Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1172~1181
e-ISSN: 2548-6861

Performance Analysis of LSTM, GRU and IndoBERT Variants for Emotion Detection in Indonesian Text

Putri Innayah Mahmid [1]*, Nouval Trezandy Lapatta [2]*
* Informatics Engineering, Tadulako University
innayahptrrr@gmail.com [1], nouval@untad.ac.id [2]

Article Info

Article history:

Received 2025-12-11
Revised 2026-02-17
Accepted 2026-02-27

Keyword:

Attention Mechanism,
Gating Mechanism,
Emotion Detection.

ABSTRACT

This study evaluates gating mechanisms, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), in comparison with attention-based models utilizing IndoBERT variants (Base, Large, and Lite) for Indonesian emotion detection across six emotion labels. The evaluation examines accuracy, efficiency, and robustness using both in-distribution and out-of-distribution (OOD) datasets collected from social media. Statistical significance is assessed through confidence interval estimation and bootstrap paired tests, and a detailed error analysis is conducted to identify model limitations. The results indicate that IndoBERT Large achieves superior performance, with a Macro F1-Score of 80.05% and greater robustness to domain shifts, whereas gating models exhibit substantial performance degradation on unseen data. In contrast, GRU outperforms LSTM and achieves the lowest inference latency, with training times up to 131 times faster than IndoBERT Large. Statistical tests confirm that the performance gap between IndoBERT variants and RNN-based models is significant. These findings highlight a key trade-off: attention mechanisms provide state-of-the-art accuracy and robustness, while GRU offers a practical and efficient solution for resource-constrained settings.

I. INTRODUCTION

Emotions are a fundamental aspect of human life. These affective states influence decision-making processes and facilitate effective communication. Emotion detection, or emotion recognition, refers to the identification of an individual's specific emotion or emotional state, such as joy, sadness, or anger [1]. The application of emotion detection plays a significant role in decision-making across various disciplines. In the context of education, this technology can identify students' emotions through their feedback and interactions on online learning platforms, thereby revealing their emotional attitudes toward learning engagement. Such insights help educators tailor instructional approaches and potentially enhance learning quality [2]. In the business domain, these technologies enable improved customer segmentation and the provision of personalized experiences [3].

In computing, text-based emotion detection is a text classification task and is regarded as a key element of Natural Language Processing (NLP) [4], This approach has emerged as an integral component of NLP due to its ability to categorize text data into predefined classes. Text classification enables a wide range of applications, including topic categorization, spam filtering, and sentiment analysis [5].

For sequential text classification, various architectures have been developed. Early architectures, such as the Recurrent Neural Network (RNN), were susceptible to gradient vanishing. They were therefore unable to capture long-term dependencies [6]. To overcome this, models with gating mechanisms, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), were introduced [6]. These models have proven effective and have long stood as strong baselines for text classification tasks, including sentiment analysis and emotion detection [7], [8]. A more modern architecture is Bidirectional Encoder Representations from Transformers (BERT), which is based on a Transformer architecture and uses the attention mechanism [9]. This allows the model to process the entire

sequence at once and capture contextual relationships. The BERT model and its language-specific variants, such as IndoBERT, have achieved state-of-the-art (SOTA) results on various Indonesian NLP tasks [10]-[11].

Although Transformer-based models such as IndoBERT frequently yield superior accuracy metrics, prior research has largely overlooked computational considerations. Contemporary literature prioritizes maximizing metrics such as the F1 Score or accuracy, with limited attention to resource utilization and latency. In practical implementations, particularly on devices constrained by limited resources or that require real-time processing, the effectiveness of a model depends not only on predictive accuracy but also on computational efficiency. Conventional architectures, such as LSTM and GRU, although less complex, may offer a practical balance between processing speed and predictive accuracy for specific datasets.

This research not only aims for high accuracy but also conducts a comprehensive performance evaluation of five representative architectures: LSTM, GRU, IndoBERT, IndoBERT Lite, and IndoBERT Large. The evaluation encompasses multiple dimensions, including classification metrics, training speed, and inference latency. The analysis aims to provide empirical insights into the trade-offs between model complexity and operational efficiency, and to recommend the optimal architecture for emotion detection in Indonesian informal texts.

II. METHOD

This study uses a comparison method between LSTM, GRU, IndoBERT Base, IndoBERT Lite and IndoBERT Large for the task of emotion classification in Indonesian informal texts. This study aims to compare the performance and effectiveness of the five models. The flow of the research is illustrated with the flowchart in Figure 1.

A. Dataset

This study uses the "Emotion dataset from public opinion", which is publicly available on GitHub. The dataset consists of 7080 lines of Indonesian informal texts. These are classified into six labels: neutral (2001), joy (1275), anger (1130), sad (1003), fear (911), and love (760). The data has been preprocessed by lower-casing sentences, removing hashtags, and eliminating unused information such as mentions, URLs, emoticons, and non-emoticon symbols [12].

In addition to the primary dataset, a secondary dataset was incorporated specifically for generalization and out-of-distribution (OOD) testing. This dataset consists of 60 text samples manually collected via scraping from Facebook and Instagram, with a distribution of joy (20), anger (17), neutral (11), sad (10), fear (1), and love (1). Strictly excluded from the training and validation phases, this dataset serves as a benchmark for evaluating the model's resilience to domain shift and Cross-Platform Variance. Cross-Platform Variance refers to the distinct linguistic differences across platforms and the model's ability to handle variations in linguistic style and platform-specific contexts.

