Journal of Information Technology and Applications Research

Journal homepage: https://jitar.org/index.php/jitar/article

# Fine-Tuning IndoBERTweet for Sentiment Analysis of Indonesian Restaurant App Reviews: A Case Study of McDonald's Indonesia

*Hafiz Mahardika[1*], Dhian Sweetania[2], Bambang Yulianto[3]*

Gunadarma University, Indonesia

*Correspondence: E-mail: hafizmahardika44@gmail.com

## ABSTRACT

The increasing adoption of mobile applications in the restaurant industry has generated large volumes of user reviews that provide valuable insights into customer satisfaction and application performance. However, manually analyzing thousands of reviews is inefficient, highlighting the need for automated sentiment analysis techniques. This study aims to develop and evaluate a fine-tuned IndoBERTweet model for sentiment analysis of user reviews of the McDonald's Indonesia mobile application on Google Play Store while identifying dominant complaint patterns that can support application improvement. A total of 4,000 Indonesian-language reviews were collected using Google Play Scraper. After removing boycott-related reviews and performing text preprocessing, 3,897 reviews were retained for analysis. Sentiment labels were automatically assigned based on star ratings, and the training dataset was balanced using Random Oversampling and class-weighted optimization. The IndoBERTweet model was fine-tuned using an 80:20 train–test split for three training epochs in Google Colaboratory. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis, while negative reviews were further examined through unigram and bigram frequency analysis. The proposed model achieved an accuracy of 85.38%, with a weighted precision of 88.16%, a weighted recall of 85.38%, and a weighted F1-score of 86.64%. Negative sentiment was classified most accurately (F1-score = 91.72%), whereas neutral sentiment remained challenging because of severe class imbalance. Complaint analysis revealed that the most common user issues involved login authentication, application errors, slow performance, device compatibility, promotions, and loyalty-point redemption. These findings demonstrate that fine-tuned IndoBERTweet is effective for analyzing Indonesian restaurant application reviews and can provide actionable insights for improving digital service quality and user experience.

## ARTICLE INFO

***Article History:***
*Submitted/Received 15 Jan 2026*
*First Revised 20 Feb 2026*
*Accepted 09 Mar 2026*
*Publication Date 30 May 2026*

***Keyword:***
*IndoBERTweet; Sentiment Analysis; Google Play Store; Mobile Application Reviews; Restaurant Applications*

## 1. INTRODUCTION

The rapid growth of mobile commerce has significantly transformed the food and beverage industry by enabling customers to access restaurant services through mobile applications. Restaurant applications now function not only as ordering platforms but also as customer engagement channels that integrate digital payments, loyalty programs, personalized promotions, and delivery services. As the number of users increases, online reviews submitted through application marketplaces such as Google Play Store have become an important source of customer feedback, reflecting user satisfaction, service quality, and application performance. Unlike traditional surveys, online reviews are generated continuously and voluntarily, providing valuable real-time insights for service providers to evaluate customer experiences and identify software-related issues (Tarwoto, Nugroho, Azka, & Graha, 2025). Consequently, automatically analyzing large volumes of user-generated reviews has become an essential task for organizations seeking to improve digital services.

Sentiment analysis has emerged as one of the most widely adopted Natural Language Processing (NLP) techniques for extracting opinions and emotions from textual data. Recent advances in deep learning, particularly Transformer-based language models, have substantially improved sentiment classification by capturing contextual semantic representations more effectively than conventional machine learning approaches. Pre-trained language models such as BERT and its variants have demonstrated superior performance across various text classification tasks (Devlin, Chang, Lee, & Toutanova, 2019). For Indonesian-language sentiment analysis, IndoBERTweet has attracted considerable attention because it was pre-trained using approximately 409 million Indonesian Twitter posts, allowing it to better understand informal expressions, abbreviations, slang, and conversational language commonly found in user-generated content (Koto, Lau, & Baldwin, 2021). Previous studies have reported that IndoBERTweet consistently outperformed conventional deep learning models such as Long Short-Term Memory (LSTM) in classifying Indonesian social media sentiment (Setiawan, Lhaksmana, & Bunyamin, 2023). Furthermore, IndoBERT and IndoBERTweet have shown competitive performance in analyzing user reviews from mobile applications and social media platforms (Jayadianti et al., 2022; Dewi, 2025).

