JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 11 Number 1 | June 2026: 14-26
DOI: 10.15575/join.v11i1.1704

Comparative Performance of Fine-Tuned IndoBERT BASE and LARGE Variants for Emotion Detection in Indonesian Tweets

Sri Winarno[1], Ika Novita Dewi[2], Adhitya Nugraha[3], Fahri Firdausillah[4], Maulatus Shaffira Fitri[5], Talitha Olga Ramadhani[6], Erna Amalia Widhiyanti [7], Ainur Rahma Miftakhul Rizqi[8]
[1,2,3,4,5]Research Center for Intelligent Distributed Surveillance and Security (IDSS), Universitas Dian Nuswantoro, Semarang, Indonesia
[6,7,8]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Article Info

Article history:

Received June 23, 2025
Revised September 01, 2025
Accepted October 06, 2025
Published April 25, 2026

Keywords:

Emotion Detection
IndoBERT
Indonesian Text
Optimization
Transformer Model

ABSTRACT

In the digital era, where emotions play a crucial role in shaping human behavior, communication, and decision-making, their expressions are often conveyed through short and informal texts on platforms such as Twitter. This research aims to improve the accuracy of emotion detection in Indonesian text using the IndoBERT-BASE-P2 and IndoBERT-LARGE-P2 transformer models. The dataset consists of 7,080 tweets annotated with six basic emotion categories (anger, fear, joy, love, neutral, and sad). The research methodology included text preprocessing, class balancing using SMOTE, and fine-tuning with optimized training parameters. Evaluation results show that IndoBERT-BASE-P2 achieved an accuracy of 84.43% and a macro F1-score of 84.33%, surpassing previous studies, while the larger IndoBERT-LARGE-P2 model tended to overfit and offered no meaningful improvement. Error analysis showed the neutral class was the most difficult to classify. These findings demonstrate that with effective preprocessing and parameter optimization, a smaller model can be a highly efficient solution for emotion classification in Indonesian text, especially in resource-constrained conditions.

Corresponding Author:

Ika Novita Dewi,
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang
Jalan Imam Bonjol No. 207, Pendrikan Kidul, Kec. Semarang Tengah, Kota Semarang, Jawa Tengah, 50131
Email: ikadewi@dsn.dinus.ac.id

1. INTRODUCTION

Emotions are a fundamental aspect of human life, profoundly influencing decision-making, behavior, and communication. In the digital age, these emotions are frequently expressed through text on platforms like Twitter. Despite being a rich data source, Twitter data presents significant challenges for automated analysis due to its brevity, informal language, and inherent contextual nuances [1]. Consequently, emotion detection in text has emerged as a critical research area in Natural Language Processing (NLP), with diverse applications spanning public opinion analysis, customer service, and mental health monitoring [2], [3].

Early studies in emotion detection predominantly utilized traditional machine learning methods such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes [4]. However, these approaches often rely on labor-intensive manual feature engineering and frequently struggle to capture the complex contextual meaning, especially in short, informal texts like tweets.

Furthermore, prior research has largely concentrated on high-resource languages, particularly English, overlooking languages like Indonesian that possess unique linguistic and emotional structures [5].

The advent of transformer-based models, such as BERT, has revolutionized NLP, significantly advancing performance across various tasks. Models like BERT, RoBERTa, and XLNet have consistently outperformed both traditional machine learning and earlier deep learning methods (e.g., LSTM and CNN) in emotion detection [6]. For the Indonesian language, the IndoBERT model has demonstrated robust capabilities in handling informal social media text [5],[7]. This trend is also observed in other low-resource languages; for instance, ensemble transformer models like TEmoX have achieved high performance in Bengali [8] and transformer-based models have surpassed deep learning baselines across Arabic dialectical and standard corpora, underscoring their effectiveness in capturing contextual understanding for emotion classification [9]. To address the identified challenges and advance Indonesian emotion detection, this research employs fine-tuned variants of IndoBERT. Specifically, we adopt IndoBERT-BASE-P2 as a primary model, which has been customized and optimized for Indonesian Twitter data by Putri et al. [5] and shown promising results. We also explore IndoBERT-LARGE-P2, a higher-capacity variant, to systematically evaluate the impact of model size on emotion recognition performance. This aligns with prior successful applications of IndoBERT for large-scale transfer learning in this domain [10].

Effective model performance also hinges on appropriate training hyperparameters. Drawing from best practices in previous research, this research implements a maximum input length of 128 tokens, incorporates a 500-step warm-up strategy, and applies a weight decay of 0.01. These settings are crucial for enhancing model stability and generalization capabilities [5]. Another significant challenge in real-world datasets is class imbalance, which can lead to biased model predictions. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to generate synthetic samples for minority emotion classes. This method has proven effective in improving classifier performance in imbalanced Twitter datasets [11].

Furthermore, the choice of the emotion model used for annotation is critical. Languré and Zareei [12] argue that emotion models should be treated as tunable hyperparameters, noting that performance differences often stem more from label schema variation than model architecture. Finally, this research evaluates IndoBERT-BASE-P2 and IndoBERT-LARGE-P2 alongside other transformer models using precision, recall, and F1-score to measure effectiveness in classifying six basic emotions: anger, fear, joy, love, neutral, and sad.