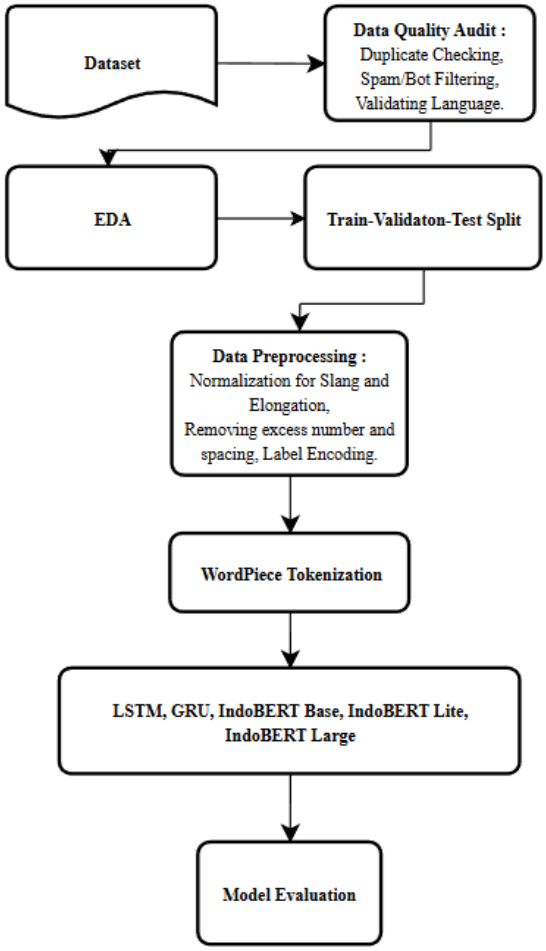

Figure 1. Research Flow

B. Data Quality Audit

Data from public opinion, such as social media, is likely to be noisy. Before this data is used for model training, a quality audit and data cleansing process is carried out. This process includes three main steps: checking for duplicate data, filtering out spam or bots, and validating language.

C. Exploratory Data Analysis (EDA)

The Analysis stage begins by examining the dataset structure. The dataset contains 7080 rows and two columns: "tweet" and "label" Both columns have object data types. There are no missing values.

Furthermore, The word count distribution shows that the average line length ranges from 5 to 20 words. The shortest line contains 1 word, while the longest has 59. Figure 2's histogram illustrates this distribution.

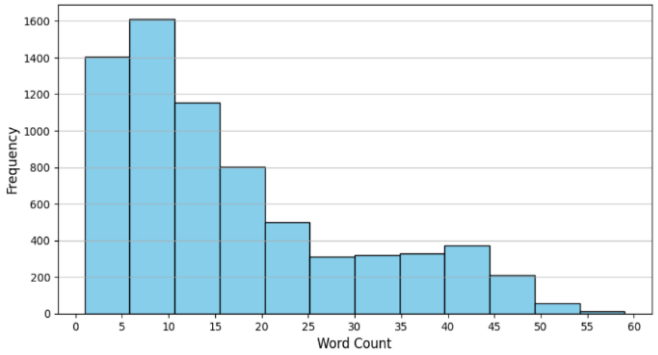

Figure 2. Distribution of Word Count in the Emotion Dataset from Public Opinion

Furthermore, the analysis of the label distribution shows that the dataset is imbalanced. A visualization of the class distribution is presented with a bar chart in Figure 3.

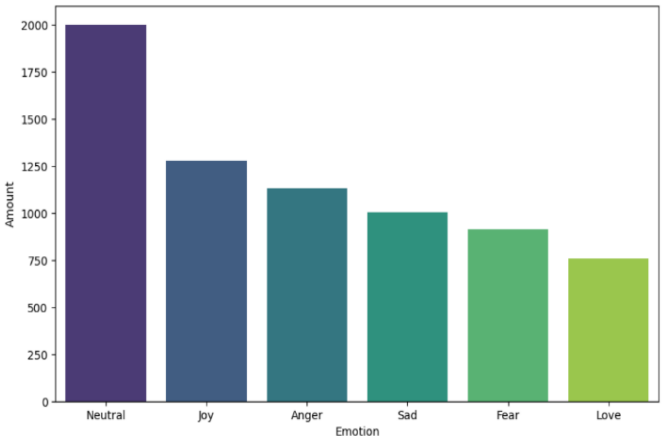

Figure 3. Distribution of Emotion Labels in the Emotion Dataset from Public Opinion

This imbalance is an important consideration for special handling at the modelling stage.

D. Data Pre-processing

Pre-processing data is a key step in transforming noisy, redundant, and incomplete raw data into clean, structured data ready for use by the model [13]. The The same Pre-Processing strategy is applied across the tested architectures. This ensures that resulting performance differences fully reflect the model's ability to extract features, rather than variations in input data quality. The data cleansing process is tailored to the input standards of context-sensitive Transformer-based models. The preprocessing stages include normalization for slang and elongation. They also involve removing excess numbers and spacing. After cleanup, all data is converted to numerical format using a WordPiece-based tokenization and label encoding.

E. Modelling and Hyperparameter Configurations

1) Class Imbalance handling with Class Weight

Analysis of label distribution at the Exploratory Data Analysis stage showed a significant class imbalance among the emotion labels. This imbalance can cause the model to become biased. As a result, the model may perform better at predicting the majority class but poorly on the minority class. To address this, the loss function is adjusted during training. Samples from the minority class are given higher weights, while those from the majority class receive lower weights. Thus, prediction errors in minority classes result in greater losses. This forces the model to focus more on learning patterns in those classes. This method was chosen to maintain the integrity of the original data. Weights are automatically calculated using the "balanced" formula in Scikit-learn, which makes the weights inversely proportional to class frequencies.

