SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 219-228, DOI: 10.59190/stc.v6i3.372

An IndoBERT-based framework for emotion classification in Indonesian song lyrics

Agustar Alfonso[1], Fitri Insani[1], Okfalisa[1,*], Muhammad Fikry[1], Fitra Kurnia[1], Sri Wahyuni[2]
[1]Department of Informatics Engineering, UIN Sultan Syarif Kasim, Pekanbaru 28293, Indonesia
[2]Department of Psychology, UIN Sultan Syarif Kasim, Pekanbaru 28293, Indonesia

ABSTRACT

Emotion classification in song lyrics represented a significant research area within natural language processing, yet studies targeting Indonesian-language lyrics remained scarce due to the limited availability of labeled datasets and the absence of domain-specific models. This study developed and evaluated an emotion classification model for Indonesian song lyrics using fine-tuned IndoBERT-base-p2, a transformer-based language model pre-trained on a large Indonesian corpus. A dataset of 1,025 labeled lyric entries was compiled from Kaggle, Genius, and KapanLagi, covering four emotion categories: joy, sadness, fear, and anger. Preprocessing encompassed duplicate removal, case folding, structural marker removal, and non-alphabetic character cleaning. Nine fine-tuning experiments were conducted by systematically varying learning rate and dropout rate, with early stopping applied based on validation loss. The optimal configuration employed a learning rate of 3×10^{-5} and a dropout rate of 0.1, achieving 75.73% accuracy and 75.85% macro-averaged F1-score on the held-out test set. Joy and anger were classified most reliably, attaining F1-scores of 82.76% and 76.47% respectively, while sadness presented the greatest challenge, exhibiting the lowest precision of 64.10% alongside a recall of 80.65%, indicating a systematic tendency of the model to over-predict this class. These findings demonstrated that IndoBERT-base-p2, when fine-tuned with appropriate hyperparameter configuration, served as an effective approach for domain-specific emotion classification in Indonesian song lyrics.

ARTICLE INFO

Article history:

Received May 26, 2026
Revised Jun 7, 2026
Accepted Jun 8, 2026

Keywords:

Emotion Classification
Fine-Tuning
IndoBERT
Song Lyrics
Transformer Model

* **Corresponding Author**
E-mail address: okfalisa@uin-suska.ac.id

1. INTRODUCTION

Music serves as an expressive medium in which melody and lyrics, reflecting two distinct human cognitive abilities, are typically combined to convey emotions [1]. Musical characteristics such as tempo and melody play a key role in shaping listeners' emotional responses, where fast-paced and upbeat music tends to elevate mood and energy, while slow and melancholic music is more likely to induce reflection or a sense of calm [2, 3]. Beyond acoustic elements, song lyrics have also been shown to play an important role in the personal meaning of a song, enabling listeners to work through their emotions and potentially view them in a new light [4]. The richness of these emotional dimensions has motivated growing interest in developing computational methods capable of automatically identifying the mood embedded in music.

The recognition and classification of emotion in music, known as Music Emotion Recognition (MER), has become an active area of research in recent years [5]. While early MER studies focused predominantly on acoustic features, growing evidence suggests that lyrics carry rich affective connotations that can be effectively mined using advanced Natural Language Processing (NLP) techniques [6]. However, prior approaches to lyrics-based MER have largely relied on word

embedding methods such as Word2Vec, GloVe, and FastText to represent lyrical content [7]. The advancement of transformer-based models, particularly BERT and its variants, has further enabled more accurate and context-aware emotion classification, outperforming such traditional approaches [8].

Progress in this area, however, has been largely limited to English-language data. Research on emotion classification for Indonesian song lyrics remains sparse, with existing studies still relying on conventional machine learning approaches such as Support Vector Machine (SVM) with Particle Swarm Optimization (PSO) rather than transformer-based models [9]. While Indonesian text emotion classification using IndoBERT has been explored across multiple text domains [10-12], no prior study has specifically fine-tuned IndoBERT on Indonesian song lyrics as the primary domain. This represents a critical and unaddressed research gap that the present study aims to fill.