Despite these advances, several research gaps remain. First, most previous studies have focused on social media platforms, public service applications, or general mobile applications rather than restaurant ordering applications, whose reviews exhibit unique linguistic characteristics combining informal language with application-specific technical complaints. Second, existing studies generally emphasize sentiment classification accuracy while providing limited discussion regarding dominant complaint patterns that can directly support application improvement and business decision-making. Third, highly imbalanced sentiment distributions, particularly the scarcity of neutral reviews, remain a significant challenge that may reduce model generalization. Although techniques such as Random Oversampling and class weighting have been proposed to mitigate class imbalance (Alvin & Winarsih, 2025), their integration with fine-tuned IndoBERTweet for Indonesian restaurant application reviews has received limited attention. Therefore, further investigation is needed to evaluate the robustness of IndoBERTweet within this specific application domain.

Based on these research gaps, this study proposes a fine-tuned IndoBERTweet model for sentiment analysis of Google Play Store reviews of the McDonald's Indonesia mobile application. The study collected 4,000 user reviews, removed boycott-related comments to maintain analytical objectivity, performed text preprocessing, automatically labeled

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

sentiment using user ratings, balanced the training dataset using Random Oversampling, and fine-tuned IndoBERTweet using class-weighted optimization. In addition to evaluating classification performance through Accuracy, Precision, Recall, and F1-score, this research identifies dominant complaint patterns extracted from negative reviews. The findings are expected to contribute both theoretically by extending the application of IndoBERTweet beyond social media domains and practically by providing actionable insights for improving restaurant mobile application quality and user experience.

## 2. METHODS

### 2.1 Research Design

This study employed a quantitative computational approach with a case study design to classify the sentiment of Indonesian-language user reviews of the McDonald's Indonesia mobile application. The analysis focused on three sentiment classes—negative, neutral, and positive—and subsequently examined the dominant complaint patterns contained in negative reviews. The complete research procedure consisted of data acquisition, data filtering, text preprocessing, automatic sentiment labeling, dataset partitioning, class balancing, tokenization, IndoBERTweet fine-tuning, model evaluation, and complaint-pattern analysis. All computational procedures were implemented using Python in Google Colaboratory.

