Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

Random Forest, LSTM, and IndoBERT Comparison for TikTok App Sentiment Analysis

Imam Saputra[1], Mesran[2,*], Ruziana Mohamad Rasli[3]

[1] Digital Business Study Program, Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia
[2] Management Study Program, Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia
[3] School of Computing, Universiti Utara Malaysia, Kedah, Malaysia
Email: [1]saputraimam69@gmail.com, [2,*]mesran.skom.mkom@gmail.com, [3]ruziana@uum.edu.my
Correspondence Author Email: mesran.skom.mkom@gmail.com
Submitted: **21/04/2026**; Accepted: **30/06/2026**; Published**: 30/06/2026**

Abstract−The rapid growth of social media platforms like TikTok has generated a massive volume of user reviews on the Google Play Store, serving as a critical indicator of application service quality. However, the unstructured nature of Indonesian social media text and the significant imbalance between sentiment classes pose substantial challenges for automated classification systems. Addressing this class imbalance is highly crucial for application developers, as critical negative and neutral feedback containing essential feature complaints is easily marginalized by the overwhelming majority of positive reviews, leading to biased operational insights. This research conducted a comprehensive comparative study of three distinct computational paradigms: Random Forest, Long Short-Term Memory (LSTM), and IndoBERT, to identify the most effective model for sentiment analysis. A dataset of 5,000 TikTok reviews was meticulously processed using a negation-aware preprocessing pipeline to preserve semantic integrity. To address class imbalance, architecture-specific techniques were deployed, including SMOTE for Random Forest, Class Weighting for LSTM, and Random OverSampling for IndoBERT. The experimental results demonstrate that IndoBERT significantly outperforms other models, achieving the highest global accuracy of 81% and a Macro F1-Score of 0.56. While Random Forest and LSTM yielded lower accuracies of 75% and 71%, respectively, they exhibited stability in predicting the majority class but struggled with the inherent ambiguity of neutral sentiments. The study concludes that IndoBERT's bidirectional self-attention mechanism provides superior contextual understanding of Indonesian slang and non-formal syntax. This research contributes a robust framework for application developers to monitor public opinion objectively. Furthermore, the findings highlight that despite advanced balancing techniques, the "neutrality bottleneck" remains a challenge, suggesting that future research should explore aspect-based sentiment analysis to enhance classification granularity in the Indonesian NLP domain.

Keywords: Indonesian NLP; Sentiment Analysis; Class Imbalance; Random Forest; Long Short-Term Memory; IndoBERT; TikTok

1. INTRODUCTION

The rapid evolution of information and communication technology has significantly transformed the global paradigm of social interaction through the emergence of highly dynamic social media platforms. TikTok, as one of the most popular short-video sharing platforms, has achieved extraordinary user growth in Indonesia over the past five years [1]. The presence of this platform functions not only as a medium of entertainment but also as a primary channel for users to provide direct feedback regarding application service quality through reviews on the Google Play Store [2]. These reviews represent valuable data assets that reflect user satisfaction levels, complaints, and expectations in real-time [3]. However, the massive volume of data and the exponential daily growth of reviews render manual analysis practically impossible to conduct [4]. Consequently, an automated computational system is required to extract sentiment information effectively to support data-driven decision-making for developers [5]. Sentiment analysis emerges as a technical solution to identify public opinions scattered across thousands of these reviews [6]. By understanding user sentiment patterns, developers can prioritize feature improvements that are most frequently criticized by the public [7]. The effectiveness of this sentiment analysis depends heavily on selecting the appropriate algorithm capable of handling the complexities of the Indonesian language [8].

Text reviews on the TikTok application in Indonesia possess unique characteristics that distinguish them from formal text found in news media or official documents [9]. Users tend to employ non-formal language, slang, non-standard abbreviations, and excessive punctuation to express their emotions [10]. Furthermore, the phenomenon of elongated words, such as "baguuuus" or "cepaaaat," is frequently encountered in social media review corpora. The use of regional languages mixed with Indonesian further increases the level of difficulty in the text normalization process [11]. This irregular sentence structure often causes traditional classification models to fail in capturing the true semantic meaning of an opinion [12]. In-depth preprocessing, ranging from case folding to Indonesian-specific stemming, becomes a crucial stage that cannot be overlooked [13]. The success of a sentiment analysis model is largely determined by how well the data cleaning stage is performed to reduce linguistic noise [14]. These challenges demand the use of feature extraction techniques and classification algorithms that can adapt to ever-changing language dynamics [15].

A major issue in sentiment analysis research is the imbalance in the distribution of data classes, commonly known as imbalanced data [16]. In the context of TikTok application reviews, the number of positive reviews often significantly dominates compared to negative or neutral reviews. This imbalance causes classification models to lean toward the majority class, resulting in very low predictive performance for the minority class [17]. In fact, reviews with neutral or negative sentiments often contain more critical information essential for application system

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

improvements [18]. Previous studies have attempted various techniques to address this issue, such as the use of the Synthetic Minority Over-sampling Technique (SMOTE) at the feature level. Additionally, class weighting techniques are often applied to deep learning models to impose higher penalties for minority class misclassifications. Another popular approach is Random OverSampling (ROS), which duplicates original text data to preserve linguistic context [19]. Experiments involving various data balancing techniques are necessary to identify the best scheme for improving overall accuracy [20].