Beyond the choice of model architecture, hyperparameter configuration plays a critical role in determining classification performance. Systematic tuning of parameters such as learning rate and regularization has been empirically demonstrated to yield significant improvements in accuracy, precision, recall, and F1-score across various text classification settings [13-15]. In the context of fine-tuning transformer-based models, this is particularly relevant, as pre-trained weights are sensitive to gradient updates during downstream adaptation, making the selection of appropriate hyperparameter values essential for achieving optimal generalization.

This study addresses that gap by implementing and evaluating a fine-tuned IndoBERT model for Indonesian song lyric emotion classification across four primary categories: joy, sadness, fear and anger. The objectives are: (1) to develop an IndoBERT-based classification model for Indonesian song lyrics, and (2) to evaluate and compare model performance across four emotion categories, namely joy, sadness, fear, and anger, using standard classification metrics including accuracy, precision, recall, and F1-score.

2. RESEARCH METHODS

This study follows a systematic pipeline consisting of five stages: data collection, text preprocessing, emotion labeling, IndoBERT fine-tuning, and model evaluation. Figure 1 illustrates the overall research workflow.

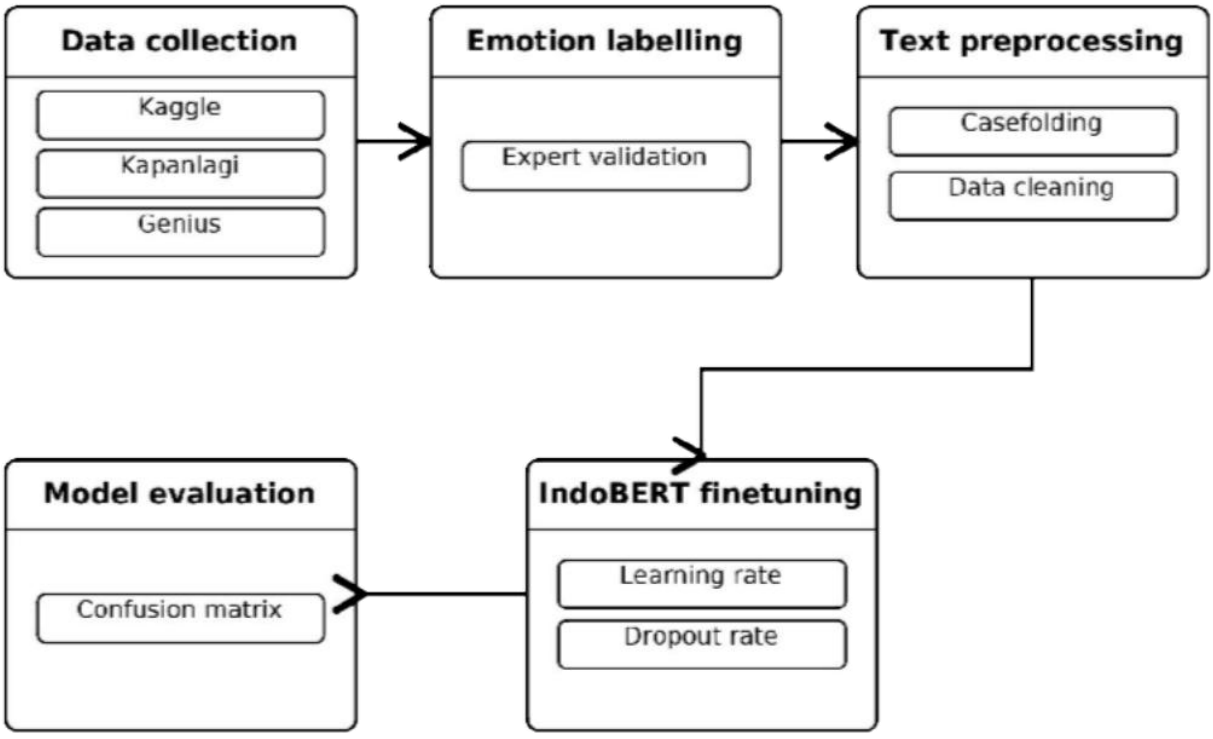

Figure 1. Research workflow.