2) *Architecture and Hyperparameter Configuration*

In Parameters for each architecture type are tailored to its fundamental characteristics. This ensures that each model converges optimally. Configuration differences are applied to the Learning Rate and Optimizer. Pre-trained models such as IndoBERT require a lower learning rate and a warm-up phase to retain the weight of prior knowledge. In contrast, LSTM and GRU models trained from scratch require a more aggressive learning pace. To ensure comparability, external control variables such as Batch Size and Sequence Length are set uniformly across all models. Table I-III presents the architecture and hyperparameter configuration of each model.

TABLE I
MODEL ARCHITECTURES FOR LSTM AND GRU

Layer	**Spesification**	**Configuration**
Input Layer	Input Shape	256
Embedding Layer	Pre-trained model	indobert-large-p1
	Embedding Dim	1024
	Trainable	False
Recurrent Layer	LSTM/GRU	1 Layer, 128 Units
Regularization	Dropout	0.5
Dense Layer	Hidden Units	64 Units, ReLU
Output Layer	Dense Units	6 Units, Softmax

TABLE II
HYPERPARAMETER CONFIGURATION FOR LSTM AND GRU

Parameter	**Configuration**
Optimizer	Adam
Learning Rate	1e-3
Loss Function	catergorical_crossentropy
Batch Size	32
Epoch	30

Early Stop	Monitor : val_f1_score, Patience : 3

TABLE III
HYPERPARAMETER CONFIGURATION FOR INDOBERT VARIANTS

Parameter	**Configuration**
Optimizer	AdamW
Learning Rate	2e-5
Loss Function	CrossEntropyLoss
Batch Size	32
Epoch	30
Early Stop	Monitor : eval_f1_score, Patience : 3
Max Sequence Length	256

F. Long Short-Term Memory (LSTM)

As mentioned earlier, LSTM is designed to address the vanishing gradients problem that often occurs in conventional RNNs. This makes LSTMs more effective at modeling long-term dependencies. With its ability to process sequential data and store information from previous steps, LSTM is well-suited for tasks involving long-term dependencies [6]. LSTMs consist of memory cells that allow for long-term storage of information, enabling the network to capture long-term dependencies in sequential data. They also include three gate mechanisms: an input gate that regulates the addition of new information to the memory cell, a forget gate that determines which information from the previous cell state to retain or discard, and an output gate that controls the information passed on in the next time step [14].

G. Gated Recurrent Unit (GRU)

GRU is a popular model for NLP tasks because it offers performance comparable to LSTM but with a simpler architecture. It is designed to enhance LSTM performance through two main gate mechanisms. The update gate combines the input and forget gate functions found in LSTM. The reset gate regulates how much information from previous memory will be forgotten. This simpler design makes the GRU more computationally efficient [15].

H. IndoBERT

IndoBERT is a pre-trained Transformer model. It has been specially trained on the Indonesian corpus [16], giving it a deep understanding of grammar, semantics, and contextual nuances in Indonesian. This model comes in several major variants, each offering a trade-off between accuracy and computational efficiency. The Base variant is a standard version that balances performance and efficiency. It is much smaller in size than the Large variant. This makes it faster to train and more resource-efficient, while still delivering powerful performance. The Lite variant is the smallest and most efficient option. It is specifically designed for high inference speeds and minimal memory usage. The Large variant has a deeper architecture and more parameters. These models can capture the most complex and accurate feature representations, but require the most computational resources and training time. IndoBERT works with an attention mechanism. Unlike previous models, this mechanism directly models the relationship between distant words. The representation of each word is influenced by the overall content of the sentence, not just nearby words [9]. This results in a two-way understanding of context, in contrast to the gating mechanism, which processes sequences in only one direction.

I. Evaluation Metrics

The The purpose of model evaluation is to assess the quality of the model. This includes its accuracy, ability to represent comprehensively, relevant similarity, suitability, and other forms [17]. The model evaluation used a confusion matrix. This matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These components form the basis for the following evaluation metrics:

1) Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2) Precision

$$\text{Precision} = \frac{TP}{TP+FP}$$

3) Recall

$$\text{Recall} = \frac{TP}{TP+FN}$$

4) F1-Score

$$\text{F1-Score} = 2\times \frac{Recall \times Precission}{Recall + Precission}$$

J. Error Analysis

To thoroughly assess the model's limitations, an error analysis procedure was employed, comprising the following stages:

1) Prediction Distribution Mapping

A confusion matrix was used to map classification success across all emotion classes for each tested architecture. This process aimed to identify the classes most challenging for the models.

2) Representative Sample Analysis

A manual inspection was conducted on test samples that exhibited systematic misclassification across all models. These samples were selected to identify technical causes of failure, including ambiguity between emotionless text or neutral text containing emotional keywords.

3) Identification of Linguistic Phenomena

Prediction errors were categorized according to linguistic phenomena, including syntactic complexity and semantic ambiguity.

K. Statistical Significance Test

Statistical significance analysis is performed to ensure that the differences in performance between models are truly

significant and not caused by random variations in the test data. Two bootstrap-based methods are used: Confidence Interval Estimation to assess the performance range of each model, and Bootstrap Paired Test for paired comparisons.