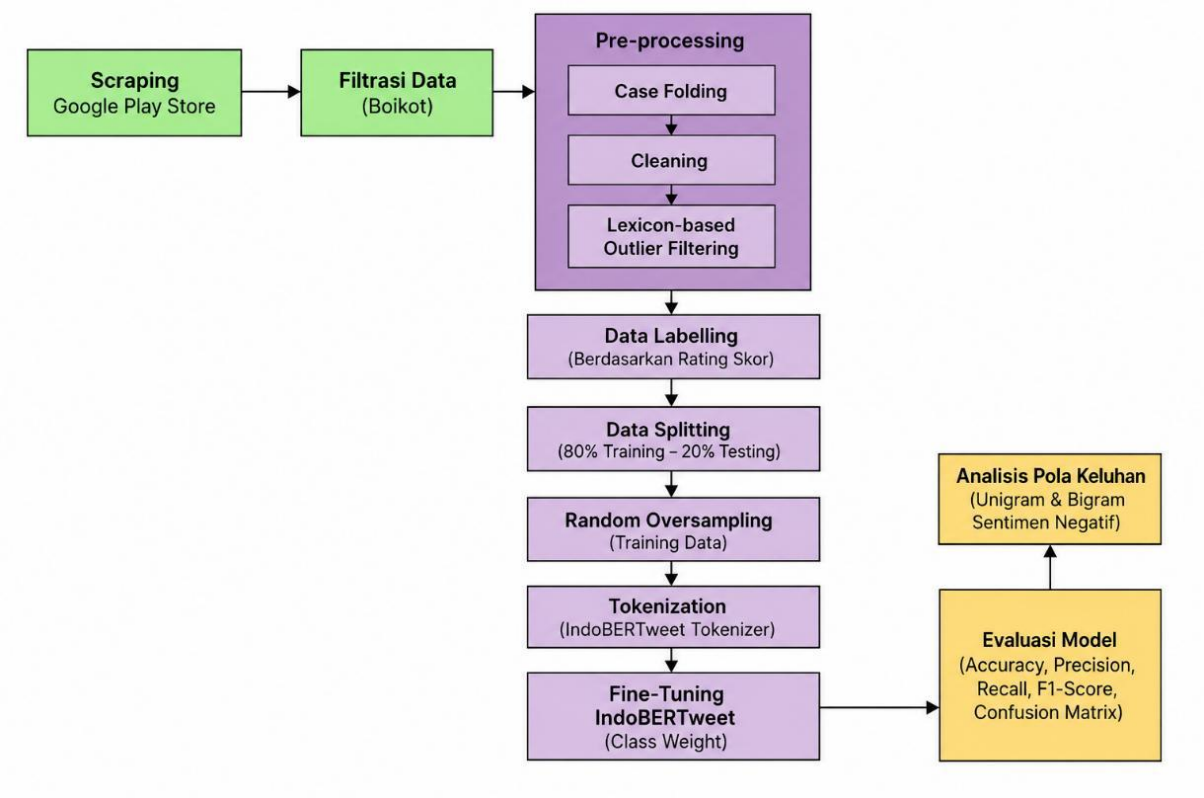


**Figure 1.** Research workflow for IndoBERTweet-based sentiment analysis

The workflow shown in Figure 1 should be retained because it provides a concise visual representation of the entire methodological process. However, the figure labels should be translated into English and standardized for journal publication.

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

### 2.2 Data Source and Collection

The research data consisted of Indonesian-language user reviews obtained from the McDonald's Indonesia application page on the Google Play Store. A total of 4,000 reviews were collected automatically using the google-play-scraper Python library. The scraping configuration used the application package identifier com.mcdonalds.mobileapp, with the language and country parameters both set to Indonesia and the reviews arranged from the newest entries.

The extracted variables were limited to the review text and the associated star rating. The original review content was stored as review_text, while the numerical star score was stored as rating. Google Play Store reviews were selected because they contain direct user evaluations of the application's technical performance, accessibility, ordering process, payment system, promotional features, and loyalty services.

After the initial collection, reviews containing the keyword *boikot* were removed. This filtration was conducted because boycott-related reviews may express opinions influenced by external social or political issues rather than users' direct experiences with the application's technical functions. The original dataset contained 4,000 reviews, of which 34 boycott-related reviews were identified and excluded, leaving 3,966 reviews for further preprocessing.

The Google Play Scraper library was used because it enables the automated collection of review content and application metadata from Google Play Store (JoMingyu, 2024).

### 2.3 Text Preprocessing

Text preprocessing was conducted to transform unstructured user reviews into cleaner textual inputs suitable for the IndoBERTweet tokenizer. The preprocessing procedure included case folding, URL removal, non-alphabetic character removal, whitespace normalization, and lexicon-based outlier filtering.

Case folding converted all uppercase characters into lowercase characters. This process ensured that words with different capitalization patterns were treated as identical linguistic units. URLs were removed using regular expressions because links did not contribute directly to the sentiment classification task. Numbers, emojis, punctuation marks, and other non-alphabetic characters were also removed. Multiple spaces resulting from the cleaning procedure were subsequently normalized into single spaces.

Unlike conventional machine-learning pipelines, stemming and general stopword removal were not applied before model training. These processes were intentionally avoided because IndoBERTweet is a contextual Transformer-based language model that benefits from retaining natural sentence structure, informal language, function words, and morphological variations. Excessive preprocessing may eliminate contextual information needed by pre-trained language models to interpret user sentiment accurately (Khairani, Mutiawani, & Ahmadian, 2024).

An additional lexicon-based filtering procedure was applied to remove semantically irrelevant outliers. Reviews consisting exclusively of a single food-related word, such as *enak*,

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

*kenyang*, *ayam*, *burger*, *sedap*, or *mantap*, were excluded because they did not provide sufficient information regarding application functionality. Nevertheless, these terms were retained when they appeared as part of a longer review containing technical or service-related information. Following all filtering and preprocessing stages, 3,897 usable reviews remained.