2. METHOD

This section describes the overall methodology used in building the emotion detection system, including dataset preparation, text preprocessing, class balancing, model architecture, and training configuration. The process flow of the system is illustrated in Figure 1, which provides a step-by-step overview from raw data to final model evaluation. Transformer-based models were fine-tuned using labeled Indonesian-language tweets, with several enhancements applied to address class imbalance and optimize training performance.

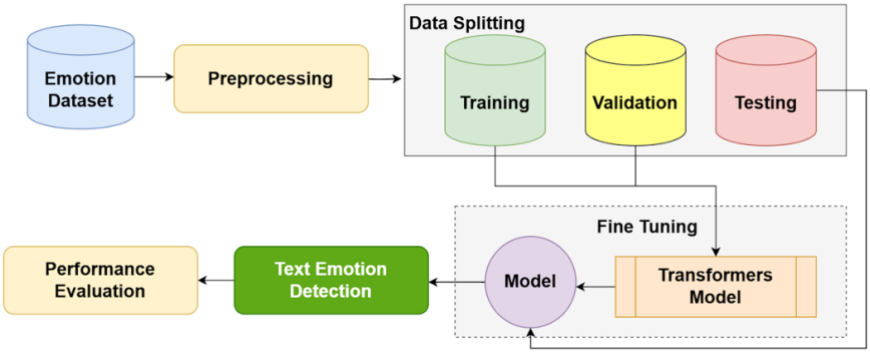

Figure 1. Research method

2.1 Dataset

This research utilizes a custom emotion dataset comprising 7,080 Indonesian public opinion tweets. This dataset has undergone careful text cleaning and manual annotation [13]. Data collection was performed via the Twitter API, leveraging the Tweepy library. To ensure relevance and avoid redundancy, tweets were retrieved using specific keywords tailored to each emotion category, and the "-filter:retweets" parameter was applied to exclude duplicated content.

The collected tweets offer a rich and diverse reflection of prevalent discourse in Indonesia. They span a wide array of topics, including social issues, political discourse, entertainment, and everyday conversations. Each tweet within the dataset is labeled with one of six discrete emotion categories: anger, fear, joy, love, sad, and neutral. This categorical scheme is structured based on Parrott's basic emotion theory. The annotation process was conducted by multiple trained human annotators, achieving a high degree of inter-annotator consistency. This robust annotation methodology ensures the reliability and validity of the labels, positioning this dataset as a valuable resource for Indonesian language emotion analysis research, a domain still characterized by limited publicly available annotated resources. The granular distribution of emotion classes within the dataset is presented in Table 1.

Table 1. Emotion classes distribution of the Indonesian public opinion dataset

Emotion class	Counts
Neutral	2,001
Joy	1,275
Anger	1,130
Sad	1,003
Fear	911
Love	760
Total	**7,080**

2.2 Preprocessing

The data preprocessing stage is critical for preparing raw text for effective model input, involving both cleaning and normalization techniques. Initially, a series of standard text cleaning operations were applied: duplicate tweets were identified and removed, all text was converted to lowercase, and punctuation, numbers, and superfluous whitespace were eliminated. User handles and common stop words were also removed to focus on relevant semantic content. Furthermore, stemming was employed to reduce words to their morphological root forms, enhancing lexical generalization. This initial cleaning process reduced the dataset from 7,080 to 7,027 unique tweet entries.

Crucially, to ensure a fair and direct comparison with the IndoBERT-BASE-P2 model as replicated from reference [5], no preprocessing steps (e.g., cleaning, normalization, or stemming) were applied to the baseline model's input. Instead, these comprehensive preprocessing steps were exclusively implemented for the custom model proposed in this research. This differential treatment directly accounts for the anticipated performance disparity: the preprocessed text provides a clearer, less noisy representation of the data, which is hypothesized to significantly improve the custom model's ability to extract meaningful features and achieve superior results compared to the unpreprocessed baseline.

In addition to text cleaning, class balancing was performed using the Synthetic Minority Over-sampling Technique (SMOTE) to address the inherent class imbalance. SMOTE was applied at the preprocessing stage, prior to data splitting and tokenization, ensuring that each emotion class achieved a balanced representation across both training and evaluation sets. This technique mitigates bias towards majority classes by generating synthetic samples for minority classes through interpolation between existing data points and their nearest neighbors in the feature space [14]. The generation of synthetic data is defined by equation (1).

$$Xnew = x_i + \lambda \times (x_{nn} - x_i) \quad (1)$$

Where x_i represents a data point from the minority class, X_{nn} is one of its nearest neighbors from the same class, and λ is a random number between 0 and 1. This approach effectively expands the

coverage of underrepresented classes without simply duplicating existing instances. In this research, SMOTE was applied to the minority classes ('anger', 'fear', 'joy', 'love', 'neutral', and 'sad') such that each emotion category achieved a balanced count of 1,992 samples. This resulted in an expanded total dataset size of 11,952 entries. This step was essential to prevent the model from overfitting to the dominant 'Neutral' class and to enhance its generalization capabilities across all emotional categories.

Overall, the combination of text cleaning and strategic class balancing demonstrably had a significant positive impact on the final model performance. This is particularly evident in the superior results achieved by the custom model compared to the baseline, especially in the accurate detection of emotions with inherently smaller data representation, such as 'love' and 'fear'. Therefore, preprocessing is not merely a preparatory phase but a pivotal factor directly influencing the success and robustness of the emotion classification model.