1) Confidence Interval Estimation

CI estimation aims to provide the most likely range of Macro F1-Score performance values for each model. This method uses Bootstrap Resampling with 1000 iterations. This process involves random sampling with replacement of test data 1000 times to form a 1000 bootstrap dataset. The F1-score is then calculated on each of these datasets. This distribution of 1000 F1-Scores is used to determine the 95% Confidence Interval. A 95% CI indicates that we are 95% sure that the actual F1-Score model is within that range.

2) Bootstrap Paired Test

This method directly tests whether the difference in performance between two models (e.g., Model A vs. Model B) is statistically significant. Unlike comparing individual CIs, this test is much more accurate for comparison. The process is similar, but instead of calculating the F1-Score, the metric is the F1-Score difference (Model A Score - Model B Score) across 1000 bootstrap datasets. This results in a distribution of the score difference. From this distribution, the 95% confidence interval for the difference is calculated. The interpretation is as follows: If this CI range does not include a value of zero (0), then the difference is statistically significant ($p < 0.05$). Conversely, if this CI range includes a value of zero (0), it means that the difference is Not Statistically Significant ($p >= 0.05$).

L. Generalization Test

In addition to standard evaluation using a test set from the main dataset, this study also applies a Generalization Test with a secondary dataset to measure the model's resilience to data outside the training distribution. This test aims to assess the consistency of the model's performance across different language styles and contexts. The generalization test procedure includes the following stages:

1) Secondary Data Collection: Using additional datasets that have different linguistic characteristics from the training dataset.
2) Pre-Processing: Secondary data is processed with the same preprocessing flow as the trained data to maintain input consistency.
3) Labelling: The data is then manually annotated by the researcher to obtain a label.
4) Performance Evaluation: All trained models are tested on a sample of the secondary data. The accuracy is calculated to measure the performance drop compared to the primary dataset.

III. Result and Discussion

This chapter presents a comprehensive performance analysis of Gating and Attention-based models for emotion detection in Indonesian texts. The evaluation focused on five model architectures. LSTM and GRU represent Recurrent Neural Networks, while three IndoBERT variants (Base, Lite, Large) represent Transformers. The test was carried out in stages, starting with evaluation on the primary test set. It continued with a generalization test using secondary datasets to measure the model's resilience to data outside the training distribution. All tests use standard metrics: Accuracy, Precision, Recall, and F1-Score. To ensure valid results, statistical analysis using Confidence Interval Estimation and a Bootstrap Paired Test was performed. These methods assess the stability and significance of performance differences between models.

A. Model Performance Analysis

Testing on the test set, which comprises 15% of the total data, evaluates model performance using both classification metrics and computational efficiency measures. These measures include training time and inference latency. Table II summarizes the performance of the five models tested.

TABLE II
Model Performance

Model	Metric	Test Set (15%)	Training Time	Latency 128 Samples
LSTM	Accuracy	71.97%	19.93 s	80.73 ms
	Precision	72.61%		
	Recall	73.85%		
	F1-Score	72.69%		
GRU	Accuracy	72.93%	13.92 s	74.81 ms
	Precision	73.35%		
	Recall	75.61%		
	F1-Score	73.70%		
Indo-BERT Base	Accuracy	76.36%	973.52 s	197.41 ms
	Precision	77.41%		
	Recall	77.83%		
	F1-Score	77.26%		
Indo-BERT Lite	Accuracy	76.26%	713.85 s	194.23 ms
	Precision	76.99%		
	Recall	78.18%		
	F1-Score	77.39%		
Indo-BERT Large	Accuracy	79.31%	1834.34 s	680.88 ms
	Precision	79.50%		
	Recall	81.23%		
	F1-Score	80.05%		

Testing on the test set showed that the Transformer-based architecture consistently excelled at capturing the nuances of emotion in informal Indonesian texts. IndoBERT Large recorded the highest performance across all evaluation metrics, with a Macro F1-Score of 80.05% and an Accuracy

of 79.31%. This analysis identified a significant performance gap of about 7% between the IndoBERT variant and the RNN-based models (LSTM and GRU), even though both RNN models used word embeddings from IndoBERT Large to ensure a fair comparison. These findings suggest that the performance limitations of classical models stem primarily from the intrinsic limitations of gating architectures that process text sequentially, rather than from the quality of word representation. On the other hand, the Attention mechanism in IndoBERT has been shown to be more effective at globally mapping contextual relationships between words, thereby distinguishing the emotional ambiguity that often appears in complex sentences.

In the context of RNN-based models, GRUs show superior performance to LSTMs in both accuracy and efficiency. Specifically, GRU achieved an F1-Score of 73.70%, surpassing LSTM's 72.69%. Furthermore, this advantage is evident in efficiency: GRU completes training in 13.92 seconds, about 30% faster than LSTM at 19.93 seconds, and achieves the fastest inference latency at 74.81 ms. This GRU advantage in the current experiment can be attributed to its simpler architectural structure (i.e., two gates) compared to the LSTM's three gates. This allows for faster convergence and helps reduce the risk of overfitting on medium-sized datasets.

The IndoBERT variant highlights the parameter efficiency of IndoBERT Lite. Despite a much smaller parameter count, IndoBERT Lite achieves an F1-Score of 77.39%, statistically equivalent to or slightly higher than IndoBERT Base (77.26%). However, time analysis shows that the inference latency of IndoBERT Lite (194.23 ms) does not differ significantly from that of IndoBERT Base (197.41 ms). These findings show that parameter reduction through factorized embedding parameterization in the ALBERT architecture underlying IndoBERT Lite effectively conserves memory, but does not linearly speed up execution time compared to standard BERT architectures.

