Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Performance Evaluation of the BERT Model in Sentiment Analysis of DANA Application User Reviews

Hazael Susanto[1], Yanto[2], Weiskhy Steven Dharmawan[3], Riski Annisa[4], Lady Agustin Fitriana[5]*

[1,2,4] Program Studi Informatika,
[3] Program Studi Sistem Informasi Akuntansi,
[5] Program Studi Sistem Informasi
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Lady.lag@bsi.ac.id

Abstract

The rapid growth of digital wallets in Indonesia generates a large volume of user reviews on platforms such as the Google Play Store that cannot be efficiently analyzed manually. This study aims to evaluate the performance of the BERT (Bidirectional Encoder Representations from Transformers) model in sentiment classification tasks on a dataset of DANA application user reviews collected from the Google Play Store. The BERT model is fine-tuned using labeled Indonesian-language data with three sentiment classes: positive, negative, and neutral. Specialized preprocessing strategies are applied to handle the characteristics of informal text, abbreviations, and code-switching phenomena prevalent in Indonesian user reviews. Evaluation is conducted using accuracy, precision, recall, and F1-score metrics. Experimental results indicate that the fine-tuned IndoBERT model achieves an accuracy of 91.24% with a weighted F1-score of 0.91 on a test dataset of 6,106 samples. The Negative class achieves the highest performance with an F1-score of 0.95, followed by the Positive class (0.88) and Neutral class (0.84). This study provides empirical evidence of the effectiveness of the IndoBERT Transformer architecture for sentiment analysis in the Indonesian-language fintech domain and can serve as a reference for developing deep learning-based NLP systems in similar contexts.

Keywords: *Sentiment Analysis, BERT, Digital Wallet, DANA, Natural Language Processing, Transformer*

1. Introduction

The development of financial technology (fintech) in Indonesia has undergone significant acceleration over the past decade, driven by increasing internet penetration and high smartphone adoption rates among the population [1]. Digital wallets have emerged as one of the fastest-growing fintech segments, with platforms such as DANA, GoPay, and OVO dominating the national digital payment ecosystem [2]. DANA, launched in 2018 as a collaboration between Emtek Group and Ant Financial, has grown into one of Indonesia's leading digital payment platforms with millions of registered users [3]. This high volume of active users generates an enormous data ecosystem, including thousands of user feedback entries distributed across various digital platforms every day.

One of the most significant sources of user feedback is reviews on the Google Play Store, where the DANA application actively receives a large volume of comments from its users [4]. In this study, review data were collected using the google-play-scraper library, targeting up to 10,000 of the most recent Indonesian-language reviews, which were subsequently stored in CSV format for further analysis. The sheer volume of these reviews renders manual analysis infeasible, both in terms of time and resources [5]. Without a reliable automated mechanism, the strategic insights embedded in user review data cannot be fully leveraged by application developers.

Sentiment analysis is a branch of Natural Language Processing (NLP) concerned with automatically identifying and extracting opinion orientations from text, commonly classified into positive, negative, or neutral categories [6]. This study employs a lexicon-based sentiment labeling approach using the InSet (Indonesian Sentiment Lexicon) dictionary developed by Koto and Rahmaningtyas, downloaded directly from its official GitHub repository. The lexicon contains 3,609 positively valenced words and 6,609 negatively valenced word [7]. Sentiment scores are computed based on the difference between the count of positive and negative words in each review, yielding three label classes: positive, negative, and neutral. This approach enables automated large-scale labeling without the need for manual annotation.

Prior to model training, all review data underwent a comprehensive text preprocessing pipeline designed to address the informal characteristics of Indonesian review text [8]. The preprocessing stages, applied sequentially, include: removal of duplicate entries; removal of URLs, HTML tags, emojis, symbols, numbers, and punctuation; case folding; normalization of non-standard words using a standardized

word dictionary sourced from GitHub; space-based tokenization; stopword removal using the NLTK Indonesian stopword list; and stemming using the Sastrawi library. Upon completion of all preprocessing and labeling stages, the clean dataset used for model training comprised 6,106 review entries [9].