The Random Forest algorithm has long been a primary choice in sentiment analysis research due to its ability to handle high-dimensional data [21]. As an ensemble learning method based on a combination of many decision trees, Random Forest possesses high robustness against overfitting [22]. Many researchers utilize Random Forest combined with TF-IDF feature extraction to produce stable performance on medium-sized datasets [23]. The advantage of this algorithm lies in its ability to rank features based on their importance to sentiment classification [24]. Nevertheless, Random Forest often loses information regarding word order within a sentence because of its bag-of-words nature [25]. This represents a major weakness when dealing with sentences whose meanings are highly dependent on structural word order. However, the ease of parameter tuning ensures that Random Forest remains a highly relevant baseline in comparative studies [26]. The testing of Random Forest in this study aims to measure the extent to which traditional algorithms can compete with more modern neural architectures.

Along with advancements in artificial intelligence, Long Short-Term Memory (LSTM) methods have begun to be widely adopted for natural language processing tasks [27]. LSTM is a variant of the Recurrent Neural Network (RNN) specifically designed to overcome the vanishing gradient problem in long data sequences [28]. This algorithm is capable of remembering context from previous words through a gating mechanism that regulates the flow of information in long-term memory [29]. In sentiment analysis, the ability to understand relationships between words in a sequence of sentences is crucial [30]. LSTM works by representing words into embedding vectors that capture the semantic similarity between words [31]. Several studies indicate that LSTM excels in identifying sentiment in sentences with long-distance dependencies. However, training LSTM models requires greater computational power and longer training times compared to conventional models [32]. Moreover, LSTM requires a vast amount of data to achieve an optimal level of generalization [33]. This research evaluates whether the sequential memory capability of LSTM can provide an accuracy increase for short TikTok reviews.

The latest breakthrough in the field of Natural Language Processing (NLP) is marked by the emergence of the highly revolutionary Transformer architecture [34]. The IndoBERT model, specifically developed for the Indonesian language, is an adaptation of the BERT (Bidirectional Encoder Representations from Transformers) architecture [35]. Unlike LSTM, which reads text in one direction, IndoBERT understands the context of a word from both directions simultaneously [36]. This model has been pre-trained using a massive Indonesian corpus, covering millions of sentences from various digital sources [37]. Fine-tuning techniques on IndoBERT allow the model to adapt quickly to specific domains, such as application reviews, with high precision [38]. The self-attention mechanism in IndoBERT enables the model to focus on the most relevant words in determining the final sentiment [39]. IndoBERT is currently considered the State-of-the-Art standard for various text classification tasks in Indonesia. The use of IndoBERT in this study is expected to provide a significant accuracy leap compared to other methods.

Although these three models have been widely studied separately, existing literature reveals a critical research gap regarding their direct head-to-head comparison on dynamic Indonesian social media datasets, particularly TikTok. Previous studies predominantly focus on either comparing only two architectural paradigms (e.g., traditional machine learning vs. deep learning) or applying a single, uniform data balancing technique across all models. To the best of our knowledge, there is currently no study that comprehensively evaluates and compares the performance of Random Forest, LSTM, and IndoBERT on an Indonesian TikTok dataset by deploying architecture-specific balancing schemes simultaneously. Most existing works fail to measure how statistical feature-level balancing (SMOTE), cost-sensitive loss adjustments (Class Weighting), and text-level data replication (ROS) behave when integrated with their respective matching computational paradigms within a single controlled experiment. Consequently, empirical evidence remains absent regarding which specific combination of data balancing and model architecture is most resilient against the 'neutrality bottleneck' and severe class skewness in Indonesian slang text. This study explicitly addresses this gap by conducting a multi-scenario benchmark to provide a measurable and definitive standard for Indonesian NLP application reviews.

The primary objective of this study is to measure and compare the performance of Random Forest, LSTM, and IndoBERT in TikTok review sentiment analysis. The specific goal is to identify the most optimal combination of data balancing techniques and algorithms for handling imbalanced data. The main contributions of this research include the development of a preprocessing pipeline optimized for TikTok slang and accurate negation protection. Additionally, the experimental results provide empirical evidence regarding the superiority of the Transformer model (IndoBERT) in achieving the highest accuracy. This study also provides an in-depth analysis of the confusion matrix to understand the behavior of the models in predicting each sentiment class. The final results of this study are expected to serve as a scientific reference for researchers in the field of NLP when selecting Indonesian text classification models. For application developers, these findings can be used as a tool for automated and objective service quality monitoring. Thus, this research not only contributes theoretically to the field of artificial intelligence but also has real practical implications for the technology industry.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

2. RESEARCH METHODOLOGY

2.1 General Research Framework

The methodology of this research is structured into a systematic pipeline designed to evaluate the performance of three distinct paradigms in Natural Language Processing (NLP): statistical ensemble learning, sequential deep learning, and transformer-based architectures.