2.3. Splitting

The splitting process was performed using the train_test_split function from the Scikit-learn library with a stratified sampling strategy to maintain equal class proportions in each subset. First, the dataset was split into 80% temporary training data and 20% testing data. From the temporary training data, 10% was further separated to serve as a validation set. This resulted in 72% of the total data being used for training, 8% for validation, and 20% for testing. The stratification approach ensured that each emotion class remained proportionally represented in all three subsets, preserving the class balance established during the SMOTE process.

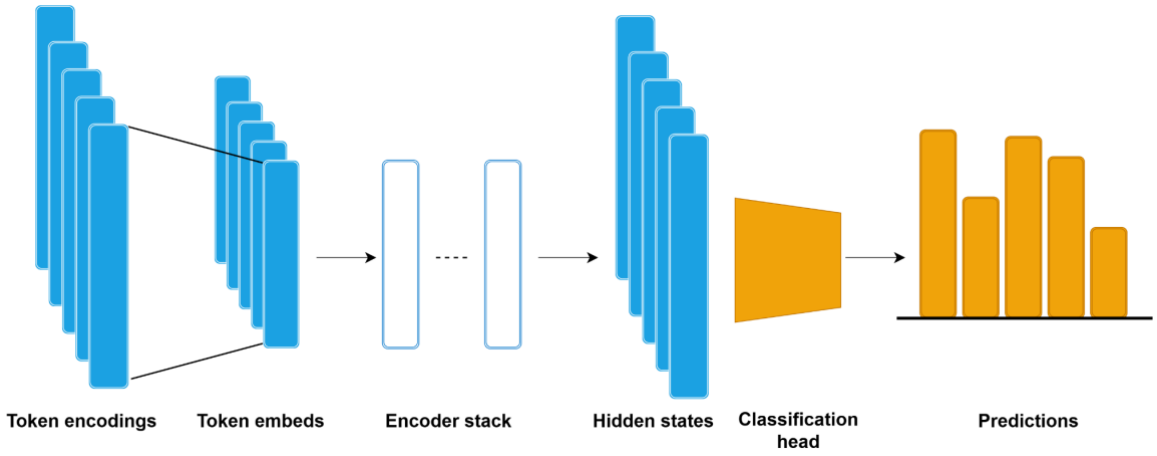

Figure 2. Model architecture [15]

The stratified train-validation-test split, detailed in Table 2, is crucial in emotion classification tasks, as it prevents data leakage and ensures that the model is evaluated on unseen and representative data. The final split comprised 8,604 samples for training, 957 samples for validation, and 2,391 samples for testing.

Table 2. Number of samples in train, validation, and test set

X_train.shape	X_val.shape	X_test.shape
8,604	957	2,391

2.4 Transformers-based Emotion Detection

The architecture of the fine-tuned transformer model used for emotion classification is illustrated in Figure 2. The input text is first tokenized into embeddings and then passed through a stack of encoder layers. The resulting hidden states are fed into a classification head to produce emotion predictions.

This research utilizes two transformer-based models: IndoBERT-BASE-P2 and IndoBERT-LARGE-P2. These models are fine-tuned to classify six emotion categories: 'anger', 'fear', 'joy', 'love', 'neutral', and 'sad'. Mathematically, the transformer generates a hidden representation $h = f_{transformer}(X)$, which is then mapped to a probability distribution over emotion labels using a fully connected layer and a Softmax function:

$$y = softmax(Wh + b) \quad (2)$$

The IndoBERT-BASE-P2 model in Figure 2 follows the architecture from [5], comprising 12 encoder layers and using ReLU activations in the hidden layers. A softmax function is applied at the output layer to produce the final classification.

2.5 Experiment Settings

The model fine-tuning was conducted using the Hugging Face Training Arguments API. While the core architecture follows the IndoBERT-BASE-P2 configuration used in [5], several training parameters were adjusted to better suit this research. Specifically, the maximum input sequence length was set to 128 tokens, which differs from the 256 token configuration in [5], in order to reduce computational load and memory usage. Additionally, we manually configured the weight decay to 0.01 and applied a warmup strategy of 500 steps to improve model generalization and training stability [16]. The optimizer used in this research was AdamW (Adam with decoupled weight decay), which is well-suited for transformer-based models due to its ability to decouple L2 regularization from the optimization steps, thereby improving convergence and generalization performance.

Table 3. Hyperparameter settings

Hyperparameter	Values
Hidden Layer	12
Number of Input	[128]
Number of Output	[6]
Weight Decay	0.01
Warmup Steps	500
Optimizer	Adam
Activation Function	ReLu (input layer), softmax (output layer)

A summary of the key hyperparameters utilized during the model fine-tuning process is presented in Table 3. These include architectural settings such as the number of hidden layers and input/output dimensions, as well as training configurations like weight decay, warmup steps, optimizer choice, and activation functions. The selected hyperparameters were carefully adjusted to optimize the performance of the IndoBERT-BASE-P2 model while maintaining computational efficiency within the limitations of the available resources. Both models were trained for 10 epochs with a batch size of 8 for both training and evaluation. Model evaluation and checkpoint saving were performed at the end of each epoch, with intermediate evaluation every 500 steps and logging every 100 steps. All training outputs were saved in the ./results directory.

