science, and decision support systems. With several years of experience in both teaching and research, he has been actively involved in developing data-driven academic projects and innovative curriculum design. He has published research papers in reputable journals and regularly contributes to academic forums and seminars. He can be contacted at email: yudi.ramdhani@univ.satu.ac.id.

**Febriyanti Panjaitan** earned her doctoral degree from the Faculty of Engineering, Universitas Sriwijaya, with a concentration in Informatics Engineering, in 2023. Her main areas of expertise include data science, machine learning, and deep learning, with a research focus on developing AI models for complex data analysis. She is a lecturer in the Informatics Engineering Program Faculty of Creative Technology at Satu University, and she is actively involved in teaching, research, and community service. She can be contacted at email: pebrianti.panjaitan@univ.satu.ac.id.