The results demonstrate a significant correlation between architectural complexity and model generalization capability. IndoBERT Large is particularly well-suited for accuracy-centric real-world applications in which precision is prioritized over speed, such as government sentiment analysis systems that monitor public opinion on critical policies. In these contexts, processing generally occurs on server-side platforms equipped with robust GPU infrastructure, making the latency of 680.88 ms and the extended training time of 1,834 seconds acceptable trade-offs for achieving optimal prediction quality.The trade-off analysis reveals that the accuracy gains offered by IndoBERT are often outweighed by the associated computational costs in specific scenarios. For instance, on low-specification mobile devices or in real-time content moderation tasks that require processing thousands of texts per second, IndoBERT Large proves inefficient. Its 680.88 ms latency leads to significant performance delays and increased battery consumption. In these cases, GRU represents a more practical alternative; while its accuracy is approximately 6.35% lower than IndoBERT Large, it provides inference that is nine times faster (74.81 ms) and training that is 131 times more efficient (13.92 seconds).

Therefore, model selection must be tailored to the requirements of the intended infrastructure. For applications such as interactive chatbots or edge device data streams, GRU or IndoBERT Lite, which offers a more favorable balance between performance (77.39% F1-Score) and latency (194.23 ms) compared to the Large variant, constitutes a more technically sustainable choice. Conversely, IndoBERT Large should be reserved for controlled server environments where accuracy is the primary consideration and device resource limitations are negligible.

