IJASST
International Journal of
Applied Sciences and Smart Technologies
Volume 07
Issue 02
p-ISSN 2655-8564, e-ISSN 2685-943, pp. 371-384, 2025

Fine-Tuned IndoBERT-Based Sentiment Analysis for Old Indonesian Songs Using Contextual and Generating Augmentation

Gilang Ramdhani[1], Siti Yuliyanti [1*]

[1]*Department of Informatics, Faculty of Engineering, Siliwangi University, Tasikmalaya, West Java, Indonesia*
[*]*Corresponding Author: sitiyuliyanti@unsil.id*

(Received 05-06-2025; Revised 10-07-2025; Accepted 26-07-2025)

Abstract

This study examines sentiment analysis of traditional Indonesian songs using a fine-tuned IndoBERT model, which has been enhanced through the incorporation of contextual and textual data augmentation. The dataset consists of user comments related to classic Indonesian songs, categorized into positive, negative, and neutral sentiments. Two augmentation strategies were applied: textual augmentation using text generation techniques and contextual augmentation leveraging semantic similarity. Evaluation was conducted using accuracy, precision, recall, and F1-score metrics. Results show that the model trained on the original dataset achieved balanced and stable performance (accuracy: 0.86). Textual augmentation, despite generating high data variation, reduced model accuracy (0.63) and introduced a bias toward negative sentiment. In contrast, contextual augmentation-maintained performance stability and even slightly improved precision (0.87). These findings indicate that contextual augmentation is more effective for enriching sentiment datasets without compromising model performance. The findings highlight the effectiveness of integrating pre-trained language models and data augmentation strategies for sentiment analysis in low-resource.

Keywords: IndoBERT, Sentiment Analysis, Old Indonesian Songs, Contextual Augmentation, Textual Augmentation.

1 Introduction

Sentiment analysis is a widely applied natural language processing (NLP) technique that seeks to identify and classify emotional tones expressed in text[1], [2]. In recent years, the increasing availability of user-generated content, such as comments on music platforms and social media, has opened new opportunities to analyze public perception toward songs, including those from past decades[3], [4]. Old Indonesian songs,

such as "Tuhan yang Aneh – Apa Elo Tega", often carry deep emotional narratives and have regained popularity in online spaces. However, the informal, poetic, and sometimes metaphorical nature of song-related commentary poses challenges for traditional sentiment analysis methods. The criteria for what we consider “old Indonesian songs” were implicitly based on temporal and stylistic attributes, but we acknowledge that this was not explicitly stated in the background section. In our study, *old Indonesian songs* are defined as songs that were released before the 1990s and are recognized as part of the classic or nostalgic repertoire of Indonesian music, typically characterized by lyrical richness, poetic language, and traditional musical arrangements. These songs are often still referenced in modern media and evoke a certain cultural sentiment tied to past eras. In future revisions, we will clarify this definition in the introduction to provide better contextual grounding.

Pretrained language models like IndoBERT, developed specifically for the Indonesian language, have shown great promise in capturing the linguistic nuances required for high-quality text classification[5], [6]. Yet, the performance of such models depends heavily on the diversity and richness of training data. To address this, data augmentation techniques can be employed to enhance model generalization [7], [8]. This study explores two augmentation strategies—textual augmentation using synonym and paraphrase generation, and contextual augmentation leveraging semantic embeddings—to fine-tune IndoBERT for sentiment analysis tasks related to "Tuhan yang Aneh – Apa Elo Tega".

Despite the growing interest in analyzing public sentiment toward music, especially emotionally intense or controversial songs like *"Tuhan yang Aneh – Apa Elo Tega"*, sentiment analysis in the Indonesian language remains underexplored. Most existing models are trained on general-purpose datasets and often fail to capture the informal, figurative, or context-dependent expressions commonly found in song-related commentary. These limitations reduce model accuracy and can lead to biased or misleading sentiment predictions.

This is an important point, and we appreciate the critique. The reason for focusing on a single song in the experimental stage was to create a controlled case study to assess the effectiveness of IndoBERT when fine-tuned with contextual and generative data

augmentation methods. While the title implies a broader scope, we acknowledge that the data was limited to one song due to constraints in labeled sentiment datasets for old Indonesian lyrics. We do not claim that one song can fully represent all old Indonesian songs, but rather that this study serves as a proof of concept. Future work will expand to include a larger and more diverse corpus of old Indonesian songs to validate the generalizability of the findings. We will revise the title and discussion to reflect this scope limitation better.