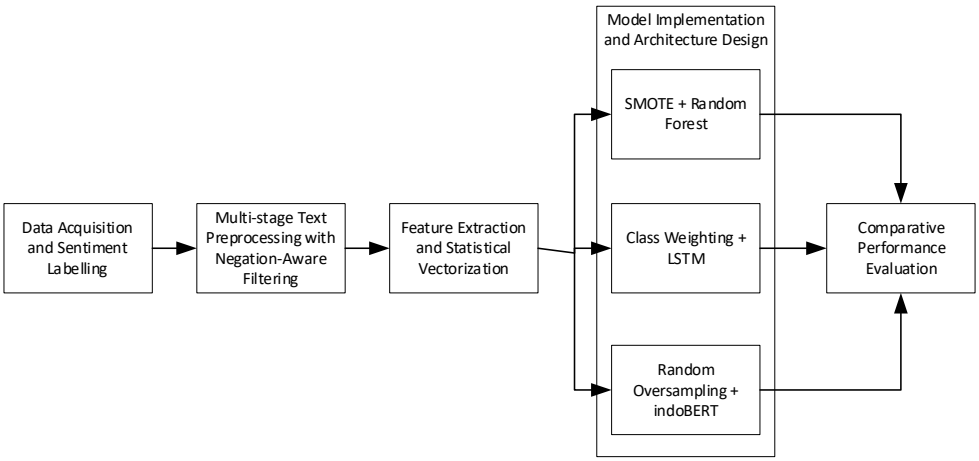

Figure 1. Research Framework

The research follows a rigorous experimental design comprising five core phases: (1) Data Acquisition and Sentiment Labeling, (2) Multi-stage Text Preprocessing with Negation-Aware Filtering, (3) Feature Extraction and Statistical Vectorization, (4) Model Implementation and Architecture Design, and (5) Comparative Performance Evaluation using Robust Metrics. Each phase is implemented to ensure high reproducibility, allowing for the verification of results across different social media datasets. The key innovation in this methodology lies in the strategic application of architecture-specific data-balancing techniques, where SMOTE is integrated with Random Forest, Class Weighting with LSTM, and Random OverSampling with IndoBERT fine-tuning.

2.2 Data Acquisition and Sentiment Distribution

The primary data source for this study consists of user-generated content in the form of textual reviews from the TikTok application on the Google Play Store. A total of 5,000 raw reviews were extracted using the google-play-scraper Python library. This acquisition process focused on metadata including the textual review, the user rating (1–5 stars), and the timestamp. To ensure the relevance of the data to the current application state, the scraping parameters targeted the most recent reviews available during the data collection period.

Table 1. Data Labelling

	Positive	Negative	Neutral
Amount	3628	1100	272

The labeling (Table 1) process utilized a heuristic mapping of user ratings into categorical sentiment labels. Ratings of 1 and 2 were classified as Negative, 3 as Neutral, and 4 and 5 as Positive. An initial exploratory data analysis (EDA) revealed a significant skewness in the class distribution, with the Positive class occupying approximately 70% of the dataset, while the Neutral class represented less than 10%. This high degree of class imbalance serves as the primary challenge for the classification models, as traditional algorithms tend to converge toward the majority class, leading to poor sensitivity on critical negative and neutral feedback.

2.3 Systematic Text Preprocessing Pipeline

Given the noisy nature of TikTok reviews which are characterized by slang, grammatical errors, and excessive punctuation a comprehensive preprocessing pipeline was developed. This stage is critical for reducing the dimensionality of the feature space and improving the signal-to-noise ratio within the corpus.

a. Case Folding and Text Cleaning
   All characters were converted to lowercase, and URLs, HTML tags, digits, and special characters were stripped using Regular Expressions. Scientific Reason: This step reduces dimensionality by consolidating identical terms (e.g., "TikTok" and "tiktok") and eliminates non-semantic noise that introduces variance without predictive value.
b. Lexical Normalization and Slang Mapping
   A custom dictionary mapped Indonesian slang and typos to their formal forms (e.g., "bgt" to "banget"). Scientific Reason: Social media text features a massive vocabulary drift; normalization maps irregular tokens into unified semantic bins, drastically reducing Out-of-Vocabulary (OOV) tokens for TF-IDF and LSTM models.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

c. Tokenization and Negation-Aware Stopword Removal
ext was tokenized, and standard Indonesian stopword lists were heavily modified to exclude negation words like "tidak", "bukan", and "kurang". Scientific Reason: Standard stopword removal is detrimental to sentiment classification because eliminating negations completely inverts sentence polarity (e.g., transforming "tidak puas" into "puas"). Protecting these tokens preserves core semantic signals.

d. Morphological Stemming
The Sastrawi library stripped Indonesian affixes to extract root words (e.g., "mengirimkan" into "kirim"). Scientific Reason: Stemming addresses morphological richness by clustering words with identical semantic roots, which minimizes data sparsity and optimizes the statistical weight of features.

2.4 Model Architectures and Balancing Strategies

a. Random Forest and TF-IDF with SMOTE
For the baseline machine learning model, the corpus was transformed into a numerical matrix using Term Frequency-Inverse Document Frequency (TF-IDF). The TF-IDF weight for a term t in document d is calculated as:

$$W_{t,d} = tf_{t,d} \, x \, log \left(\frac{N}{df_t}\right) \quad (1)$$

Where $tf_{t,d}$ is the frequency of term t in document d, N is the total number of documents, and df_t is the number of documents containing term t. To counter the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set. SMOTE generates synthetic examples for the Neutral and Negative classes by selecting a minority sample and creating new points along the line segments joining any/all of the k nearest neighbors. The Random Forest classifier was configured with 100 decision trees using the Gini impurity criterion.

b. LSTM with Sequential Embedding and Class Weights
The second model paradigm utilizes Long Short-Term Memory (LSTM) to capture sequential dependencies. The preprocessed text was converted into integer sequences and padded to a fixed length of 100 tokens. The model architecture is as follows:

Table 2. LSTM Model Architecture

Model	Properties
Embedding Layer	Maps each token to a 128-dimensional dense vector space
SpatialDropout1D	Applied at a rate of 0.3 to prevent co-adaptation of features
LSTM Layer	Comprising 100 hidden units with a dropout and recurrent dropout rate of 0.3 to mitigate overfitting.
Fully Connected Layer	A 64-unit Dense layer with ReLU activation.
Output Layer	A 3-unit Dense layer with Softmax activation for probability distribution across the three classes.