B. Confussion Matrix

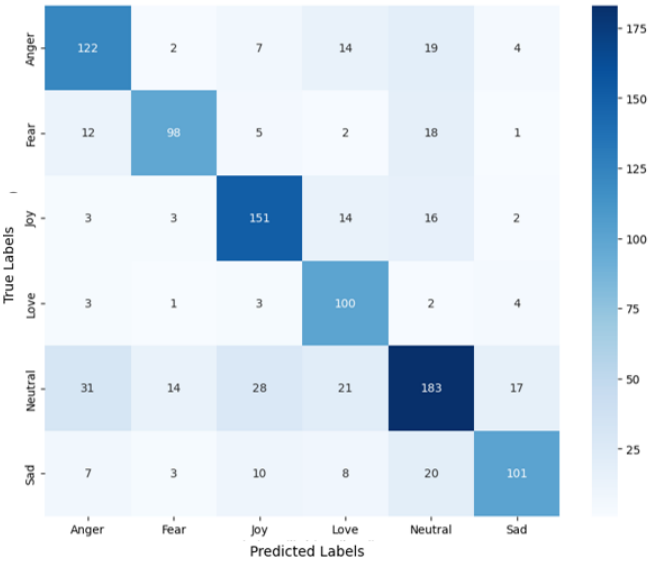

Figure 4. Confussion Matrix LSTM

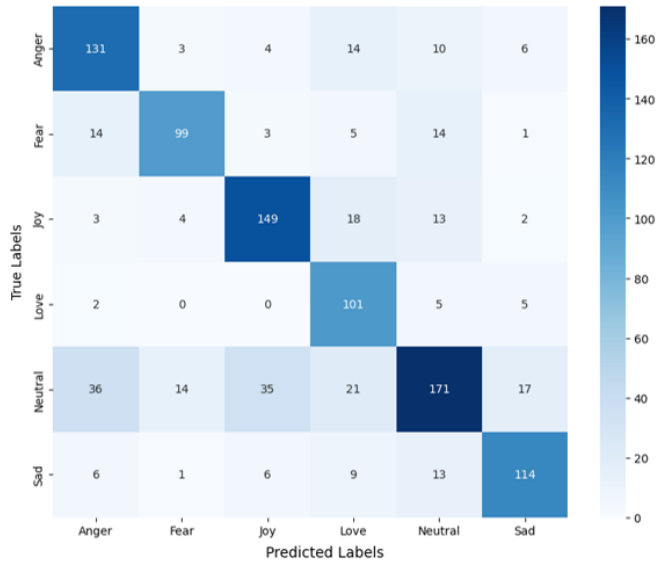

Figure 5. Confussion Matrix GRU

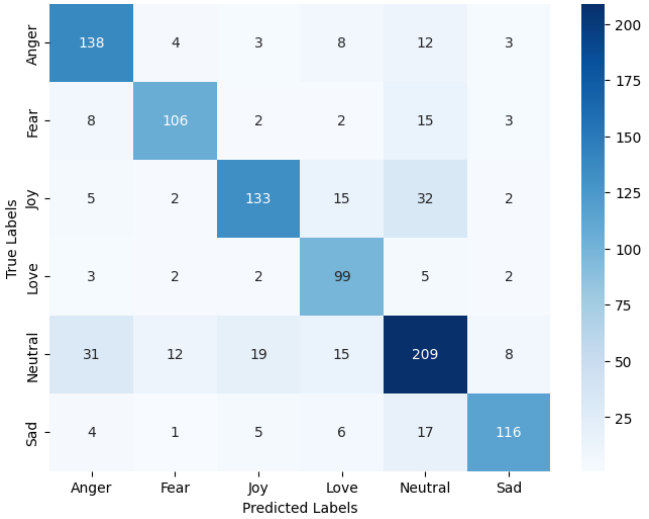

Figure 6. Confussion Matrix IndoBERT

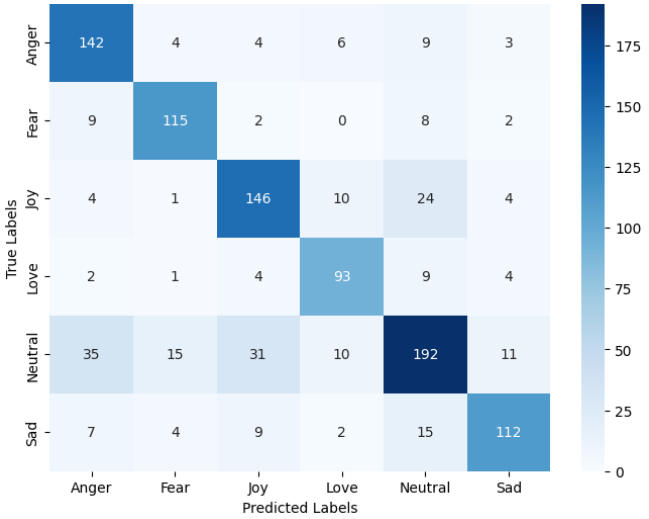

Figure 7. Confussion Matrix IndoBERT Lite

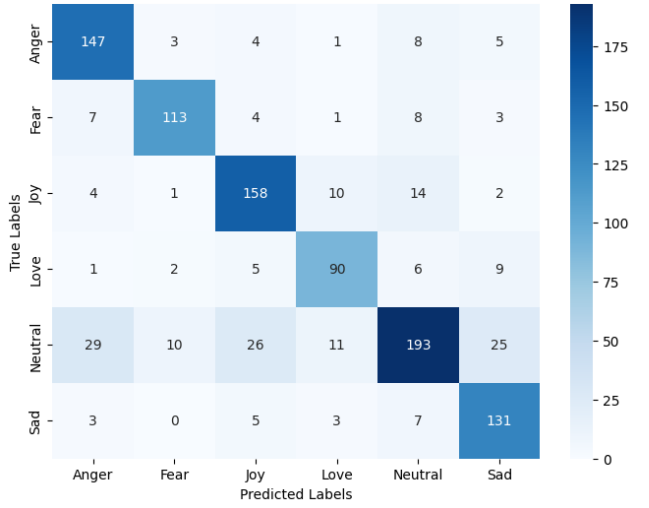

Figure 8. Confussion Matrix Indobert Large

The Confusion Matrix identifies the specific performance characteristics of each architecture. Notably, RNN-based models, such as LSTM and GRU, show significant limitations in capturing neutral contexts, as evidenced by high misclassification rates in the Neutral class, which is often confused with other emotions such as Anger and Joy. Meanwhile, the performance visualization for IndoBERT Large shows higher predictive stability. these findings indicate that the Attention mechanism can overcome semantic ambiguity in implicit sentences, particularly between the Sad and Neutral classes, a major point of failure in sequential models.

C. Error Analysis

This section provides a comprehensive analysis of the model's failure to accurately classify emotions.

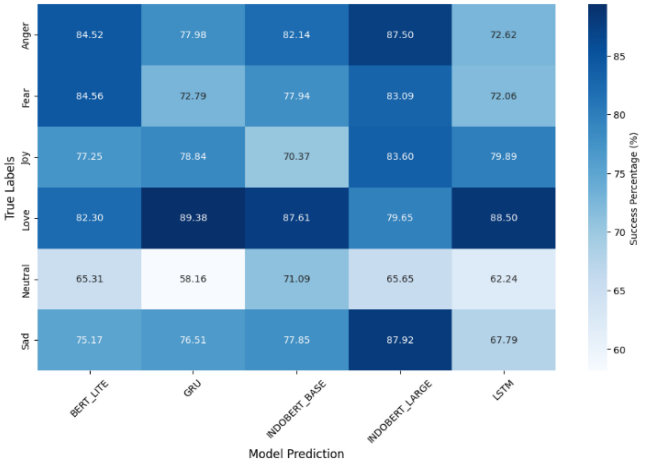

Figure 9. Model Success Rate per Emotion Class

Figure 9 demonstrates significant variation in classification success rates across different emotions. The Neutral class is the most challenging to detect, with the GRU model exhibiting the highest failure rate at 41.84%. This difficulty arises from the ambiguity between emotionless text and neutral text containing emotional keywords. Architecturally, the Gating model achieves the highest accuracy in the Love class, likely due to the consistent sequential structure of affectionate expressions. Conversely, the IndoBERT model achieves superior performance in the Anger class, indicating that the attention mechanism is more effective at capturing the syntactic complexity of these emotions.

To investigate the technical causes of classification failures, representative test samples exhibiting systematic errors across all models were selected.

TABLE III
MISCLASSIFICATION SAMPLES BASED ON LINGUISTIC PHENOMENA

Line	Text	Ground Truth	Predicted
501	"marah marah mlu, hamil y"	Neutral	Anger

509	“enak banget tapi hanya bisa melihat”	Sad	Joy
397	“dih bego kok bangga”	Neutral	Anger
407	“ya kali baru kaget sekarang”	Neutral	Fear

Table III identifies several key factors contributing to classification failures:

1) Lexical Bias: The model demonstrates high sensitivity to the repetition of specific emotional words. In line 501, the repeated use of the word “marah” automatically triggers an Anger prediction, even though the sentence is pragmatically a casual, neutral question.
2) Emotional Shift: The use of the contrastive conjunction “tapi” in line 509 poses a significant challenge. The model frequently focuses on the initial positive sentiment “enak banget” and overlooks subsequent expressions of disappointment, leading to critical errors in final label assignment.
3) Slang and Idiomatic Complexity: The use of Indonesian idioms such as “ya kali” (line 407) and sarcastic expressions (line 397) is challenging for the model to interpret, as their semantic meanings differ significantly from their literal word meanings.