Sentiment analysis methodologies have undergone substantial evolution, progressing from lexicon-based and classical machine learning approaches to more advanced Transformer-based deep learning architectures [10]. The Transformer architecture, introduced by Vaswani et al., revolutionized language modeling through a self-attention mechanism capable of bidirectionally and in parallel modeling contextual dependencies between tokens, thereby overcoming the limitations of conventional recurrent models [10]. This advancement gave rise to a variety of pre-trained language models that can be fine-tuned for specific NLP tasks, including sentiment classification [11].

IndoBERT (indobert-base-p1) is a BERT-based model developed specifically for the Indonesian language by Koto et al. through the IndoLEM project, pre-trained on a large-scale corpus of Indonesian text [11]. In this study, IndoBERT is fine-tuned using the HuggingFace Transformers library with the following configuration: three training epochs, a learning rate of 2e-5, a batch size of 8, and a weight decay of 0.01. The data are split into training and test sets at an 80:20 ratio, and the best-performing model is selected based on the highest accuracy metric at the end of each epoch using the load_best_model_at_end strategy [12].

Although research on Transformer-based sentiment analysis has expanded rapidly, studies that specifically evaluate IndoBERT on Indonesian-language fintech application review datasets remain limited in the existing literature [13]. Indonesian user review text exhibits unique linguistic characteristics, including informal language, abbreviations, and code-switching phenomena that present challenges at the tokenization and semantic representation stages [14]. Based on this identified research gap, the present study aims to construct a comprehensive preprocessing pipeline for Indonesian-language review text, perform automated sentiment labeling using the InSet lexicon, fine-tune the IndoBERT (indobert-base-p1) model on a dataset of DANA application user reviews from the Google Play Store and evaluate model performance using accuracy, precision, recall, F1-score, and confusion matrix metrics.

2. Research Methods

This study consists of six sequential main stages: data collection, text preprocessing, sentiment labeling, data partitioning, IndoBERT model fine-tuning, and model evaluation. The research workflow is illustrated in Figure 1:

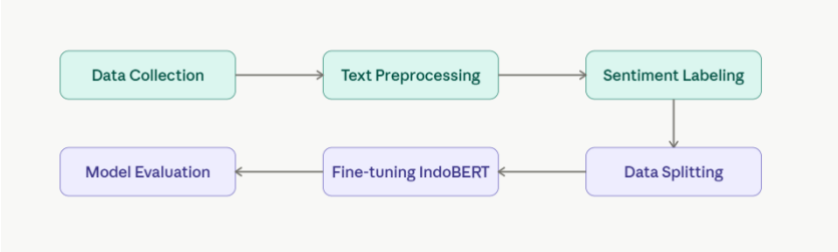


Fig. 1: Research workflow diagram

2.1. Data Collection

The data used in this study consist of DANA application user reviews collected from the Google Play Store using the google-play-scraper library. Data collection was performed with the parameters app_id = 'id.dana', language set to Indonesian ('id'), country set to Indonesia ('id'), sorting by most recent reviews (Sort.NEWEST), and a target quantity of 10,000 reviews. Each retrieved review contains the following attributes: reviewId, userName, score (star rating on a 1–5 scale), content (review text), and at (review date). A total of 10,000 reviews were successfully collected and subsequently stored in CSV format under the filename 'hasil_dana.csv' using the csv.DictWriter library.

```
from google_play_scraper import reviews, Sort

app_id = 'id.dana'

def get_reviews(app_id, lang='id', count=10000, sort=Sort.NEWEST, filter_scor
    try:
        result, continuation_token = reviews(
            app_id,
            lang=lang,
            country='id',
            sort=sort,
            count=count,
            filter_score_with=filter_score_with,
            filter_device_with=filter_device_with,
            continuation_token=continuation_token
        )

        return result, continuation_token
    except Exception as e:
        print("Error:", e)
```

Fig. 2: Data scraping process

2.2. Text Preprocessing

The preprocessing stage aims to clean and standardize review texts prior to model input. The process begins with the removal of duplicate entries using the drop_duplicates method applied to the 'Review Text' column, reducing the dataset from 10,000 to 6,280 unique reviews. Subsequent cleaning steps are applied sequentially using regular expressions (regex) and include: removal of URLs (https?://\S+|www\.\S+), removal of @mention usernames, removal of HTML tags, removal of emojis and Unicode symbols, removal of non-alphanumeric symbols, and removal of numeric characters.