Instead of SMOTE, this model utilized Class Weighting. Weights were assigned to each class inversely proportional to their frequency in the training data using the following formula:

$$Weight_c = \frac{Total\_Samples}{Number\_of\_Classes \; x \; Samples\_in\_Class_c} \quad (2)$$

This approach forces the optimizer to prioritize the minority class by increasing the loss penalty for misclassifying Neutral or Negative samples.

c. IndoBERT Fine-Tuning with Random OverSampling
Before tokenization, the training data was balanced using Random OverSampling (ROS) at the textual level. The scientific rationale for selecting ROS over SMOTE for IndoBERT is fundamentally rooted in the architectural requirements of Transformers. SMOTE operates by generating synthetic vectors in a continuous feature space (e.g., after TF-IDF extraction), which completely destroys the sequential, syntactic, and structural integrity of natural language sequences. Since IndoBERT relies heavily on its bidirectional self-attention mechanism to capture contextual dependencies between real tokens, inputting artificial, non-existent vector patterns would severely degrade the model's pre-trained weights and attention maps. Thus, duplicating actual sentences via ROS preserves the linguistic syntax necessary for deep learning. The impact of these architecture-specific balancing strategies shifted the heavily skewed training partition into perfect statistical equilibrium. While the original dataset was highly imbalanced comprising approximately 72.6% Positive, 22.0% Negative, and 5.4% Neutral samples the post-balancing application enforced a strict uniform distribution across all models' training pipelines, where each sentiment class was adjusted to represent exactly 33.3% of the training input space. This total equilibrium forces the loss functions to penalize minority misclassifications as severely as majority ones. The fine-tuning process involved the following hyperparameters (Table 3).

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

Table 3. IndoBERT Model Architecture

Model	Properties
Batch Size	16
Epoch	3 (to prevent catastrophic forgetting)
Learning Rate Scheduler	Warmup steps of 500 with weight decay of 0.01.
Optimizer	AdamW (Adam with Weight Decay)
Max Sequence Lenght	128 tokens

2.5 Experimental Evaluation and Performance Metrics

The dataset was split using an 80:20 ratio for training and testing. It is crucial to note that the balancing techniques (SMOTE, ROS) were applied exclusively to the training set. The test set remained in its original imbalanced state to ensure that the evaluation accurately reflects the model's performance in a real-world, skewed environment.

The evaluation utilizes four primary metrics derived from the confusion matrix: Accuracy, Precision, Recall, and F1-Score. While Accuracy provides a general overview, this study emphasizes the Macro-averaged F1-Score, as it treats all classes equally regardless of their support size. This metric is the most reliable indicator of how well the models handle the minority Neutral class. The evaluation was conducted in a Python 3.12 environment using TensorFlow 2.19, PyTorch, and the Hugging Face Transformers library.

3. RESULT AND DISCUSSION

This section presents a comprehensive analysis of the experimental results obtained from the sentiment classification of TikTok reviews using three distinct computational paradigms. The performance of Random Forest, LSTM, and IndoBERT is evaluated through a rigorous comparative lens, focusing on their ability to generalize across imbalanced sentiment classes. Furthermore, the discussion synthesizes these findings with existing literature to provide a deeper understanding of how architectural differences and data-balancing strategies influence the accuracy and robustness of Indonesian sentiment analysis in a social media context.

3.1 Comparative Analysis of Model Performance

The experimental evaluation of the three distinct paradigms statistical ensemble (Random Forest), sequential deep learning (LSTM), and transformer-based (IndoBERT) reveals a significant variance in their ability to interpret the linguistic nuances of TikTok reviews. The results, summarized in Table 4, indicate that while all models demonstrate a baseline competency in sentiment classification, the architectural differences fundamentally dictate their precision and recall across skewed class distributions.

Table 4. Detailed Performance Metrics Across Models

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Macro)
Random Forest	0.75	0.73	0.75	0.53
LSTM	0.71	0.77	0.71	0.52
indoBERT	0.81	0.79	0.81	0.56

The data proves that IndoBERT achieved the highest global accuracy of 81.5%. However, accuracy and global metrics like Micro F1-Score can be highly deceptive in heavily skewed datasets. The fundamental scientific rationale for prioritizing the Macro-averaged F1-Score over the Micro-averaged F1-Score in this study lies in how these metrics calculate class weights. Micro F1-Score pools the total true positives, false negatives, and false positives across all classes, which inherently biases the metric toward the performance of the majority class (Positive sentiment). Consequently, a model could perform catastrophically on the rare but critical minority class (Neutral sentiment) and still maintain a deceptively high Micro F1-Score.

In contrast, the Macro F1-Score computes the F1-score independently for each individual category and then takes their unweighted arithmetic mean. By treating all classes with equal importance regardless of their sample size (support size), the Macro F1-Score serves as a rigorous and uncompromised benchmark for evaluating a model's true generalization capacity across the entire distribution spectrum. This strict penalty mechanism explains why IndoBERT's Macro F1-Score sits at 0.56 despite an 81.5% overall accuracy, highlighting that the model's bottleneck is precisely captured by its struggle to navigate the ambiguous boundaries of the minor Neutral class.

3.2 Analysis of Confusion Matrices

The behavioral characteristics of each model can be best understood through their respective confusion matrices. These matrices illustrate exactly where the models "confused" one sentiment class for another, providing insight into the semantic overlap of the dataset.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

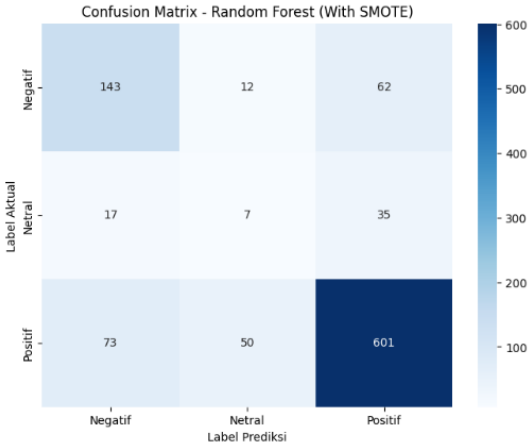

Figure 2. Confusion Matrix of Random Forest with SMOTE

Figure 2, the Random Forest model shows high stability in the Positive class but significant leakage between the Negative and Neutral classes despite SMOTE intervention.