Furthermore, high-quality annotated datasets for Indonesian sentiment analysis, particularly in the context of music and culture, are limited in both size and variety[9]. This scarcity of training data can limit the ability of even powerful models, such as IndoBERT, to generalize across diverse expressions and sentiment patterns. While data augmentation techniques offer a potential solution, not all augmentation methods are equally effective[10], [11]. Textual augmentation using random synonym replacement or text generation may introduce noise or distort sentiment polarity, while contextual augmentation strategies remain underutilized in Indonesian-language NLP tasks [12], [13].

The goal of this research is to evaluate the effectiveness of contextual and textual data augmentation in improving sentiment classification performance, while also understanding how public sentiment toward emotionally charged old songs is represented online. The findings are expected to contribute to better sentiment analysis practices in the domain of cultural and musical content, particularly in low-resource language settings like Bahasa Indonesia.

2 Material and Methods

The stages of this research include data collection, pre-processing, data sharing, IndoBERT Modeling, Fine-Tuning using Contextual and Textual Augmentation, and model evaluation as shown in Fig. 1.

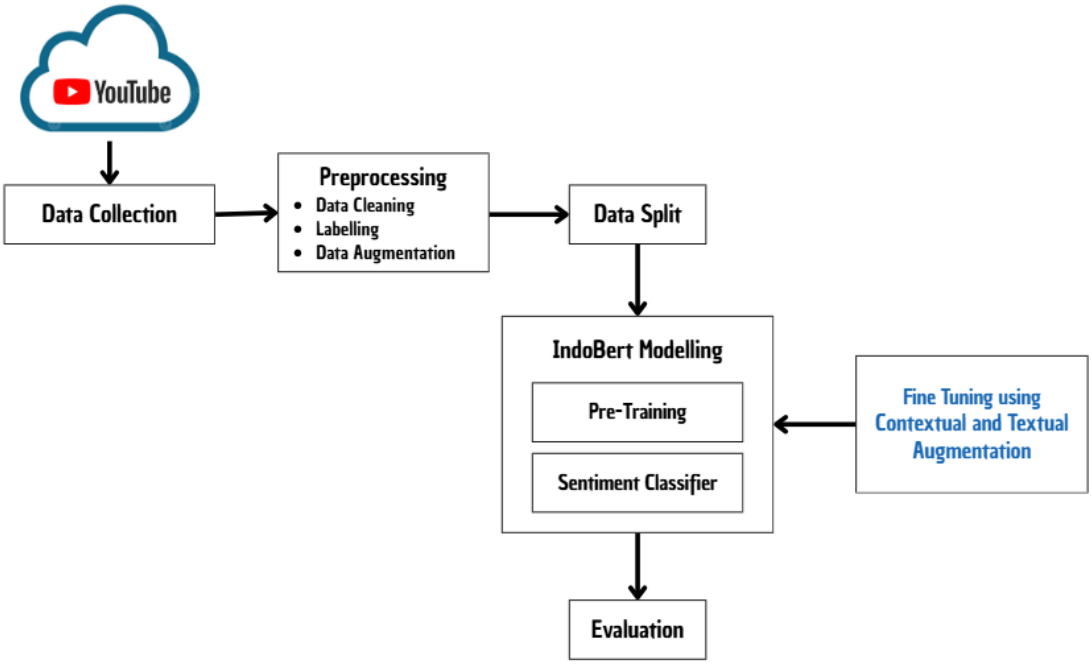

Figure 1. Research Framework

Data Collection

Comment data is gathered from the YouTube video for the song "Tuhan yang Aneh – Apa Elo Tega. " This data consists of public comments extracted via web scraping and saved in CSV format. These comments represent users' genuine thoughts and are divided into three categories: positive, negative, and neutral [8], [14], [15]. Alongside the original dataset, this project incorporates additional data to enhance and balance the collection. The comments considered for this project are sourced from the YouTube video titled "Tuhan Yang Aneh - Apa Elo Tega (Old Song 2012)," which was re-uploaded by the Pepstun YouTube channel. The video has a total of 13,509 comments, of which 4,683 were successfully gathered, excluding replies to individual comments.

Preprocessing

After the data scraping process, data cleaning is carried out. These steps include: Removing links, usernames starting with the symbol “@” (@username) or hashtags, and numeric characters[16], [17]. In addition, symbols and unnecessary punctuation are also

removed to produce cleaner data that is ready to be labeled. Next comes the stage of labeling, in which remarks are sorted by hand or partially automatically into sentiment groups (positive, negative, neutral) [3], [18], [19]. Following that is the phase of data enhancement, which involves two forms of augmentation: textual augmentation that includes synonyms or paraphrases, and contextual augmentation utilizing embedding methods to preserve the underlying meaning.