Following the cleaning stage, case folding is applied by converting all text to lowercase. Non-standard word normalization is then performed using an Indonesian standardized word dictionary downloaded from a GitHub repository (analysisdatasentiment/kamus_kata_baku) in Excel format, whereby informal forms such as 'bagussss' are mapped to their standard form 'bagus'. Tokenization is carried out using simple space-based splitting (split()), after which stopword removal is applied using the Indonesian stopword list from the NLTK library. The final step is stemming using the Sastrawi library (version 1.0.1) to reduce each word to its root form. After all processing steps are completed and empty entries (NaN) are removed via dropna(), the resulting clean dataset contains 6,174 reviews.

2.3. Sentiment Labeling

Sentiment labeling is performed automatically using a lexicon-based approach employing the InSet (Indonesian Sentiment Lexicon) dictionary developed by Koto and Rahmaningtyas. The determine_sentiment() function computes the sentiment score of each review as the difference between the count of positive and negative words found in the text. Reviews with a score greater than 0 are labeled 'Positive,' those with a score less than 0 are labeled 'Negative,' and those with a score of 0 are labeled 'Neutral.' Following labeling, the dataset used for model training comprises 6,106 reviews after final filtering and cleaning.

2.4. Data Splitting

The final labeled dataset of 6,106 reviews is split into training and test sets using the train_test_split method from the scikit-learn library, with an 80:20 ratio and random_state=42 to ensure reproducibility. This partitioning yields 4,884 training samples (80%) and 1,222 test samples (20%). Sentiment labels, originally represented as text strings (Positive, Neutral, Negative), are converted to numerical values using astype('category').cat.codes prior to tokenization and model training.

2.5. Fine-tuning Model IndoBERT

The model employed in this study is IndoBERT (indobenchmark/indobert-base-p1), loaded using the BertForSequenceClassification class from the HuggingFace Transformers library (version 5.0.0). This model is a BERT variant pre-trained exclusively on a large corpus of Indonesian text. A classification head is added on top of the IndoBERT architecture with an output of three classes (Positive, Neutral, Negative).

Input tokenization was performed using BertTokenizer with the truncation=True parameter to limit the sequence length to the model's maximum limit (512 tokens). The fine-tuning process was performed using the HuggingFace Trainer API with the following configurations: number of epochs = 3, learning rate = 2×10^{-5}, batch size = 8, weight decay = 0.01, evaluation and storage strategy per epoch (eval_strategy="epoch"), and selection of the best model based on the highest accuracy (load_best_model_at_end=True).

2.6. Model Evaluation

Model performance is evaluated on the test set (1,222 samples) using standard text classification metrics: accuracy, precision, recall, and F1-score, computed per class using the classification_report function from the scikit-learn library. Additionally, a confusion matrix is visualized using seaborn.heatmap to analyze the distribution of model predictions across each sentiment class. Final predictions are obtained by taking the index of the highest logit value (np.argmax) from the model's output on the test data.

3. Results and Discussion

3.1. Data Collection Results

The data collection process from the Google Play Store using the google-play-scraper library successfully retrieved 10,000 DANA application user reviews in Indonesian, collected in April 2026 using the Sort.NEWEST parameter to ensure that the data represent the most recent user submissions. The collected data contain comprehensive information for each review, including: a unique review identifier (reviewId), the username of the reviewer (userName), the star rating assigned by the user (score on a 1–5 scale), the review text content (content), and the review submission timestamp (at). All data were stored in CSV format with a structured column arrangement to facilitate subsequent processing stages.

A deduplication step was applied to eliminate duplicate reviews that may have arisen due to application bugs or overlapping data retrieval. Following deduplication using the drop_duplicates() method, 6,280 unique reviews remained from the original 10,000 collected. This reduction of 37.2% indicates the presence of considerable duplication in the raw dataset. Figure 3 presents a sample of the raw data obtained from the collection process, illustrating the range of information captured for each user review, spanning star ratings from 1 (negative) to 5 (positive).