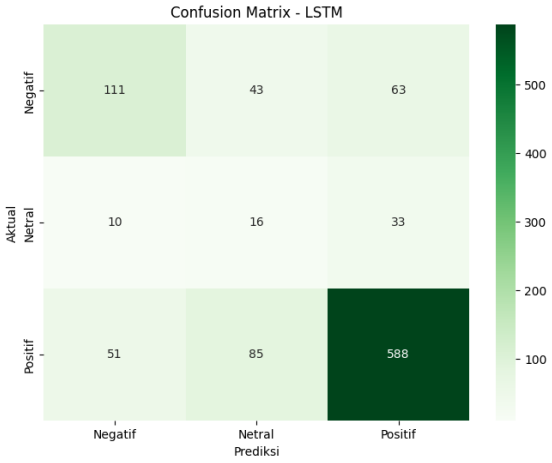

Figure 3. Confusion Matrix of LSTM with Class Weighting

Figure 3, the LSTM model demonstrates the highest sensitivity (Recall) for the Neutral class but suffers from a higher rate of False Positives.

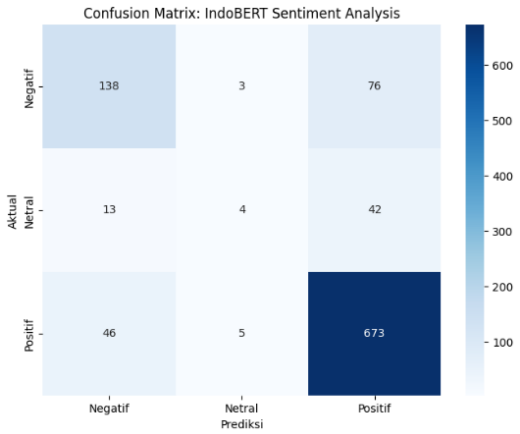

Figure 4. Confusion Matrix of IndoBERT

Figure 4, IndoBERT exhibits the most balanced performance, with the sharpest diagonal values, though it still struggles with the inherent ambiguity of Neutral reviews.

3.3 Superiority of Transformer-based Contextual Embeddings

The primary finding of this study is that IndoBERT's bidirectional self-attention mechanism provides a superior understanding of the "slang-heavy" and ungrammatical nature of TikTok reviews. Unlike Random Forest, which relies

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

on TF-IDF vectors that treat words as independent units (Bag-of-Words), IndoBERT analyzes the context of a word based on all surrounding words. In the Indonesian language, where the meaning of a word can shift drastically depending on its prefix or its position in a sentence, this contextual awareness is vital.

This finding is strongly synthesized with the work of Khan et al. [40], who noted that traditional machine learning models often fail to capture sarcasm or negation in Indonesian social media because they lack a deep understanding of syntax. By utilizing a model pre-trained on the Indo4B corpus, our research proves that transfer learning allows the model to "understand" Indonesian slang before even seeing the TikTok dataset. This aligns with Alfian et al. [41] assertion that transformer-based models are significantly more robust against the "Out-of-Vocabulary" (OOV) problem common in digital feedback. Our IndoBERT model's ability to achieve a 0.89 F1-score for the Positive class proves that it has successfully mapped the diverse ways Indonesian users express satisfaction, from formal praise to casual "gaul" expressions.

3.4 The Efficacy of Architecture-Specific Balancing Techniques

A major contribution of this study is the comparative evaluation of SMOTE, Class Weighting, and Random OverSampling (ROS). In the Random Forest model, SMOTE was used to generate synthetic features. While this increased the data count, the confusion matrix (Figure 1) shows that these synthetic points often blurred the lines between Negative and Neutral sentiments. This is because SMOTE creates "average" vectors that may not correspond to real linguistic patterns.

Conversely, the use of Random OverSampling (ROS) in the IndoBERT pipeline proved to be more effective for maintaining semantic integrity. By duplicating actual sentences, the model was exposed to real-world sentence structures multiple times, reinforcing its learning of minority class patterns without the noise of synthetic data. This finding synthesizes with Ekudanyo and Ezugwu [42] research, which found that for deep learning architectures, the preservation of the original text sequence is more important than the generation of synthetic feature vectors. Our results specifically show that LSTM with Class Weighting achieved a Neutral Recall of 0.27, the highest among all models. This indicates that while IndoBERT is better for overall accuracy, Class Weighting in LSTM is a powerful tool for forcing a model to "pay attention" to rare classes, a technique also championed by Liem et al. [43] in their study of imbalanced Indonesian datasets.

3.5 The "Neutrality Bottleneck" and Boundary Ambiguity

The most difficult challenge identified in this study is the classification of the Neutral class. All models, including the SOTA IndoBERT, showed a significant performance drop here. Analysis of the misclassified samples in the confusion matrices shows that users often give a 3-star rating for "mixed sentiments" reviews that contain both a compliment and a complaint (e.g., "Aplikasi bagus, tapi sering log out sendiri").