The experiments were conducted at Google Colaboratory using a single NVIDIA Tesla T4 GPU. This setup provided a balance between computational efficiency and resource availability, making it suitable for fine-tuning large transformer-based models such as IndoBERT-BASE-P2 within a constrained academic environment.

2.6 Performance Evaluation

To evaluate the effectiveness of the emotion classification models, this research employed standard metrics including accuracy, precision, recall, and F1-score, measured both globally and per emotion class [17]. Accuracy reflects the overall proportion of correct predictions across the dataset, calculated using equation (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

In addition to overall accuracy, class-specific precision, recall, and F1-score were computed to better assess the model's performance across the six emotion categories: ‘anger’, ‘fear’, ‘joy’, ‘love’, ‘neutral’, and ‘sad’. Precision for class *i* is given by equation (4).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

While recall is defined as shown in equation (5).

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

The F1-score, which balances precision and recall, is calculated by (6).

$$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

To ensure equal weight across classes regardless of their frequency, macro-averaging was applied to the precision, recall, and F1-score metrics. This is particularly important for datasets that were originally imbalanced and then adjusted using SMOTE. Macro-averaged F1-score remains one of the most reliable metrics for balanced evaluation in multi-class settings, especially when working with text classification in low-resource languages [18].

3. RESULT AND DISCUSSION

This section presents the evaluation results of several transformer-based models fine-tuned for emotion detection in Indonesian texts. The experiments include baseline replication, hyperparameter tuning, and a comparative analysis across multiple pretrained models. This research successfully evaluated and improved the performance of transformer models for detecting emotions in Indonesian text, with a particular focus on IndoBERT-BASE-P2 and IndoBERT-LARGE-P2. Our main findings indicate that optimized training parameter settings, such as `max_length`, `warmup_steps`, and `weight_decay`, significantly improve the model's accuracy and macro F1-score. Specifically, the IndoBERT-BASE-P2 model with our customized configuration achieved a macro F1-score of 0.8433 and an accuracy of 0.8443, far surpassing the initial configuration reported in previous studies. This improvement underscores the crucial role of careful parameter tuning in maximizing the potential of pre-trained language models for detailed tasks such as emotion detection.

Additionally, our research highlights the importance of the data preparation stage. Unlike previous basic approaches, the addition of steps such as data cleaning, stopword removal, stemming, and especially class balancing using the SMOTE technique proved to be highly beneficial. This comprehensive data preparation improves the quality of text representation, enabling the model to capture more subtle emotional nuances more effectively, especially for minority emotion classes like ‘love’ and ‘fear’. Although larger models like IndoBERT-LARGE-P2 show competitive results, the fine-tuned IndoBERT-BASE-P2 achieves comparable performance with significantly better computational efficiency, making it the optimal choice for real-world applications with limited resources. This work contributes to the field by providing a robust and efficient solution for Indonesian language emotion classification, demonstrating that the combination of proper data preparation and strategic model tuning can yield substantial performance improvements without adding architectural complexity.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Further research could explore multimodal data integration or delve into more detailed emotion detection in Indonesian language text.

3.1 Replication Result

To address class imbalance prior to model training, SMOTE was applied directly on the raw textual dataset before any data splitting and tokenization [19]. This ensured that all emotion categories had equal representation during training, validation, and testing phases. Initially, the 'neutral' class was the most dominant (1,992 samples), while the 'love' class was the smallest (754 samples). Other classes ranged between 906 and 1,258 entries. SMOTE was applied to the minority classes ('anger', 'fear', 'joy', 'love', and 'sad'), resulting in a balanced dataset with 1,992 samples for each category and a total of 11,952 entries overall. The emotion class distribution, contrasting the state before SMOTE application with the data after augmentation, is presented in Figure 3.

After the training and fine-tuning process for various Transformer model architectures described in the previous section, the next step is to evaluate the performance of each model. The main purpose of this evaluation is to measure the effectiveness of the model in accurately and fairly classifying emotions in Indonesian text. Additionally, this evaluation aims to compare the performance of models with uniform training parameters and to replicate the results of previous research [5].

The evaluation was conducted using several key metrics, namely accuracy, F1-score, train loss, and validation loss. The primary focus is on the macro F1-score metric because it provides a fair assessment of the model's performance across all emotion classes, unaffected by data imbalance per class [20]. This means that even if there are classes with dominant data (such as 'neutral') or minority data (such as 'love'), the macro F1-score still considers the model's performance evenly for each emotion label.

Figure 3. Emotion Class Distribution Before and After SMOTE Balancing

Evaluation results, detailed in Table 4, demonstrate that the retrained IndoBERT-BASE-P2 model, utilizing customized parameters, yields optimal performance, specifically an accuracy of 0.8443 and an F1-score of 0.8433. This model shows a significant improvement compared to the results in the previous reference paper, which had an F1-score of only 0.7648. This indicates that parameter adjustments, particularly reducing the maximum token input length (max_length) from 256 to 128, can enhance the model's efficiency and accuracy [5].