Data Split

The data is divided into 70% training data to train the model, and 30% test data to measure the performance of the model [20], [21], [22], and the model is trained with 3 epochs.

IndoBERT using Contextual and Textual Augmentation

The model used is indobenchmark/indobert-large-p1, which is a large version of Indobert provided by IndoNLU and accessed through the Hugging Face Transformers Library[11]. Model training is carried out using a fine-tuning approach, namely retraining the pre-trained IndoBERT model on the YouTube comment dataset that has gone through the preprocessing and labeling process. The fine-tuning process is carried out separately for three data scenarios, namely:

1. Original data (without augmentation),
2. Data with Text Generating augmentation,
3. Data with Contextual Augmentation augmentation.

Each scenario uses 70% of the total dataset for training data, and the model is trained with 3 epochs. This training uses the Google Colab Platform which uses the Tesla T4 GPU.

3 Results and Discussions

IndoBERT's effectiveness was evaluated across three different dataset scenarios: the original data (without any modifications), data enhanced with text generation, and data that has been supplemented with contextual augmentation. The evaluation criteria include accuracy, precision, recall, and f1-score, with a confusion matrix providing a

breakdown of predictions for each category. The model excels in identifying the positive category (TP: 2039), with only 102 and 24 instances misidentified as neutral and negative, respectively. In terms of the neutral category, the model still performs well (TP: 1603), although it misclassifies 367 instances as positive, suggesting an inclination to overestimate positive sentiments. Regarding the negative category, the model demonstrates strong performance with TP: 382, and only a small number are incorrectly categorized as positive or neutral.

The model performs best in identifying **positive** sentiment, with high accuracy and relatively few misclassifications. It is also fairly accurate with **neutral** sentiment, though it often confuses it with positive, as shown in Fig. 2. The **negative** class has the lowest count and highest misclassification rate, suggesting the model struggles more with identifying negative sentiment correctly. The color intensity in the heatmap helps visually represent the number of occurrences, with darker shades indicating higher values.

The model shows stable and balanced performance without augmentation. The results are evenly distributed across the three classes, with a slight tendency towards the positive class, as shown in Fig. 3.

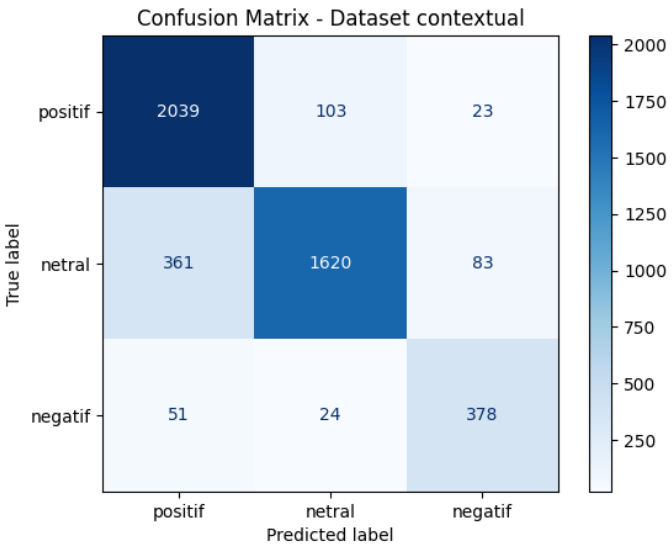

Figure 2. Confusion matrix contextual dataset

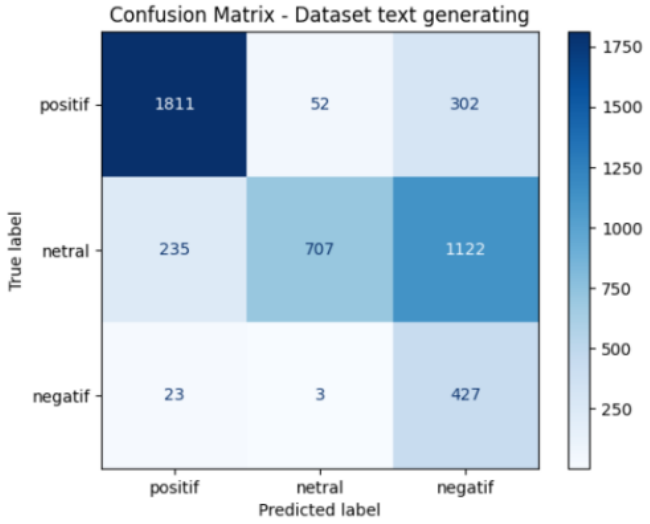

Figure 3. Confusion matrix text generating

Fig. 3 shows that the positive class is still classified dominantly (TP: 1811), but there is a spike in errors to negative by 302 comments that should be positive. The neutral class is significantly confused, with only 707 correct (TP), while 1122 are misclassified as negative. This shows the weakness of the model in recognizing neutrality when the data comes from text-generating augmentation. The negative class remains strong with TP: 427, showing the model's resilience to negative comments even after augmentation. Although the precision and F1-score values are high, the accuracy drops drastically. This indicates that text-generating augmentation increases the variation of the data but makes it difficult for the model to accurately distinguish neutral and negative comments.