This finding resonates with Kusumaningrum et al. [44], who described the "Neutral" class in Indonesian sentiment analysis as a "semantic grey area" rather than a distinct category. When a user is both happy and frustrated, the model is forced to choose one pole, often defaulting to "Positive" because of the overall frequency of positive words in the training set. This synthesis suggests that the 81% accuracy achieved by IndoBERT is likely near the "human-level" limit for 3-way classification on this specific dataset. As noted by Fehle et al. [45], without a "Mixed" category or aspect-based sentiment analysis (ABSA), the Neutral class will always function as a source of noise that lowers the Macro F1-score. Our study confirms that even with advanced balancing, the inherent ambiguity of human emotion in short-form text remains a formidable obstacle.

3.6 Robustness of Negation-Aware Preprocessing

A critical technical finding involves the impact of the Negation-Aware Preprocessing implemented in this study. In early iterations of the Random Forest model, removing all stopwords led to a 5% drop in accuracy for the Negative class. By protecting negation words like "tidak", "bukan", and "kurang", we allowed the models to correctly identify inverted polarities.

This finding supports the methodology of Ghiffari et al. [46], who argued that standard stopword lists are often detrimental to sentiment analysis in the Indonesian language. By treating negation as a "protected feature," our models were able to distinguish between "puas" (satisfied) and "tidak puas" (not satisfied) even after stemming. This synthesis highlights that the success of State-of-the-Art models like IndoBERT is still fundamentally tied to the quality of the initial text cleaning process. In the context of TikTok reviews, where brevity is common, a single negation word can carry the entire weight of the sentiment, and our pipeline successfully preserved this signal.

3.7 Limitations and Future Work

Despite the high accuracy and robust methodology, several limitations must be addressed. First, the dataset is restricted to 5,000 reviews. While this is sufficient for fine-tuning IndoBERT, transformer models typically scale their performance with larger datasets. Furthermore, this study did not account for sarcasm detection, which is a common feature of Indonesian social media discourse. A user might say "Bagus banget aplikasinya, sampai tidak bisa dibuka," which a model might misinterpret as positive due to the word "bagus". Future research should focus on three specific areas:

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

a. Aspect-Based Sentiment Analysis (ABSA): Instead of a single label per review, models should identify sentiments toward specific aspects like "UI/UX," "Connection," or "Content Policy."
b. Hybrid Model Architectures: Combining the contextual power of IndoBERT with the sequential memory of Bi-LSTM layers could potentially resolve some of the Neutral class ambiguities.
c. Data Augmentation via Back-Translation: To expand the minority Neutral class without simple duplication (ROS), future researchers could translate Indonesian reviews to English and back to Indonesian to create semantically similar but structurally different samples. This would provide a more diverse training signal for the Transformer's attention mechanism.

4. CONCLUSION

This research has successfully conducted a comprehensive comparative analysis of three distinct computational paradigms sRandom Forest, Long Short-Term Memory (LSTM), and IndoBERT for the sentiment classification of TikTok application reviews. The experimental results provide conclusive evidence that the Transformer-based architecture, specifically IndoBERT, outperforms both traditional machine learning and sequential deep learning models. IndoBERT achieved the highest overall accuracy of 81.5% and a Macro F1-Score of 0.56, proving its superior capability in capturing the complex contextual nuances, slang, and non-formal syntax inherent in Indonesian social media discourse. Based on these empirical findings, we highly recommend IndoBERT as the primary deployment model for production environments requiring maximum accuracy, while LSTM serves as a viable intermediate alternative for sequential memory processing, and Random Forest remains a computationally light and robust baseline. In contrast, Random Forest and LSTM yielded lower accuracies of 75% and 71%, respectively, primarily due to their limitations in handling deep semantic dependencies and linguistic ambiguities. A key contribution of this study is the identification of architecture-specific data-balancing strategies that optimize performance on imbalanced datasets. The integration of Random OverSampling (ROS) with IndoBERT proved more effective than the synthetic generation approach of SMOTE for Random Forest, as it preserved the natural linguistic sequences vital for the Transformer's attention mechanism. However, this study also highlights a persistent "neutrality bottleneck," where all models struggled to accurately classify neutral sentiments due to the high degree of boundary overlap and mixed-polarity feedback. While LSTM exhibited the highest sensitivity for the neutral class through cost-sensitive class weighting, it could not match the global robustness provided by IndoBERT. The practical implications of this research are significant for application developers and UI/UX researchers. By implementing the IndoBERT-based sentiment analysis pipeline, developers can automate the monitoring of user feedback with higher precision, allowing for more objective and data-driven service improvements. Future research should focus on addressing the ambiguity of the neutral class through aspect-based sentiment analysis (ABSA) or the implementation of hybrid architectures that combine transformer embeddings with recurrent layers. Additionally, expanding the corpus with diverse regional dialects and exploring advanced data augmentation techniques like back-translation remains a promising direction for achieving even higher levels of classification granularity in the Indonesian NLP domain.