| Rating Bintang | Ulasan Asli Pengguna | Hasil Teks Bersih (Cleaned) |
|---|---|---|
| 1 | Sudah di donlod karena tergiur promo, pas di donlod gak bisa di pesan, dan sialnya lagi hanya payment gopay, dan kartu kredit...ini aplikasi buat apa sih sebenarnya?? | sudah di donlod karena tergiur promo pas di donlod gak bisa di pesan dan sialnya lagi hanya payment gopay dan kartu kreditini aplikasi buat apa sih sebenarnya |
| 2 | aplikasi tdk propper, lama loginya, gejalanya banyak, diskon tdk ada biar pake apk. funanya pake apk ini sama pesan langsung apa? klo harganya sama aja. tiru tuh apk kfc, propper diskon banyak, pembayarannya banyak pilihannya. | aplikasi tdk propper lama loginya gejalanya banyak diskon tdk ada biar pake apk funanya pake apk ini sama pesan langsung apa klo harganya sama aja tiru tuh apk kfc propper diskon banyak pembayarannya banyak pilihannya |
| 3 | kurang suka | kurang suka |
| 4 | sebagai cust yang sering pakai apk. untuk update selanjutnya yang dimana, jika kode suatu barang ketika habis waktu(error dari mesin) maka si barang akan kembali ke halaman menu utama dan tidak hilang. exp : saya memesan salah satu menu promo katakan lah ice coffee float. ketika saya mau menukarkan kode tersebut ke dalam mesin melalui scan/imput kode secara manual. mesin akan menunjukkan *Kode yang anda masukkan tidak valid* saya ingin mencari promo yang sama, sudah tidak ada lagi atau HILANG | sebagai cust yang sering pakai apk untuk update selanjutnya yang dimana jika kode suatu barang ketika habis waktuerror dari mesin maka si barang akan kembali ke halaman menu utama dan tidak hilang exp saya memesan salah satu menu promo katakan lah ice coffee float ketika saya mau menukarkan kode tersebut ke dalam mesin melalui scanimput kode secara manual mesin akan menunjukkan kode yang anda masukkan tidak valid saya ingin mencari promo yang sama sudah tidak ada lagi atau hilang |
| 5 | oke mantap | oke mantap |

**Figure 2.** Examples of user reviews before and after text preprocessing

This figure is suitable for the Method section because it demonstrates how raw review texts were transformed into cleaned texts. The figure should be redesigned as an editable journal table or a high-resolution English-language figure rather than inserted as a screenshot.

### 2.4 Sentiment Labeling

The reviews were automatically labeled using their Google Play Store star ratings as proxy sentiment labels. Rating-based labeling was selected because the star score represents the evaluation explicitly assigned by users when submitting their reviews. The sentiment classes and numerical labels were defined as follows.

**Table 1.** Sentiment-labeling criteria based on star ratings

| Star rating | Sentiment class | Numerical label |
|---|---|---|
| 1–2 | Negative | 0 |
| 3 | Neutral | 1 |
| 4–5 | Positive | 2 |

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

Reviews with ratings of one or two stars were classified as negative because these scores generally represent dissatisfaction or technical difficulties. Three-star reviews were categorized as neutral because they tend to represent mixed, moderate, or non-extreme evaluations. Reviews with ratings of four or five stars were categorized as positive because they indicate general user satisfaction.

This automatic procedure enabled all collected reviews to be labeled consistently and efficiently. However, rating-based labeling may not always perfectly represent the linguistic content of a review because users may assign a rating that is inconsistent with their written comments. This limitation should be acknowledged when interpreting the model's performance.

### 2.5 Data Partitioning and Class Balancing

The labeled dataset was divided into training and testing subsets using stratified random splitting. Eighty percent of the reviews were allocated to the training set, while the remaining 20% formed the testing set. A fixed random_state value of 42 was used to ensure reproducibility. Stratification maintained the original relative distribution of negative, neutral, and positive classes in both subsets.

The initial training data showed substantial class imbalance. The negative class contained 1,992 reviews, the neutral class contained 105 reviews, and the positive class contained 1,020 reviews. The limited number of neutral reviews created a risk that the model would prioritize the majority classes and fail to learn the linguistic characteristics of neutral sentiment.

Random Oversampling was therefore applied exclusively to the training set. Neutral reviews were randomly duplicated with replacement until the number of neutral training samples reached 1,992, equal to the largest sentiment class. Oversampling was not applied to the testing set because modifying the test distribution could produce an unrealistic estimate of the model's generalization performance.

### 2.6 IndoBERTweet Tokenization and Model Configuration

The principal model used in this study was the pre-trained indolem/indobertweet-base-uncased checkpoint obtained through the Hugging Face Transformers library. IndoBERTweet was selected because it was pre-trained using a large Indonesian Twitter corpus and was specifically developed to process informal Indonesian-language expressions, abbreviations, conversational forms, and domain-specific vocabulary (Koto, Lau, & Baldwin, 2021).