Table 4. Model comparison

Model	Epoch	WS	ML	WD	Training time	Acc	F1	Epoch
IndoBERT-BASE-P2 (Paper)	50	-	256	-	-	0.79	0.78	50
IndoBERT-BASE-P2 (Baseline Paper)	10	500	256	0.01	0:18:11	0.7546	0.7648	10
IndoBERT-BASE-P2 (Custom)	10	500	128	0.01	0:47:01	0.8443	0.8433	10
IndoBERT-Large-P2 (Baseline Paper)	10	500	256	0.01	0:48:00	0.7859	0.7915	10
IndoBERT-Large-P2 (Custom)	10	500	128	0.01	2:38:35	0.8435	0.8406	10

Notes: Epoch = num_train_epochs; WS = warmup_steps; ML = max_length; WD = weight_decay; Acc = Accuracy; F1 = F1-score

The IndoBERT-LARGE-P2 model also shows competitive results, with this custom model version achieving an F1-score of 0.8406, only slightly below the BASE version. However, the training time for the LARGE model is much longer, reaching 2 hours and 38 minutes, compared to just 47 minutes for the BASE model. This indicates that although models with larger architectures have higher modeling capacity, they do not always result in proportional performance improvements relative to the time and resources used. In fact, in some cases, larger models can overfit the training data, especially if the amount of data is relatively limited.

Meanwhile, the replication results of the IndoBERT-BASE-P2 (Baseline Paper) and IndoBERT-LARGE-P2 (Baseline Paper) models successfully approached the original publication results, with F1-scores of 0.7648 and 0.7915, respectively. However, their performance remained lower than the custom version because the training parameters used were not optimized [5].

Overall, this table reinforces the finding that optimizing training parameters can have a significant impact on Transformer model performance, even for models with the same architecture. With the right fine-tuning strategy, models like IndoBERT-BASE-P2 can be an effective and efficient solution for emotion classification tasks on Indonesian-language text, outperforming more complex models in terms of both performance and training time.

3.2 Comparison Across Models

To further evaluate the effectiveness of various transformer architectures in Indonesian text emotion classification tasks, a comparative analysis was conducted on several model configurations. This comparison includes models using default parameters (baseline) as reported in previous studies, as well as models with customized hyperparameters aimed at performance optimization. The models compared include IndoBERT-BASE-P2 and IndoBERT-LARGE-P2, each evaluated using both the standard training configuration from prior research and a custom training setup optimized in this research [5].

The evaluation metrics used to assess each model include Precision, Recall, and F1-score per emotion label. Additionally, macro average and weighted average metrics are used to provide a broader performance overview. While the macro average treats each class equally, the weighted average accounts for class imbalance by weighting each class proportionally to its frequency. Overall accuracy is also reported as a supplementary metric. The performance comparison between these models, in terms of both accuracy and F1-score, is illustrated in Figure 4, where it can be observed that the custom configurations consistently outperform the baseline setups.

A comparison of five models is provided in Table 5, including the IndoBERT-BASE-P2 from its original publication, and the baseline and custom variants for both IndoBERT-BASE-P2 and IndoBERT-LARGE-P2. IndoBERT-BASE-P2 from the previous research shows that the emotion ‘anger’ achieved the highest score (F1-score 0.98), followed by ‘joy’ (0.79) and ‘sad’ (0.77). Meanwhile, the emotion ‘neutral’ recorded the lowest F1 score of 0.65, indicating that this label is more difficult to classify consistently, likely due to the small amount of data or its more complex and diverse linguistic expressions. The overall average F1-score on this baseline is 0.78, which serves as a baseline for comparison in subsequent experiments with adjusted parameters.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