D. Statistical Significance Test

The first step in statistical analysis is to estimate a Confidence Interval of 95% for the F1-Macro Score of each model.

TABLE IV
MACRO F1 SCORE MODEL

Model	F1-Score	95% CI
LSTM	0.72	[0.7001, 0.7553]
GRU	0.73	[0.7110, 0.7646]
IndoBERT	0.77	[0.7480, 0.7965]
IndoBERT Lite	0.77	[0.7475, 0.8005]
IndoBERT Large	0.80	[0.7756, 0.8235]

Table IV shows that IndoBERT Large not only has the highest average F1-Score (0.80), but also the highest overall confidence interval (CI) range [0.7756, 0.8235]. In addition, the CI ranges of IndoBERT Lite [0.7475, 0.8005] and IndoBERT Base [0.7480, 0.7965] overlap, indicating that the performance of the two models is statistically equivalent. On the other hand, the CI ranges of LSTM and GRU models are lower, indicating a consistent performance gap compared to the IndoBERT family.

To build on these findings and verify previous observations more rigorously, hypothesis tests were conducted using the Bootstrap Paired Test. The results of the Bootstrap Paired Test Significance Test are presented in Table V.

TABLE V
BOOTSTRAP PAIRED TEST SIGNIFICANCE TEST RESULT

Model	Score Difference	95% CI	p-value
IndoBERT Large - GRU	+0.0634	[+0.0383, +0.0876]	Significant
IndoBERT Large - LSTM	+0.0736	[+0.0464, +0.0999]	Significant
IndoBERT Large - IndoBERT Base	+0.0279	[+0.0034, +0.0509]	Significant
IndoBERT Large - IndoBERT Lite	+0.0266	[+0.0030, +0.0506]	Significant
IndoBERT Base - IndoBERT Lite	-0.0013	[-0.0234, +0.0211]	Not Significant
GRU - LSTM	+0.0102	[-0.0081, +0.0297]	Not Significant

Findings Analysis of the results of statistical tests shows several important findings. Firstly, IndoBERT Large has proven to be significantly superior to all comparison models. The all-positive Confidence Interval range confirms that this advantage reflects the ability of IndoBERT Large’s architecture to capture complex emotional nuances, rather than mere chance. Furthermore, when comparing IndoBERT Base to Lite, the results are inconclusive, with a CI range of difference [-0.0234, +0.0211] that includes zero. These results show that IndoBERT Lite matches the performance of the Base variant despite having a much smaller parameter count, making it an efficient alternative. In a similar vein, a similar pattern was found in the classical model, where the difference between GRU and LSTM was not statistically significant. Although the GRU recorded a higher average score, the difference was insufficient to establish absolute superiority in prediction accuracy. However, GRU remains a more pragmatic option when considering time efficiency metrics.

E. Generalization and Model Robustness Test

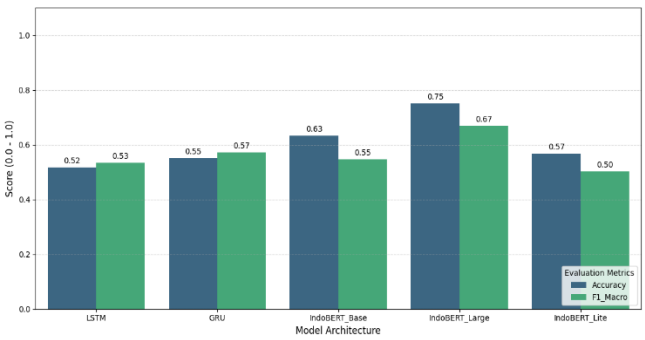

Figure 10. Test Result Generalization on Secondary Data

The robustness test was performed using a secondary dataset with a highly unbalanced class distribution, where the Joy, Neutral, Sad and Anger labels dominated most of the data, while the Fear and Love labels had very limited representation. The challenging conditions of this dataset demonstrate high performance stability for IndoBERT Large, which maintained an Accuracy of 75%. The model’s

ability to perform well across dominant classes despite language style shifts demonstrates that large parameter capacity Transformer architectures have optimal semantic generalization capabilities and are driven by an understanding of universal emotional context, rather than simply memorizing patterns from the main training dataset.

In contrast to IndoBERT Large, the sequential architecture reveals significant performance vulnerabilities to changes in data distribution. LSTM experienced the greatest performance degradation, with accuracy falling to 51.67% and prediction error rates reaching 29 out of 60 samples, indicating the model's failure to distinguish the emotional nuances of the new text due to strong bias towards the majority class or an inability to capture unfamiliar contexts. Meanwhile, GRU showed unique performance characteristics, achieving 55% accuracy but a higher F1-Score of 61.47%. This anomaly indicates that the inequality in the number of labels, especially Love and Fear, is likely to penalize the average score, the GRU has a better sensitivity in detecting minority classes such as Fear and Love than the LSTM. The drastic decline in performance in these two RNN models confirms the presence of overfitting symptoms in the training data, limiting their generalization capabilities.

Building on these findings, in the small model, the performance of IndoBERT Base and Lite was recorded at 63.33% and 56.67%, respectively. An important finding in this analysis is that, despite being a compressed model and facing uneven data challenges, its architecture still maintains more stable generalizations than LSTMs. This confirms that the attention mechanism in the Transformer family provides a more reliable foundation for language understanding for real-world applications with high linguistic variability.