The prediction distribution shown in Fig. 4 closely resembles the initial dataset. The model effectively identifies positive remarks, achieving 2039 true positives, while only misclassifying 103 as neutral and 23 as negative. Regarding neutral remarks, there is an enhancement in performance when compared to text generation. The misclassification numbers are lower, with 361 mistakenly labeled as positive and 83 as negative, resulting in 1620 true positives. The negative category is also fairly well-balanced, with 378 true positives, suggesting that contextual enrichment does not negatively impact the model's efficiency.

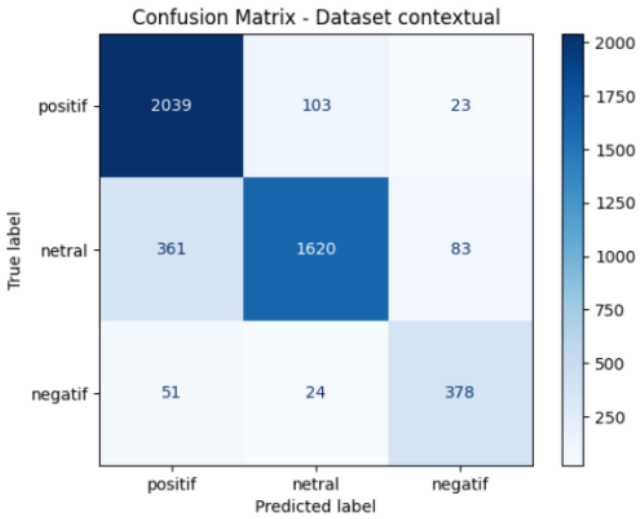

Figure 4. Confusion matrix text contextual

Assessment of model effectiveness following fine-tuning is presented in Table 1, which utilizes accuracy, precision, recall, and F1-score metrics, along with an evaluation of the confusion matrix across the three dataset scenarios. It can be inferred that the dataset enhanced with contextual augmentation yields the highest and most even performance.

Although the original dataset (without any enhancements) demonstrated strong results (both accuracy and F1-score of 0.86), this method faces constraints due to the limited quantity and diversity of the initial data. Conversely, contextual augmentation effectively enhanced the dataset without compromising the stability of the model, yielding evaluation outcomes that matched those of the original dataset (accuracy: 0.86, F1-score: 0.86), while maintaining a balanced classification

Table 1. Performance comparison table with augmentation

Dataset	Accuray	Precission	Recal	F1-Score	Description
Initial Dataset	0.86	0.86	0.86	0.86	Stable, no augmentation, balanced prediction
Text Generating	0.63	0.84	0.83	0.85	High variation but lower accuracy, biased to negative
Text Contextual	0.86	0.87	0.86	0.86	Variation maintained, performance remains stable

across sentiment categories.

In contrast, augmentation through text generation achieved a relatively high F1-score (0.85), but accuracy dropped significantly to 0. 63. This suggests that the model struggles to differentiate between neutral and negative feedback, a challenge that is evident in the confusion matrix. This decline might stem from the generated text being less reflective of the original comments' structure and context.

Therefore, employing contextual augmentation techniques appears to be the most effective method for this project, as it enhances data variety without diminishing the quality of the model's classifications. This strategy is ideal for expanding the dataset while preserving both the accuracy and the generalization capabilities of the IndoBERT model.

4 Conclusions

This study proposed a sentiment analysis approach for old Indonesian songs using a fine-tuned IndoBERT model, incorporating both contextual and generative data augmentation techniques. Based on the evaluation results, the model trained with generative augmentation outperformed the one using contextual augmentation, achieving an accuracy of 87%, compared to the contextual model’s performance. This indicates that generative augmentation provides more diverse and effective training data, which enhances the model's ability to generalize and classify sentiments more accurately in this specific domain. For future work, the following recommendations are proposed: incorporate emotion classification, further improve accuracy, develop a web-based interactive dashboard, and apply to other social media platforms.

Acknowledgements

The author would like to express his deepest gratitude to all parties who have assisted in this research.