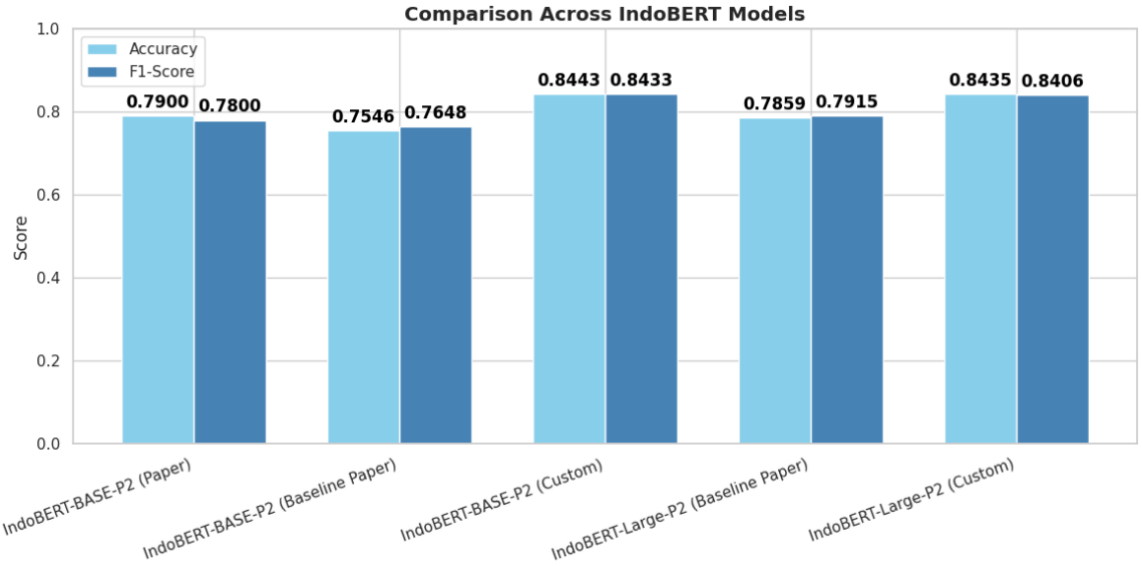

Figure 4. Performance comparison across IndoBERT models

Table 5. Comparison of emotional outcomes between models

Model	Evaluation Matrix	Emotion						Acc	Macro Avg	Weighted Avg
		Anger	Fear	Joy	Love	Neutral	Sad			
IndoBERT-BASE-P2 (Paper Reference)	Precision	**1.00**	0.76	0.81	0.78	0.61	0.79	-	0.79	-
	Recall	**0.97**	0.73	0.78	0.75	0.70	0.76	-	0.78	-
	F1-Score	**0.98**	0.75	0.79	0.77	0.65	0.77	-	0.78	-
IndoBERT-BASE-P2 (Baseline Paper)	Precision	0.82	0.86	**0.84**	0.79	0.64	0.73	-	0.78	0.76
	Recall	0.76	0.82	0.71	0.74	**0.77**	0.72	-	0.75	0.75
	F1-Score	0.79	0.84	0.77	0.76	0.70	0.72	0.75	0.76	0.76
IndoBERT-BASE-P2 (Custom)	Precision	0.84	**0.88**	0.83	**0.92**	0.71	**0.85**	-	**0.84**	**0.84**
	Recall	0.88	0.94	0.85	**0.92**	0.61	**0.89**	-	**0.85**	**0.84**
	F1-Score	0.86	**0.91**	0.84	**0.92**	0.66	**0.87**	**0.84**	**0.84**	**0.84**
IndoBERT-LARGE-P2 (Baseline Paper)	Precision	0.79	0.84	0.76	0.72	**0.82**	0.77	-	0.78	0.79
	Recall	0.90	0.88	0.80	0.85	0.64	0.76	-	0.81	0.78
	F1-Score	0.84	0.86	0.78	0.78	**0.73**	0.76	0.78	0.79	0.78
IndoBERT-LARGE-P2 (Custom)	Precision	0.83	0.87	0.83	0.91	0.75	0.83	-	0.84	0.84
	Recall	0.89	**0.95**	**0.87**	0.91	0.58	0.87	-	0.85	0.84
	F1-Score	0.86	0.91	**0.85**	0.91	0.65	0.85	0.84	0.84	0.83

Based on Figure 4 and Table 5, it can be seen that the use of adjusted parameters results in a significant improvement in performance for both model variants. IndoBERT-BASE-P2 with custom parameters achieved a macro F1-score of 0.8433 and an accuracy of 0.8443, an improvement from the baseline, which only achieved 0.7648 for F1 and 0.7546 for accuracy. This improvement indicates that adjusting parameters such as max length, learning rate, number of epochs, and warmup steps has a significant impact on the final performance of the model, without the need to change its underlying architecture.

A similar improvement was also observed in IndoBERT-LARGE-P2. The custom version of this model achieved a macro F1-score of 0.8406, up from 0.7915 in the baseline. However, despite the relatively high score, the improvement in the LARGE version is not as sharp as the improvement observed in the BASE version. This suggests that larger models like IndoBERT-LARGE-P2 may be more prone to overfitting, especially when trained on datasets that are not very large. As a result, efficiency and stability are more pronounced in the optimized BASE version.

Analysis of each emotion label shows that emotions such as 'love', 'fear', and 'sad' consistently achieve high F1-scores across all models, particularly in custom configurations. This suggests that the linguistic expressions of these emotions tend to be more explicit or follow specific patterns that can be

captured by the model. Conversely, the 'neutral' label presents its own challenges. Even in the best-performing model, such as IndoBERT-LARGE-P2 Custom, the F1 score for 'neutral' only reached 0.6534. This value is significantly lower than other emotions such as 'fear' (0.9110) or 'love' (0.9145). The low performance on the 'neutral' label aligns with previous studies findings that this emotion is semantically more ambiguous and tends to overlap with weaker emotions like 'sad' or 'joy'.

From the perspective of aggregate metrics, the nearly identical macro average and weighted average values in the IndoBERT-BASE-P2 Custom model (0.8433 and 0.8412, respectively) indicate that this model not only performs well on majority classes but also maintains stable performance on minority classes. This confirms that parameter tuning can help the model achieve better generalization across all emotion classes, even when the data distribution is imbalanced.

Overall, these results show that IndoBERT-BASE-P2 with tuned parameters is the most balanced model in terms of accuracy and efficiency. Although IndoBERT-LARGE-P2 offers competitive performance, its higher computational requirements and longer training time make it less suitable for real-world applications with resource constraints. Therefore, IndoBERT-BASE-P2 Custom is the optimal choice for implementing Indonesian language emotion classification systems, particularly in practical applications such as academic chatbots or text-based emotional support systems.

3.3 Model Performance Analysis

Across all settings, IndoBERT-BASE-P2 models consistently outperform in classifying Indonesian-language emotional texts. This superior performance is attributed to IndoBERT's pretraining on a large Indonesian corpus, making it more semantically aligned with the linguistic nuances present in the dataset [21].

The custom parameter settings resulted in a notable performance boost compared to the original settings used in the paper. For example, the IndoBERT-BASE-P2 (Custom Params**)** achieved a macro F1-score of 0.8433 and an accuracy of 0.8443, outperforming the version using original paper parameters, which only reached an F1-score of 0.7648 and an accuracy of 0.7546. This indicates that hyperparameter tuning has a significant impact on optimizing model performance, especially in emotion classification tasks involving subtle language features.