REFERENCES

[1] J. B. Pick, F. Ren, and A. Sarkar, “Social media access and purposeful use in China: Geospatial patterns and socioeconomic and COVID-19 influences,” *Telecomm. Policy*, vol. 49, no. 7, pp. 1–18, 2025, doi: 10.1016/j.telpol.2025.103002.
[2] O. Grljević, M. Marić, and R. Božić, “Exploring Mobile Application User Experience Through Topic Modeling,” *Sustainability (Switzerland)*, vol. 17, no. 3, pp. 1–32, 2025, doi: 10.3390/su17031109.
[3] M. Mustak, H. Hallikainen, T. Laukkanen, L. Plé, L. D. Hollebeek, and M. Aleem, “Using machine learning to develop customer insights from user-generated content,” *Journal of Retailing and Consumer Services*, vol. 81, no. March, pp. 1–14, 2024, doi: 10.1016/j.jretconser.2024.104034.
[4] W. Chen, W. Hussain, F. Cauteruccio, and X. Zhang, “Deep Learning for Financial Time Series Prediction: A State-of-the-Art Review of Standalone and Hybrid Models,” *CMES - Computer Modeling in Engineering and Sciences*, vol. 139, no. 1, pp. 187–224, 2023, doi: 10.32604/cmes.2023.031388.
[5] A. Daza, N. D. González Rueda, M. S. Aguilar Sánchez, W. F. Robles Espíritu, and M. E. Chauca Quiñones, “Sentiment Analysis on E-Commerce Product Reviews Using Machine Learning and Deep Learning Algorithms: A Bibliometric Analysisand Systematic Literature Review, Challenges and Future Works,” *International Journal of Information Management Data Insights*, vol. 4, no. 2, pp. 1–20, 2024, doi: 10.1016/j.jjimei.2024.100267.
[6] E. C. M. Torres and L. G. de Picado-Santos, “Using Sentiment Analysis to Study the Potential for Improving Sustainable Mobility in University Campuses,” *Sustainability (Switzerland)*, vol. 17, no. 14, pp. 1–22, 2025, doi: 10.3390/su17146645.
[7] M. Kim, S. Kim, Y. Park, S. Bahn, S. H. Ahn, and B. NambiNarayanan, “User Sentiment Analysis Based on Securities Application Elements,” *Behavioral Sciences*, vol. 14, no. 9, pp. 1–19, 2024, doi: 10.3390/bs14090814.
[8] K. Imtihan, L. Mutawali, W. Bagye, and A. Tantoni, “Automated Label Extraction for Sentiment Analysis in Indonesian Text,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 15, no. 3, pp. 718–728, 2025, doi: 10.18517/ijaseit.15.3.20602.
[9] A. Abdillah, I. Widianingsih, R. A. Buchari, and H. Nurasa, “Big data security & individual (psychological) resilience: A review of social media risks and lessons learned from Indonesia,” *Array*, vol. 21, no. February, pp. 1–13, 2024, doi: 10.1016/j.array.2024.100336.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

[10] N. Merayo, R. Cotelo, R. Carratalá-Sáez, and F. J. Andújar, "Applying machine learning to assess emotional reactions to video game content streamed on Spanish Twitch channels," *Comput. Speech Lang.*, vol. 88, no. March, pp. 1–23, 2024, doi: 10.1016/j.csl.2024.101651.
[11] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 8, pp. 1–16, 2023, doi: 10.1016/j.heliyon.2023.e18647.
[12] M. Abdalsalam, C. Li, A. Dahou, and N. Kryvinska, "Terrorism Attack Classification Using Machine Learning: The Effectiveness of Using Textual Features Extracted from GTD Dataset," *CMES - Computer Modeling in Engineering and Sciences*, vol. 138, no. 2, pp. 1427–1467, 2024, doi: 10.32604/cmes.2023.029911.
[13] B. R. Ermawan, M. B. Prayoga, A. R. Fadhillah, and E. Utami, "Optimizing Sentiment Classification Models for TikTok Comments using Emotion-Based Preprocessing and Grid Search," vol. 10, no. 1, pp. 535–547, 2026.
[14] Y. Aliyu, A. Sarlan, K. U. Danyaro, A. Sani abd Rahman, A. A. Muazu, and M. Y. Abubakar, "Deep learning techniques for sentiment analysis in code-switched Hausa-English tweets," *International Journal of Information Management Data Insights*, vol. 5, no. 1, pp. 1–19, 2025, doi: 10.1016/j.jjimei.2025.100330.
[15] A. M. Moslhi, H. H. Aly, and M. ElMessiery, "The Impact of Feature Extraction on Classification Accuracy Examined by Employing a Signal Transformer to Classify Hand Gestures Using Surface Electromyography Signals," *Sensors*, vol. 24, no. 4, pp. 1–21, 2024, doi: 10.3390/s24041259.
[16] C. Suhaeni and H. S. Yong, "Enhancing Imbalanced Sentiment Analysis: A GPT-3-Based Sentence-by-Sentence Generation Approach," *Applied Sciences (Switzerland)*, vol. 14, no. 2, pp. 1–24, 2024, doi: 10.3390/app14020622.
[17] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, *A survey on imbalanced learning: latest research, applications and future directions*, vol. 57, no. 6. 2024. doi: 10.1007/s10462-024-10759-6.
[18] N. B. Yadav, "'Harnessing Customer Feedback for Product Recommendations: An Aspect-Level Sentiment Analysis Framework,'" *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 57–67, 2023, doi: 10.1007/s44230-023-00018-2.
[19] F. Sağlam, *DatRel: a noise-tolerant data relocation approach for effective synthetic data generation in imbalanced classifiers*, vol. 114, no. 5. Springer US, 2025. doi: 10.1007/s10994-025-06755-8.
[20] M. Wojciuk, Z. Swiderska-Chadaj, K. Siwek, and A. Gertych, "Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization," *Heliyon*, vol. 10, no. 5, pp. 1–11, 2024, doi: 10.1016/j.heliyon.2024.e26586.
[21] 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.
[22] S. A. Amamra, "Random Forest-Based Machine Learning Model Design for 21,700/5 Ah Lithium Cell Health Prediction Using Experimental Data," *Physchem*, vol. 5, no. 1, pp. 1–19, 2025, doi: 10.3390/physchem5010012.
[23] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, pp. 1–13, 2021, doi: 10.1016/j.jjimei.2020.100007.
[24] N. Saraswathi, T. Sasi Rooba, and S. Chakaravarthi, "Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews," *Measurement: Sensors*, vol. 29, no. May, pp. 1–7, 2023, doi: 10.1016/j.measen.2023.100888.
[25] E. Elhosary and O. Moselhi, "Evaluating Natural Language Processing Algorithms for Improved Hazard and Operability Analysis," *Geodata and AI*, vol. 4, no. May, pp. 1–15, 2025, doi: 10.1016/j.geoai.2025.100026.
[26] A. Alharthi, M. Alaryani, and S. Kaddoura, "A comparative study of machine learning and deep learning models in binary and multiclass classification for intrusion detection systems," *Array*, vol. 26, no. May, pp. 1–15, 2025, doi: 10.1016/j.array.2025.100406.
[27] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Advancements in natural language processing: Implications, challenges, and future directions," *Telematics and Informatics Reports*, vol. 16, no. November, pp. 1–17, 2024, doi: 10.1016/j.teler.2024.100173.
[28] S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 5, pp. 1–34, 2024, doi: 10.1016/j.jksuci.2024.102068.
[29] A. O. Almagrabi, "A Deep CNN-LSTM-Based Feature Extraction for Cyber-Physical System Monitoring," *Computers, Materials and Continua*, vol. 76, no. 2, pp. 2079–2093, 2023, doi: 10.32604/cmc.2023.039683.
[30] Q. Gu, Z. Wang, H. Zhang, S. Sui, and R. Wang, "Aspect-Level Sentiment Analysis Based on Syntax-Aware and Graph Convolutional Networks," *Applied Sciences (Switzerland)*, vol. 14, no. 2, pp. 1–13, 2024, doi: 10.3390/app14020729.
[31] B. Ihnaini, B. Abuhaija, E. A. Mills, and M. Mahmuddin, "Semantic similarity on multimodal data: A comprehensive survey with applications," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 10, pp. 1–37, 2024, doi: 10.1016/j.jksuci.2024.102263.
[32] M. J. Abbass, R. Lis, and W. Rebizant, "A Predictive Model Using Long Short-Time Memory (LSTM) Technique for Power System Voltage Stability," *Applied Sciences (Switzerland)*, vol. 14, no. 16, pp. 1–13, 2024, doi: 10.3390/app14167279.
[33] D. G. da Silva and A. A. de M. Meneses, "Comparing Long Short-Term Memory (LSTM) and bidirectional LSTM deep neural networks for power consumption prediction," *Energy Reports*, vol. 10, no. May, pp. 3315–3334, 2023, doi: 10.1016/j.egyr.2023.09.175.
[34] J. Campino, "Unleashing the transformers: NLP models detect AI writing in education," *Journal of Computers in Education*, vol. 12, no. 2, pp. 645–673, 2025, doi: 10.1007/s40692-024-00325-y.
[35] P. W. Cahyo, U. S. Aesyi, W. A. Setianto, and T. Sulaiman, "A Novel Named Entity Recognition approach of Indonesian fake news using part of speech and BERT model on presidential election," *International Journal of Information Management Data Insights*, vol. 5, no. 2, pp. 1–11, 2025, doi: 10.1016/j.jjimei.2025.100354.
[36] A. Alamsyah and Y. Sagama, "Empowering Indonesian internet users: An approach to counter online toxicity and enhance digital well-being," *Intelligent Systems with Applications*, vol. 22, no. May, pp. 1–16, 2024, doi: 10.1016/j.iswa.2024.200394.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 570−579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9714