The cleaned review texts were transformed into numerical model inputs using the IndoBERTweet tokenizer. The tokenization process produced input_ids and attention_mask tensors. A maximum input length of 128 tokens was specified because Google Play Store reviews are generally short. Reviews containing fewer than 128 tokens were padded to a fixed length, whereas longer reviews were truncated.

The model was configured for sequence classification with three output labels representing negative, neutral, and positive sentiment. The tokenized training and testing inputs were encapsulated in custom PyTorch dataset objects that paired the encoded text with the corresponding numerical sentiment labels.

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

### 2.7 Model Fine-Tuning

IndoBERTweet was fine-tuned in Google Colaboratory using a T4 Graphics Processing Unit. Training was performed for three epochs with a training batch size of 8 and an evaluation batch size of 8. The model was evaluated at the end of each epoch, and a checkpoint was saved at the same interval.

A customized Trainer class was implemented to integrate class weights into the cross-entropy loss function. The weighted loss increased the penalty imposed when the model incorrectly classified reviews belonging to underrepresented sentiment classes. The relevant training configuration included:

**Table 2.** Model Fine-Tuning

| Parameter | Configuration |
|---|---|
| Pre-trained model | indolem/indobertweet-base-uncased |
| Number of output classes | 3 |
| Maximum token length | 128 |
| Training epochs | 3 |
| Training batch size | 8 |
| Evaluation batch size | 8 |
| Train–test split | 80:20 |
| Random state | 42 |
| Hardware accelerator | GPU T4 |
| Class-balancing methods | Random Oversampling and class weighting |
| Model selection | Lowest validation loss |

The load_best_model_at_end parameter was activated so that the final evaluation used the checkpoint with the lowest validation loss rather than automatically using the parameters obtained from the final epoch. This configuration was particularly important because continued training may improve the fit to the training data while reducing generalization to unseen reviews.

## 3. RESULTS AND DISCUSSION

### 3.1 Dataset Preparation and Class Distribution

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

The initial scraping process collected 4,000 Indonesian-language reviews from the McDonald's Indonesia application page on Google Play Store. After removing 34 reviews containing boycott-related content, 3,966 reviews remained. The preprocessing and lexicon-based outlier filtering stages subsequently removed 69 short food-related reviews that did not provide information about application performance. Therefore, the final dataset consisted of 3,897 reviews.

The dataset was divided into 80% training data and 20% testing data using stratified splitting. Before Random Oversampling, the training dataset contained 1,992 negative reviews, 105 neutral reviews, and 1,020 positive reviews. This distribution indicated severe class imbalance, particularly for neutral sentiment. Random Oversampling increased the neutral training samples to 1,992, equal to the largest class, while the testing set remained unchanged to preserve the original data distribution.

**Table 3.** Distribution of training and testing data

| Sentiment class | Training data before ROS | Training data after ROS | Testing data |
|---|---|---|---|
| Negative | 1,992 | 1,992 | 499 |
| Neutral | 105 | 1,992 | 26 |
| Positive | 1,020 | 1,020 | 255 |
| **Total** | **3,117** | **5,004** | **780** |

The oversampling procedure substantially improved the representation of neutral reviews during training. However, it did not generate new linguistic variations because Random Oversampling only duplicated existing samples. Consequently, the neutral class remained limited in semantic diversity despite having an equal numerical representation in the training data.

The original bar chart comparing class composition before and after oversampling does not need to be included because Table 3 already presents the same information more efficiently.

### 3.2 Fine-Tuning Performance

The IndoBERTweet model was fine-tuned for three epochs using a batch size of 8. The training loss decreased consistently from 0.1449 in the first epoch to 0.0796 in the second epoch and 0.0306 in the third epoch. This pattern indicated that the model increasingly fitted the training dataset.

In contrast, the validation loss increased from 1.4802 in the first epoch to 1.6095 in the second epoch and 1.6796 in the third epoch. The opposing movement between training and validation loss indicated overfitting. The model became increasingly accurate on the training data but showed a declining ability to generalize to unseen reviews. Because load_best_model_at_end=True was activated, the final model used the checkpoint from the first epoch, which produced the lowest validation loss.