Interestingly, the IndoBERT-BASE-P2 variant slightly outperformed the IndoBERT-LARGE-P2 variant in macro F1-score across custom settings (0.8433 vs 0.8406), suggesting that the larger model may be prone to overfitting when trained on relatively small datasets. While the IndoBERT-LARGE-P2 variant did show improvements in individual class scores (e.g., fear: 0.9110 F1), this came at the cost of generalization on other labels, particularly the 'neutral' class.

3.4 Error Analysis by Class

Among all emotion classes, 'neutral' consistently received the lowest F1-scores across both model variants. For instance, in the best-performing setting IndoBERT-LARGE-P2 (Custom), the 'neutral' class only achieved an F1-score of 0.6534, which was significantly lower than other emotions, such as 'fear' (0.9110) or 'love' (0.9145). This result aligns with previous findings [4], indicating that 'neutral' tends to be a semantically ambiguous category, often overlapping with weakly expressed emotions such as 'sad' or 'joy'.

Another noteworthy observation is the clear improvement in the 'neutral' class (from the paper baseline, F1 = 0.65), which was not directly evaluated in the custom setting due to label adjustments. However, similar patterns are observed in classes like 'sad' and 'neutral', which are often challenging due to subtler linguistic cues or class imbalance.

3.5 Influence of Data Training Configuration

Performance improvement can also be attributed to data preprocessing and custom training parameters [22]. The use of stratified data splits and balanced batch sampling helped mitigate the effects of class imbalance, particularly in underrepresented emotions like 'love' and 'sad'. Moreover, reducing the learning rate and increasing the number of training epochs contributed to better convergence in

p-ISSN: 2528-1682
e-ISSN: 2527-9165

later training stages. This suggests that even with a pretrained language model, domain-specific fine-tuning strategies remain essential to achieve optimal performance. Hyperparameter tuning, balanced sampling, and detailed data cleaning are critical factors for enhancing results.

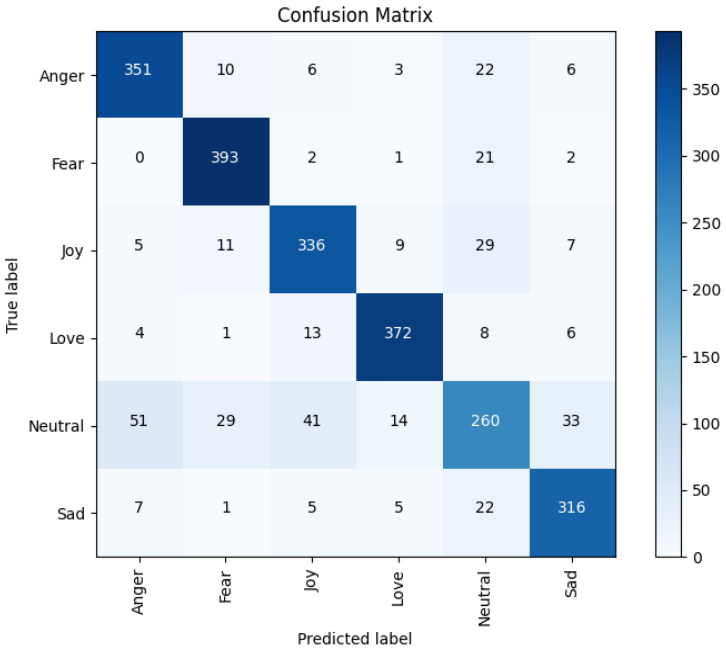

Figure 5. Confusion Matrix for Best Model

A confusion matrix in Figure 5, was also generated for the best-performing model IndoBERT-BASE-P2 (Custom), illustrating frequent misclassifications between 'neutral', 'sad', 'joy', and 'love', which further confirms the semantic overlap among these emotional categories.

4. CONCLUSION

This research successfully evaluated and improved the performance of transformer models for detecting emotions in Indonesian text, with a particular focus on IndoBERT-BASE-P2 and IndoBERT-LARGE-P2. Our main findings indicate that optimized training parameter settings, such as max_length, warmup_steps, and weight_decay, significantly improve the model's accuracy and macro F1-score. Specifically, the IndoBERT-BASE-P2 model with our customized configuration achieved a macro F1-score of 0.8433 and an accuracy of 0.8443, far surpassing the initial configuration reported in previous studies. This improvement underscores the crucial role of careful parameter tuning in maximizing the potential of pre-trained language models for detailed tasks such as emotion detection.

Additionally, our research highlights the importance of the data preparation stage. Unlike previous basic approaches, the addition of steps such as data cleaning, stopword removal, stemming, and especially class balancing using the SMOTE technique proved to be highly beneficial. This comprehensive data preparation improves the quality of text representation, enabling the model to capture more subtle emotional nuances more effectively, especially for minority emotion classes like 'love' and 'fear'. Although larger models like IndoBERT-LARGE-P2 show competitive results, the fine-tuned IndoBERT-BASE-P2 achieves comparable performance with significantly better computational efficiency, making it the optimal choice for real-world applications with limited resources. This work contributes to the field by providing a robust and efficient solution for Indonesian language emotion classification, demonstrating that the combination of proper data preparation and strategic model tuning can yield substantial performance improvements without adding architectural complexity.