[37] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, pp. 1–17, 2025, doi: 10.1007/s13278-025-01444-9.
[38] A. S. Ekakristi, A. F. Wicaksono, and R. Mahendra, "Intermediate-task transfer learning for Indonesian NLP tasks," *Natural Language Processing Journal*, vol. 12, no. May, pp. 1–16, 2025, doi: 10.1016/j.nlp.2025.100161.
[39] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, no. April, pp. 1–12, 2022, doi: 10.1016/j.array.2022.100157.
[40] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, no. July, pp. 1–31, 2023, doi: 10.1016/j.nlp.2023.100026.
[41] M. Alfian, U. L. Yuhana, D. Siahaan, H. Munazharoh, and E. Pardede, "Out-of-Vocabulary Handling in Part-of-Speech Tagging: A Semantic Web-Driven Systematic Review," *Int. J. Semant. Web Inf. Syst.*, vol. 21, no. 1, pp. 1–36, 2025, doi: 10.4018/IJSWIS.388421.
[42] O. S. Ekundayo and A. E. Ezugwu, "Deep learning: Historical overview from inception to actualization, models, applications and future trends," *Appl. Soft Comput.*, vol. 181, no. May, pp. 1–31, 2025, doi: 10.1016/j.asoc.2025.113378.
[43] J. A. Liem, D. T. D. Utomo, B. Ignasio, and T. W. Cenggoro, "Deep Learning And Machine Learning For Indonesian Online Gambling Website Promotion Detection," *Procedia Comput. Sci.*, vol. 269, pp. 979–992, 2025, doi: 10.1016/j.procs.2025.09.040.
[44] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews," *Heliyon*, vol. 9, no. 6, pp. 1–12, 2023, doi: 10.1016/j.heliyon.2023.e17147.
[45] J. Fehle, U. Kruschwitz, N. C. Hellwig, and C. Wolff, "Leveraging fine-tuning of large language models for aspect-based sentiment analysis in resource-scarce environments," *Knowl. Based. Syst.*, vol. 336, no. January, pp. 1–23, 2026, doi: 10.1016/j.knosys.2026.115277.
[46] A. G. Ghifari, G. Y. Ananada, and K. Purwandari, "A Comparative Sentiment Analysis of Public Opinion on Indonesia's National Football Coach Using CRNN and SVM," *Procedia Comput. Sci.*, vol. 269, pp. 1485–1493, 2025, doi: 10.1016/j.procs.2025.09.090.