**Table 4.** Fine-tuning performance across epochs

| Epoch | Training loss | Validation loss | Accuracy | Weighted F1-score |
|---|---|---|---|---|
| 1 | 0.1449 | 1.4802 | 0.8538 | 0.8664 |
| 2 | 0.0796 | 1.6095 | 0.8756 | 0.8773 |
| 3 | 0.0306 | 1.6796 | 0.8692 | 0.8714 |

Although the second epoch generated a slightly higher accuracy, its validation loss was greater than that of the first epoch. Therefore, the first-epoch checkpoint was selected to reduce the risk of overfitting. This result also suggests that longer training does not necessarily

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

improve generalization, especially when the dataset contains highly repetitive oversampled observations.

The original training-log screenshot is not required. Table 4 provides a clearer and more publication-ready presentation of the training results.

### 3.3 Sentiment Classification Results

The final IndoBERTweet model achieved an overall accuracy of 85.38%. The weighted-average precision, recall, and F1-score were 88.16%, 85.38%, and 86.64%, respectively. These results indicate that the model performed well in classifying the overall sentiment of Indonesian restaurant application reviews.

**Table 5.** Classification performance of the fine-tuned IndoBERTweet model

| Sentiment | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Negative | 0.9127 | 0.9218 | 0.9172 | 499 |
| Neutral | 0.0392 | 0.0769 | 0.0519 | 26 |
| Positive | 0.9067 | 0.8000 | 0.8500 | 255 |
| **Accuracy** | | | **0.8538** | **780** |
| **Macro average** | **0.6195** | **0.5996** | **0.6064** | **780** |
| **Weighted average** | **0.8816** | **0.8538** | **0.8664** | **780** |

The negative class produced the highest F1-score of 91.72%. This performance was supported by the relatively large number of negative testing samples and the presence of explicit complaint expressions, such as *error*, *tidak bisa*, *gak masuk*, and *loading lama*. These linguistic patterns are generally easier for the model to distinguish because they contain strong negative polarity.

The positive class achieved an F1-score of 85.00%. Its precision was high at 90.67%, although its recall was lower at 80.00%. This result means that positive predictions were generally correct, but some genuinely positive reviews were classified into other classes.

The neutral class generated the weakest performance, with an F1-score of only 5.19%. Only 26 neutral reviews were available in the testing set, compared with 499 negative and 255 positive reviews. Neutral reviews also frequently contain mixed opinions, such as praise followed by criticism, making their semantic boundaries less distinct. These findings demonstrate that balancing the training quantity through oversampling did not fully overcome the limited diversity and ambiguity of neutral-language patterns.

The considerable difference between the macro-average F1-score of 60.64% and the weighted-average F1-score of 86.64% confirms that the model's overall performance was strongly influenced by the majority classes. Therefore, accuracy and weighted F1-score should not be interpreted independently from class-specific results.

### 3.4 Confusion Matrix Analysis

The confusion matrix provides a more detailed explanation of the classification errors. Of the 499 actual negative reviews, 460 were correctly classified as negative, while 20 were predicted as neutral and 19 as positive. Of the 255 actual positive reviews, 204 were correctly classified, 29 were predicted as neutral, and 22 were predicted as negative. In the neutral class, only 2 of 26 reviews were correctly identified.