Further research could explore multimodal data integration or delve into more detailed emotion detection in Indonesian language text.

ACKNOWLEDGEMENTS

This work was supported by the Institute for Research and Community Service [LPPM] of Universitas Dian Nuswantoro under Grant No. 005/A.38-04/UDN-09/I/2025

REFERENCES

[1] K. Mahor and A. K. Manjhvar, "Public Sentiment Assessment of Coronavirus-Specific Tweets using a Transformer-based BERT Classifier," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, IEEE, Oct. 2022, pp. 1559–1564. doi: 10.1109/ICECAA55415.2022.9936448.

[2] A. K. J, E. Cambria, and T. E. Trueman, "Transformer-Based Bidirectional Encoder Representations for Emotion Detection from Text," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Dec. 2021, pp. 1–6. doi: 10.1109/SSCI50451.2021.9660152.

[3] R. F. Reza, Muhmmad Thoriq, and Rd. Imam Saepul Millah, "Sentiment Analysis of Marketplace Review with Islamic Perspective using Fine-Tuning DistilBERT," *Khazanah Journal of Religion and Technology*, vol. 2, no. 2, pp. 45–54, Jan. 2025, doi: 10.15575/kjrt.v2i2.1118.

[4] S. Jayanthi and S. S. Arumugam, "Advancing Emotion Detection in a Text of Transformer-Based Models and Traditional Classifiers," in *2024 Asian Conference on Intelligent Technologies (ACOIT)*, IEEE, Sep. 2024, pp. 1–5. doi: 10.1109/ACOIT62457.2024.10939623.

[5] N. A. P. Masaling, R. R. Siswanto, and A. S. Girsang, "Indonesian Tweet Emotion Detection Using IndoBERT," in *2024 International Conference on Information Management and Technology (ICIMTech)*, IEEE, Aug. 2024, pp. 478–482. doi: 10.1109/ICIMTech63123.2024.10780847.

[6] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Recognizing Emotions from Texts Using an Ensemble of Transformer-Based Language Models," in *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, IEEE, Dec. 2021, pp. 161–164. doi: 10.1109/ICCWAMTIP53232.2021.9674102.

[7] K. Hulliyah, F. Rayyan, and N. S. A. A. Bakar, "Development Of A Chatbot For The Online Application Telegram Chat With An Approach To The Emotion Classification Text Using The Indobert-Lite Method," in *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, Oct. 2022, pp. 1–4. doi: 10.1109/ICORIS56080.2022.10031483.

[8] A. Das, M. M. Hoque, O. Sharif, M. A. A. Dewan, and N. Siddique, "TEmoX: Classification of Textual Emotion Using Ensemble of Transformers," *IEEE Access*, vol. 11, pp. 109803–109818, 2023, doi: 10.1109/ACCESS.2023.3319455.

[9] N. Mossad, Y. Mohamed, A. Fares, and A. B. Zaky, "Arabic text sentiment analysis and emotion classification using transformers," in *2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, IEEE, Dec. 2023, pp. 131–137. doi: 10.1109/JAC-ECC61002.2023.10479609.

[10] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 22, Mar. 2025, doi: 10.1007/s13278-025-01439-6.

[11] M. K. Anam, "Improved Performance of Hybrid GRU-BiLSTM for Detection Emotion on Twitter Dataset," *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 354–365, Jan. 2024, doi: 10.47738/jads.v6i1.459.

[12] A. de León Languré and M. Zareei, "Evaluating the Effect of Emotion Models on the Generalizability of Text Emotion Detection Systems," *IEEE Access*, vol. 12, pp. 70489–70500, 2024, doi: 10.1109/ACCESS.2024.3401203.

[13] Riccosan, K. E. Saputra, G. D. Pratama, and A. Chowanda, "Emotion dataset from Indonesian public opinion," *Data Brief*, vol. 43, p. 108465, Aug. 2022, doi: 10.1016/j.dib.2022.108465.

[14] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.

[15] Ajitesh Kumar, "Transformer Architecture Types: Explained with Examples," 2024.

[16] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences*, vol. 15, no. 1, p. 172, Dec. 2024, doi: 10.3390/app15010172.

[17] C. I. V and S. K J, "Text-Based Emotion Recognition Using Deep Learning," in *2024 Second International Conference on Advances in Information Technology (ICAIT)*, IEEE, Jul. 2024, pp. 1–7. doi: 10.1109/ICAIT61638.2024.10690782.

[18] V. K. Agbesi *et al.*, "Pre-Trained Transformer-Based Models for Text Classification Using Low-Resourced Ewe Language," *Systems*, vol. 12, no. 1, p. 1, Dec. 2023, doi: 10.3390/systems12010001.

[19] Arif Bijaksana Putra Negara, "The Influence Of Applying Stopword Removal And Smote On Indonesian Sentiment Classification," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 14, no. 03, pp. 172–185, Oct. 2025, doi: 10.24843/LKJITI.2023.v14.i03.p05.

[20] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, p. 122778, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.

[21] F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, "Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 6, pp. 748–757, Dec. 2024, doi: 10.29207/resti.v8i6.6100.

[22] Muhamad Ridwan and Ema Utami, "An Optimized Hyperparameter Tuning for Improved Hate Speech Detection with Multilayer Perceptron," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 4, pp. 525–534, Aug. 2024, doi: 10.29207/resti.v8i4.5949.