IV. CONCLUSION

The performance analysis of a Gating Mechanism (LSTM, GRU) and an Attention Mechanism (IndoBERT variant) based architecture for emotion detection in Indonesian informal texts showed that the Attention mechanism in IndoBERT was significantly superior to the gating mechanism in RNNs in capturing global semantic context. This advantage is reflected in IndoBERT Large, which achieved the highest and most stable performance across both internal and generalization tests. Moreover, IndoBERT Large also demonstrated a key ability to distinguish ambiguous emotions, particularly between the Neutral class and other implicit emotions, where RNN-based models often suffer from prediction failures. The robustness of this model was further confirmed through generalization tests on secondary datasets, where various IndoBERT variants maintained high accuracy under domain shifts or changes in language style, whereas LSTM and GRU models experienced significant performance degradation due to overfitting. Furthermore, the analysis highlights a clear trade-off between predictive performance and computational costs. IndoBERT Large requires the largest computing resources to achieve maximum accuracy, while GRU is the most efficient architecture, surpassing LSTM in training and inference speeds. This makes GRU a suitable solution for resource-constrained environments, albeit with a tolerance for declining accuracy. In addition, IndoBERT Lite shows that parameter reduction does not significantly compromise generalization, making it a balanced alternative between complex BERT families and less accurate RNNs. Despite these findings, several limitations should be acknowledged to ensure scientific transparency. First, the dataset remains relatively small, which may limit the model's exposure to broader linguistic variation. Second, there is potential domain bias because the primary data are sourced from specific social media environments. Third, the study relies predominantly on particular IndoBERT variants, which may not fully capture the ongoing development of Indonesian Transformer models. Given these findings and limitations, future research should evaluate this approach using a more diverse, multi-domain dataset to ensure consistent model performance and generalizability. Additionally, investigating hybrid models that integrate the efficiency of gating mechanisms with the deep contextual capabilities of attention mechanisms represents a promising direction for balancing accuracy and efficiency. Further advancements may include exploring IndoBERT distillation or quantization techniques to reduce latency, as well as incorporating multilingual or code-mixed Indonesian texts to enhance the model's robustness in addressing the complexity of contemporary digital communication.

REFERENCES

[1] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, Dec. 2021, doi: 10.1007/S13278-021-00776-6.

[2] C. Wang, "Emotion Recognition of College Students' Online Learning Engagement Based on Deep Learning," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 6, pp. 110–110, 2022, doi: 10.3991/ijet.v17i06.30019.

[3] S. V. Oprea and A. Bâra, "Extracting Emotions from Customer Reviews Using Text Mining, Large Language Models and Fine-Tuning Strategies," *J. Theor. Appl. Electron. Commer. Res. 2025, Vol. 20, Page 221*, vol. 20, no. 3, p. 221, Sep. 2025, doi: 10.3390/JTAER20030221.

[4] Winda Kurnia Sari, D. P. Rini, Reza Firsandaya Malik, and Iman Saladin B. Azhar, "Multilabel Text Classification in News Articles Using Long-Term Memory with Word2Vec," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 276–285, Apr. 2020, doi: 10.29207/RESTI.V4I2.1655.

[5] U. Mahesh, A. Prof, R. Jr, R. Kumar, S. Vm, and U. Bd, "Text Classification using RNN," *Int. J. Eng. Res. Technol.*, vol. 14, no. 5, May 2025, doi: 10.17577/IJERTV14IS050314.

[6] S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 5, p. 102068, Jun. 2024, doi: 10.1016/J.JKSUCI.2024.102068.

[7] I. G. P. M. Yusadara and I. G. A. D. Saryanti, "Classification of User Expressions on Social Media Using LSTM and GRU Models," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 14, no. 1, pp. 49–54, Jan. 2025, doi: 10.32736/sisfokom.v14i1.2370.

[8] K. G. T. Kumar, R. Anoop, S. G. Koolagudi, T. Rao, and A.

Kodipalli, "Stratification of Depressed and Non-Depressed Texts from Social Media using LSTM and its Variants," *Procedia Comput. Sci.*, vol. 235, pp. 1353–1363, Jan. 2024, doi: 10.1016/J.PROCS.2024.04.127.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Oct. 22, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805

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

[11] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.

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

[13] "Data Preprocessing - an overview | ScienceDirect Topics." Accessed: Oct. 10, 2025. [Online]. Available: https://www.sciencedirect.com/topics/engineering/data-preprocessing

[14] H. Chung and K. S. Shin, "Genetic algorithm-optimized long short-term memory network for stock market prediction," *Sustain.*, vol. 10, no. 10, 2018, doi: 10.3390/su10103765.

[15] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, p. 102946, Dec. 2024, doi: 10.1016/J.MEX.2024.102946.

[16] Wildan Amru Hidayat and V. R. S. Nastiti, "Perbandingan Kinerja Pre-Trained Indobert-Base Dan Indobert-Lite Pada Klasifikasi Sentimen Ulasan Tiktok Tokopedia Seller Center Dengan Model Indobert," *JSiI (Jurnal Sist. Informasi)*, vol. 11, no. 2, pp. 13–20, Sep. 2024, doi: 10.30656/JSII.V11I2.9168.

[17] W. S. Parker, "Model Evaluation," *Routledge Handb. Philos. Sci. Model.*, pp. 208–219, Jan. 2024, doi: 10.4324/9781003205647-19.