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

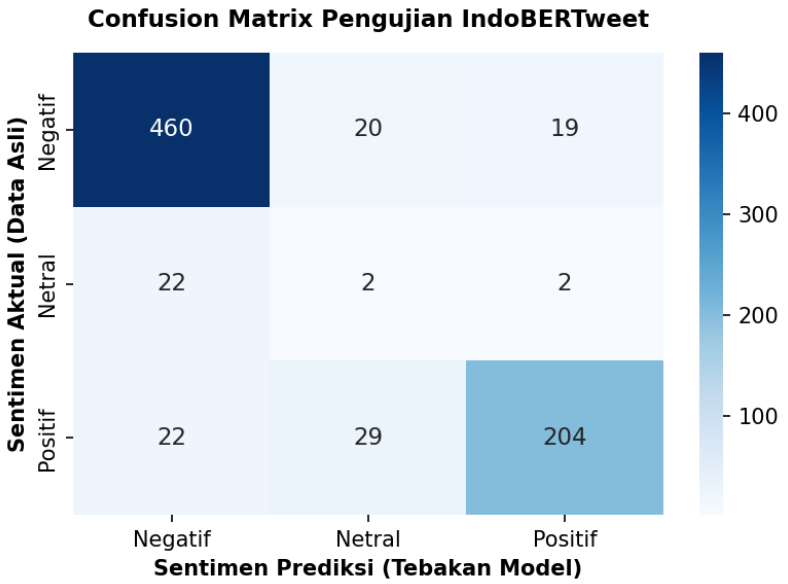


**Figure 3.** Confusion matrix of the fine-tuned IndoBERTweet model

The confusion matrix is the only classification figure recommended for inclusion because it visually complements Table 5 without duplicating the same information. The original classification-report image should be omitted because its values have already been presented in Table 5.

The confusion matrix shows that most predictions were concentrated along the diagonal for the negative and positive classes. However, the neutral class was frequently confused with the two more extreme sentiment categories. This outcome may be attributed to the use of star ratings as automatic labels. A three-star rating was treated as neutral, although the accompanying review might contain either positive or negative expressions. Thus, rating-based labeling may introduce label inconsistency when the written text does not fully correspond to the numerical score.

### 3.5 Dominant Complaint Patterns

A total of 2,491 negative reviews were analyzed using unigram and bigram frequency analysis. Negation words, including *tidak* and *gak*, were retained because they are important indicators of dissatisfaction. The analysis showed that the most frequent complaint-related words included *promo*, *login*, *masuk*, *error*, *poin*, *lama*, *update*, and *email*.

To avoid presenting two lengthy frequency tables, the unigram and bigram findings were summarized into four principal complaint themes.

**Table 6.** Main complaint themes identified from negative reviews

| Complaint theme | Representative words and phrases | Interpretation |
|---|---|---|
| Login and account verification | *kode OTP*, *gak masuk*, *tidak login*, *email* | Users experienced difficulties accessing accounts and receiving verification codes. |
| Application errors and accessibility | *sering error*, *tidak dibuka*, *gak buka*, *tidak berfungsi* | The application failed to open, crashed, or became unusable. |
| Performance and compatibility | *loading lama*, *tidak kompatibel*, *your device*, *baru download* | Users reported slow performance and device-compatibility problems, particularly after installation or updates. |

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

| Complaint theme | Representative words and phrases | Interpretation |
|---|---|---|
| Promotions and loyalty points | *promo*, *poin*, *tidak digunakan* | Users encountered difficulties accessing promotions or redeeming loyalty points. |

The most frequent bigram was *kode OTP*, which appeared 40 times. Other prominent phrases included *tidak dibuka* 29 times, *tidak digunakan* 23 times, *baru download* 22 times, and *sering error* 21 times. These patterns indicate that the strongest dissatisfaction was associated with authentication, application accessibility, and post-update stability.

The findings provide more practical value than sentiment labels alone. A negative classification only indicates dissatisfaction, whereas complaint-pattern analysis identifies the specific functional areas requiring improvement. For McDonald's Indonesia, application authentication and stability should be prioritized, followed by performance optimization, device compatibility, and clearer mechanisms for promotions and loyalty-point redemption.

### 3.6 Discussion

The overall results demonstrate that fine-tuned IndoBERTweet is effective for distinguishing strongly polarized Indonesian-language restaurant application reviews. The model's high performance for negative and positive reviews supports the suitability of IndoBERTweet for informal digital texts containing abbreviations, conversational expressions, and non-standard Indonesian vocabulary. This result is consistent with the design of IndoBERTweet, which was pre-trained on informal Indonesian social media data.

Nevertheless, the weak neutral-class performance reveals an important methodological limitation. Random Oversampling successfully balanced the number of training samples but could not create genuinely new semantic patterns. Repetition of the same neutral observations may also have contributed to overfitting, as indicated by the decreasing training loss and increasing validation loss. Future studies should therefore consider collecting more naturally occurring neutral reviews, applying manual annotation, or using data augmentation methods that generate linguistically diverse samples.

The use of star ratings as sentiment labels also requires careful interpretation. Ratings provide an efficient labeling mechanism for thousands of reviews, but rating–text inconsistency may occur. A user may provide three stars while expressing a predominantly negative complaint or provide a high rating while including criticism. Manual validation by multiple annotators would improve label reliability and help clarify ambiguous neutral reviews.