	Review ID	Username	Rating	Review Text	Date
0	5f75acfb-5258-4927-a758-06b93869cb7c	Pengguna Google	1	uang saya hilang begitu sajah saya mohon tolon...	2026-04-25 04:07:52
1	98d5728d-d4b4-4d22-96c9-f9e896670d57	Pengguna Google	5	good	2026-04-25 04:07:26
2	80ceba84-b0af-471f-9db8-8f4343acb34d	Pengguna Google	5	mantap	2026-04-25 04:04:22
3	71f40b21-cbd2-4fd9-bb47-8301f1db271b	Pengguna Google	3	nambahin sedikit karena mulai membaik	2026-04-25 04:00:33

Fig. 3: DANA application reviews collected from the Google Play Store

3.2. Data Preprocessing Results

The preprocessing pipeline was applied sequentially to the 6,280 unique reviews, encompassing: removal of URLs, HTML tags, emojis, symbols, numbers, and punctuation; case folding; non-standard word normalization using the standardized word dictionary; tokenization; stopword removal using the Indonesian NLTK stopword list; and stemming using the Sastrawi library. Figure 4 presents a comparison between the original review text and the output of the cleaning stage, while Figure 5 illustrates the cumulative output of the complete preprocessing pipeline from cleaning through stemming for the first five reviews.

	Review Text	cleaning
0	uang saya hilang begitu sajah saya mohon tolon...	uang saya hilang begitu sajah saya mohon tolon...
1	good	good
2	mantap	mantap
3	nambahin sedikit karena mulai membaik	nambahin sedikit karena mulai membaik
4	👍👍👍	

Fig. 4: Comparison Between Original Review Text And Cleaning Stage Output

	Review Text	cleaning	case_folding	normalisasi	tokenize	stopword removal	steming_data
0	uang saya hilang begitu sajah saya mohon tolon...	uang saya hilang begitu sajah saya mohon tolon...	uang saya hilang begitu sajah saya mohon tolon...	uang saya hilang begitu sajah saya mohon tolon...	[uang, saya, hilang, begitu, sajah, saya, moho...	[uang, hilang, sajah, mohon, tolong, kembalika...	uang hilang sajah mohon tolong kembali uang te...
1	good	good	good	good	[good]	[good]	good
2	mantap	mantap	mantap	mantap	[mantap]	[mantap]	mantap
3	nambahin sedikit karena mulai membaik	nambahin sedikit karena mulai membaik	nambahin sedikit karena mulai membaik	nambahin sedikit karena mulai membaik	[nambahin, sedikit, karena, mulai, membaik]	[nambahin, membaik]	nambahin baik
4	👍👍👍				[]	[]	

Fig. 5: Complete Preprocessing Pipeline Output Cleaning Through Stemming

3.3. Sentiment Label Distribution

Sentiment labeling using the InSet lexicon produced three label classes on a dataset of 6,384 reviews. Figure 6 presents sample labeling results displaying the stemming_data, Score, and Sentiment columns for the first five entries. Based on the sentiment distribution visualization in Figure 7, the Positive class dominates the dataset with 3,724 reviews (58.33%), followed by the Neutral class with 1,727 reviews (27.05%), and the Negative class with 933 reviews (14.61%).

	steming_data	Score	Sentiment
0	uang hilang sajah mohon tolong kembali uang te...	1	Positif
1	good	1	Positif
2	mantap	1	Positif
3	nambahin baik	0	Netral
5	becus urus transfer mending hapus fitur transf...	-7	Negatif