References

[1] S. Yuliyanti and Rizky, “Implementasi Algoritma Rabin Karp Untuk Mendeteksi Kemiripan Dokumen Stmik Bandung,” *J. Bangkit Indones.*, vol. 10, no. 02, p. 1, 2020, doi: 10.52771/bangkitindonesia.v10i02.124.

[2] S. Yuliyanti, E. Nur Fitriani Dewi, and A. Nur Rachman, “Optimasi Rabin Karp dengan Rolling Hash dan k-Gram pada Similarity Check Dokumen Abstrak Jurnal,” vol. 12, no. 1, 2024, doi: 10.26418/justin.v12i1.71224.

[3] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, “Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN,” *Ilk. J. Ilm.*, vol. 14, no. 3, pp. 348–354, 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[4] G. Hakim, T. N. Fatyanosa, and A. W. Widodo, “Analisis Sentimen Masyarakat terhadap Kereta Cepat Whoosh pada Platform X menggunakan IndoBERT,” vol. 1, no. 1, pp. 1–10, 2023.

[5] K. Ayu Pradani, L. Hulliyyatus Suadaa, and P. Korespondensi, “Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai Automated Essay Scoring Using Transformer-Based Semantic Textual Similarity for Essay Assessment,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1177–1184, 2019, doi: 10.25126/jtiik.2023107338.

[6] T. Wahyuningsih *et al.*, “Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice’s Coefficient,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 2, pp. 343–348, 2021, doi: 10.17762/turcomat.v12i3.938.

[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[8] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, "Fine-Tuned IndoBERT based model and data augmentation for indonesian language paraphrase identification," *Rev. d'Intelligence Artif.*, vol. 37, no. 3, pp. 733–743, 2023, doi: 10.18280/ria.370322.

[9] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile, "AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets," *CEUR Workshop Proc.*, vol. 2481, 2019.

[10] R. Silva Barbon and A. T. Akabane, "Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study," *Sensors*, vol. 22, no. 21, 2022, doi: 10.3390/s22218184.

[11] M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," *Sci. Rep.*, vol. 14, no. 1, p. 24882, 2024, doi: 10.1038/s41598-024-76079-5.

[12] T. Bey Kusuma, I. Komang, and A. Mogi, "Implementasi BERT pada Analisis Sentimen Ulasan Destinasi Wisata Bali," *J. Elektron. Ilmu Komput. Udayana*, vol. 12, no. 2, pp. 409–420, 2023.

[13] Y. A. Singgalen, "Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data," vol. 6, no. 2, pp. 976–986, 2025, doi: 10.47065/josh.v6i2.6505.

[14] I. A. Oktariansyah, F. R. Umbara, and F. Kasyidi, "Klasifikasi Sentimen Untuk Mengetahui Kecenderungan Politik Pengguna X Pada Calon Presiden Indonesia 2024 Menggunakan Metode IndoBert," *Build. Informatics, Technol. Sci.*, vol. 6, no. 2, pp. 636–648, 2024, [Online]. Available: https://ejurnal.seminar-id.com/index.php/bits/article/view/5435

[15] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11856 LNAI, no. 2, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.

[16] S. Redhu, "Sentiment Analysis Using Text Mining: A Review," *Int. J. Data Sci. Technol.*, vol. 4, no. 2, 2018, doi: 10.11648/j.ijdst.20180402.12.

[17] S. Yuliyanti, T. Djatna, and H. Sukoco, "Sentiment mining of community development program evaluation based on social media," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 4, pp. 1858–1864, 2017, doi: 10.12928/TELKOMNIKA.v15i4.4633.

[18] M. G. Villar, J. B. Ballester, I. De, T. Diez, and I. Ashraf, "Analyzing Sentiments Regarding ChatGPT Using Novel BERT : A Machine Learning Approach," pp. 1–29, 2023.

[19] D. H. Yuliyanti, Siti;Ula, "Essay Answer Detection System Uses Cosine Similarity and Similarity Scoring in Sentences," vol. 06, no. 02, pp. 337–348, 2024.

[20] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 1, p. 79, 2023, doi: 10.22146/ijccs.77354.

[21] M. Chiny, M. Chihab, and Y. Chihab, “LSTM , VADER and TF-IDF based Hybrid Sentiment Analysis Model,” vol. 12, no. 7, pp. 265–275, 2021.

[22] J. Asian, M. Dholah Rosita, and T. Mantoro, “Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods,” *J. Online Inform.*, vol. 7, no. 1, p. 132, 2022, doi: 10.15575/join.v7i1.900.

This page intentionally left