Despite these limitations, the model produced useful practical insights. The results show that sentiment classification can be combined with complaint mining to support application evaluation. The dominant problems were not related only to food or restaurant service but to digital service quality, particularly login verification, application errors, slow loading, compatibility, promotions, and loyalty points. Therefore, the proposed approach can support continuous monitoring of user experience and help application developers identify technical priorities from large-scale user feedback.

## 4. CONCLUSION

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021

This study demonstrated that fine-tuned IndoBERTweet can effectively classify Indonesian-language reviews of the McDonald's Indonesia mobile application into negative, neutral, and positive sentiment categories. Using 3,897 cleaned reviews collected from Google Play Store, the model achieved an overall accuracy of 85.38%, with a weighted precision of 88.16%, a weighted recall of 85.38%, and a weighted F1-score of 86.64%. The strongest performance was obtained for negative sentiment, with an F1-score of 91.72%, followed by positive sentiment at 85.00%. These results indicate that IndoBERTweet is well suited to processing informal Indonesian user reviews containing conversational expressions, abbreviations, and application-specific complaints.

The study also showed that combining sentiment classification with complaint-pattern analysis provides more actionable findings than sentiment labels alone. The most prominent negative-review themes were related to login and OTP verification, application errors, inaccessible features, slow loading, device compatibility, promotions, and loyalty points. These findings suggest that the main sources of user dissatisfaction were concentrated on the technical performance and usability of the application. Therefore, improvements to authentication reliability, post-update stability, loading performance, and promotion or point-redemption mechanisms should be prioritized by application developers.

However, the model performed poorly in identifying neutral sentiment, with an F1-score of only 5.19%. This limitation was influenced by the small number of neutral testing samples, the semantic ambiguity of three-star reviews, and the use of rating-based automatic labeling. Although Random Oversampling and class weighting increased the representation of the neutral class during training, these techniques did not fully address the lack of linguistic diversity. Future studies should employ a larger and more balanced dataset, incorporate manual annotation by multiple evaluators, compare IndoBERTweet with other Transformer-based models, and apply aspect-based sentiment analysis to identify user evaluations of specific application features in greater detail.

## 5. REFERENCES

Alvin, F., & Winarsih, N. A. S. (2025). Perbandingan kinerja model IndoBERT, IndoBERTweet, dan algoritma klasik pada analisis sentimen isu Indonesia Gelap. Building of Informatics, Technology and Science, 7(3), 1601–1613. https://doi.org/10.47065/bits.v7i3.8636

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019 (pp. 4171–4186)

Dewi, B. E. S. (2025). Sentiment analysis model on electric vehicles using IndoBERTweet and IndoBERT algorithm. Antivirus: Jurnal Ilmiah Teknik Informatika, 19(2), 192–202. https://doi.org/10.35457/w5r3g517

Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Dwiyanto, F. A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. ILKOM Jurnal Ilmiah, 14(3), 348–354. https://doi.org/10.33096/ilkom.v14i3.1505.348-354

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. Proceedings

of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10660–10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

Setiawan, J. C., Lhaksmana, K. M., & Bunyamin. (2023). Sentiment analysis of Indonesian TikTok review using LSTM and IndoBERTweet algorithm. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 8(3), 774–780. https://doi.org/10.29100/jipi.v8i3.3911

Tarwoto, Nugroho, R., Azka, N., & Graha, W. S. R. (2025). Analisis sentimen ulasan aplikasi Mobile JKN di Google Play Store menggunakan IndoBERT. Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), 9(2), 495–505. https://doi.org/10.35870/jtik.v9i2.3340

Aldhy, F., Salsabila, N. A., & Purwarianti, A. (2023). Sentiment analysis for Indonesian online reviews using IndoBERT and deep learning approaches. Journal of Information Systems Engineering and Business Intelligence, 9(2), 181–192. https://doi.org/10.20473/jisebi.9.2.181-192

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). A practical guide to sentiment analysis. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-55394-8

Liu, B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions (2nd ed.). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781108630016

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In China National Conference on Chinese Computational Linguistics (pp. 194–206). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253. https://doi.org/10.1002/widm.1253

DOI: https://doi.org/10.63956/jitar.v2i1.75
e-ISSN: 3089-8013; p-ISSN:3089-8021