Fig. 6: Sample sentiment labeling results using the InSet lexicon

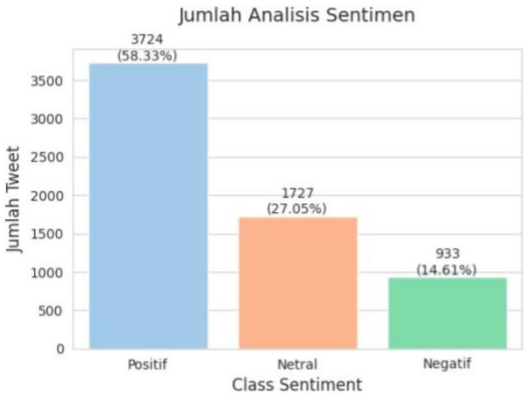


Fig. 7: Sentiment class distribution from InSet Lexicon labeling

3.4. IndoBERT Model Training and Evaluation Results

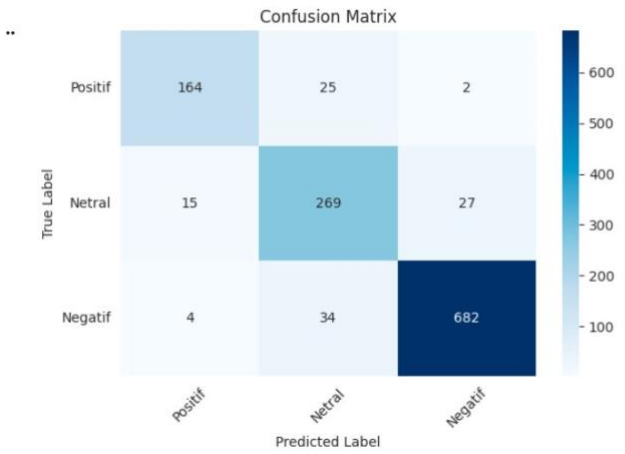
The IndoBERT (indobert-base-p1) model was trained for 3 epochs on 4,884 training samples and evaluated on 6,106 test samples. The model evaluation results, generated using the classification_report function from scikit-learn, are presented in Figure 8. The overall accuracy metric reached 91.24%, demonstrating IndoBERT's capability to effectively classify sentiment in Indonesian-language reviews following the fine-tuning process.

```
•••  Accuracy: 0.9124386252045826

     Classification Report:
                   precision    recall  f1-score   support

          Positif       0.90      0.86      0.88       191
           Netral       0.82      0.86      0.84       311
          Negatif       0.96      0.95      0.95       720

         accuracy                           0.91      1222
        macro avg       0.89      0.89      0.89      1222
     weighted avg       0.91      0.91      0.91      1222
```

Fig. 8: Accuracy output and classification report of the IndoBERT model

3.5. Confusion Matrix Analysis

A confusion matrix analysis was conducted to gain a more detailed understanding of the model's classification error patterns. The confusion matrix is visualized using seaborn.heatmap, as shown in Figure 9, with the x-axis representing predicted labels and the y-axis representing



true labels. Based on these results, of the 191 test samples in the Positive class, the model correctly predicted 164 (85.9%), misclassifying 25 as Neutral and 2 as Negative. Of the 311 test samples in the Neutral class, the model correctly predicted 269 (86.5%), misclassifying 15 as Positive and 27 as Negative. Of the 720 test samples in the Negative class, the model correctly predicted 682 (94.7%), misclassifying 4 as Positive and 34 as Neutral.

Fig. 9: Confusion Matrix From Indobert Model Classification Results

Overall, the evaluation results demonstrate that the IndoBERT model fine-tuned on the DANA application review dataset is capable of learning sentiment patterns from Indonesian-language review text that has undergone a comprehensive preprocessing pipeline. IndoBERT's ability to capture bidirectional contextual information in Indonesian text renders it a well-suited architecture for sentiment classification tasks in the fintech application review domain.

4. Conclusion

This study has successfully implemented and evaluated the IndoBERT (indobert-base-p1) model for sentiment analysis on a dataset of DANA application user reviews obtained from the Google Play Store. Overall, this research yields four primary contributions. First, a comprehensive Indonesian-language text preprocessing pipeline was constructed, encompassing deduplication, cleaning, case folding, non-standard word normalization, tokenization, stopword removal, and stemming, which effectively reduced 10,000 raw data entries to 6,106 high-quality clean samples.

Second, automated sentiment labeling using the InSet lexicon was successfully applied to the entire dataset without the need for manual annotation, producing three label classes: Positive, Neutral, and Negative. Third, the IndoBERT model was successfully fine-tuned under an optimal configuration (3 epochs, learning rate = 2×10^{-5}, batch size = 8) on a training set of 4,884 samples. Fourth, model evaluation on 1,222 test samples yielded an accuracy of 91.24% with a weighted average F1-score of 0.91, demonstrating the model's capability to recognize sentiment in informal Indonesian-language review text.

This study demonstrates that Transformer-based models pre-trained specifically for the Indonesian language, such as IndoBERT, are effective when applied to the domain of fintech application review sentiment analysis. For future research, it is recommended that: (1) manual sentiment annotation be performed to improve training data quality; (2) alternative Indonesian language models such as IndoBERT-lite or RoBERTa-base-Indonesian be evaluated as comparative baselines; (3) class imbalance handling techniques such as oversampling or class weighting be applied to improve performance on minority classes; and (4) the analytical scope be extended to aspect-based sentiment analysis to identify the specific aspects that influence user satisfaction with the DANA application.

References

[1] F. Kurniasari, A. Gunardi, F. Perdana, and A. Firmansyah, “International Journal of Data and Network Science The role of financial technology to increase financial inclusion in Indonesia,” vol. 5, pp. 391–400, 2021, doi: 10.5267/j.ijdns.2021.5.004.

[2] L. J. Ningri, M. Hamidi, and F. Adrianto, “Sentiment Analysis Against Digital Payment ‘ GoPay ’, ‘ OVO ’, ‘ DANA ’ , and ‘ ShopeePay ’ Using Na ï¨ve Bayes Classifier Algorithm,” vol. 3, no. 2, pp. 322–336, 2023.

[3] K. S. Nugroho, A. Y. Sukmadewa, F. A. Bachtiar, and N. Yudistira, “BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews,” pp. 1–10, 2020.

[4] J. Islam, R. Datta, and A. Iqbal, “Actual rating calculation of the zoom cloud meetings app using user reviews on google play store with sentiment annotation of BERT and hybridization of RNN and LSTM,” *Expert Syst. Appl.*, vol. 223, no. June 2022, p. 119919, 2023, doi: 10.1016/j.eswa.2023.119919.

[5] I. Gambo, R. Massenon, R. Oluwaseun, S. Agarwal, and W. Pak, “Heliyon Identifying and resolving conflict in mobile application features through contradictory feedback analysis,” *Heliyon*, vol. 10, no. 17, p. e36729, 2024, doi: 10.1016/j.heliyon.2024.e36729.

[6] M. Wankhade, A. Chandra, S. Rao, and C. Kulkarni, *and challenges*, no. 0123456789. 2022.

[7] F. Koto, “InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs,” pp. 391–394, 2017.

[8] A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, “Improving the accuracy of text classification using stemming method , a case of non - formal Indonesian conversation,” *J. Big Data*, pp. 1–16, 2021, doi: 10.1186/s40537-021-00413-1.

[9] Y. Asri, D. Kuswardani, W. N. Suliyanti, and Y. O. Manullang, “Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application,” vol. 38, no. 1, pp. 677–688, 2025, doi: 10.11591/ijeecs.v38.i1.pp677-688.

[10] A. Vaswani, “Attention Is All You Need,” no. Nips, 2017.

[11] F. Koto and T. Baldwin, “IndoLEM and IndoBERT : A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” pp. 757–770, 2020.

[12] H. P. Wijaya, A. Anugrah, M. Afrina, and A. Ibrahim, “Sentiment Analysis of JMO Application Reviews on the Google Play Store Using BERT,” vol. 8, no. 1, pp. 643–649, 2026, doi: 10.61992/jiem.v8i1.250.

[13] A. Aprinando, P. Simarmata, and T. B. Sasongko, “Sentiment Analysis on BRImo Application Reviews Using IndoBERT,” vol. 9, no. 3, 2025.

[14] A. F. Aji *et al.*, “One Country , 700 + Languages : NLP Challenges for Underrepresented Languages and Dialects in Indonesia”.

[15] M. C. Kenton, L. Kristina, and J. Devlin, “BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding,” pp. 4171–4186, 2019.