2.1. Data Collection

The dataset was compiled from three primary sources. The first is the Indonesian Song Lyric Emotion Dataset obtained from Kaggle [16], which contains Indonesian song lyrics with pre-assigned emotion labels. The second source is Genius, from which additional lyrics were collected via the official Genius API. The third source is KapanLagi, a popular Indonesian lyric platform, from which lyrics were gathered through a custom web scraping pipeline using BeautifulSoup and httpx, targeting song pages structured under the /artis/{artist}/{title}/ URL pattern. Scraped data were stored in CSV format and normalized to ensure structural consistency across all three sources. The combined dataset comprises 1,031 lyric entries distributed across four emotion categories, as summarized in Table 1.

Table 1. Dataset distribution by source and emotion class.

Platform	Joy	Sad	Anger	Fear	Total
Kaggle	130	151	47	33	361
Genius	173	149	90	63	475
Kapanlagi	1	2	80	112	195
Total	304	302	217	208	1031

2.2. Emotion Labeling

This study adopts four emotion categories: joy, sadness, fear, and anger. Lyrics sourced from the Kaggle dataset were initially assigned labels based on the original dataset annotations. All labels, regardless of source, were subject to revision if identified as misaligned during expert validation. For lyrics collected through web scraping, manual annotation was performed by the researcher, with each lyric assigned a single emotion label based on its semantic meaning, narrative context, and dominant affective tone.

To ensure annotation validity, a portion of the labeled entries, covering both the Kaggle-sourced annotations and the manually annotated lyrics, were subjected to expert validation by a licensed psychologist affiliated with Universitas Islam Negeri Sultan Syarif Kasim Riau. The expert independently assessed each label against its lyric content, and labels that did not align with the expert's assessment were revised accordingly. Labels identified for revision were corrected prior to model training. This validation step was conducted to strengthen the psychological grounding of the emotion labels and ensure their consistency with established affective frameworks.

2.3. Text Preprocessing

Prior to model training, all lyric texts underwent a two-stage preprocessing pipeline tailored for IndoBERT compatibility.The first stage is case folding, which converts all characters to lowercase to ensure consistent token representation across the corpus.

The second stage is data cleaning, which encompasses three operations: duplicate removal to eliminate entries with identical songs, reducing the dataset from 1,031 to 1,025 entries; structural marker removal to strip platform-specific artifacts such as section headers (e.g., [Verse 1], [Chorus]), navigation markers (e.g., Back to Chorus, Kembali ke Reff), and contributor metadata introduced by platforms such as Genius and KapanLagi; and removal of non-alphabetic characters including numbers, punctuation, emojis, and symbols. Stopword removal and stemming were deliberately omitted, as IndoBERT's subword tokenizer relies on full sentence context to produce optimal token representations [17].

Following preprocessing, emotion labels were encoded as integer indices: joy (0), sadness (1), fear (2), and anger (3). The dataset was then partitioned into training, validation, and test sets at an 80:10:10 ratio using stratified splitting to preserve class distribution across all subsets, yielding 797 training samples, 100 validation samples, and 100 test samples.

2.4. IndoBERT Fine-Tuning

IndoBERT is a monolingual pre-trained language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically developed for the Indonesian language as part of the IndoNLU benchmark [18, 19]. Being trained exclusively on Indonesian-

language corpora, IndoBERT is able to capture language-specific morphological, syntactic, and semantic patterns more effectively than general-purpose multilingual models.

IndoBERT follows the BERT-Base configuration, consisting of 12 transformer encoder layers, each with a hidden dimension of 768, 12 self-attention heads, and feed-forward layers with an inner dimension of 3,072, totalling approximately 124.5 million parameters. The model employs a SentencePiece tokenizer with byte-pair encoding (BPE) and a vocabulary of 30,522 subword tokens. Pre-training was performed using masked language modeling (MLM) and next sentence prediction (NSP) objectives on TPUv3-8 hardware. The model was trained on the Indo4B dataset, a large-scale Indonesian corpus of approximately four billion words (~23 GB) aggregated from diverse sources including online news, Wikipedia, social media, blogs, subtitles, and parallel corpora. Training proceeded in two phases: the first phase used a maximum sequence length of 128 tokens, and the second phase extended this to 512 tokens to capture longer contextual dependencies.

The bidirectional self-attention mechanism enables IndoBERT to generate contextual representations of each token conditioned on its full surrounding context, making it well-suited for tasks requiring deep semantic understanding such as emotion classification. In this study, the IndoBERT-base-p2 variant is employed, representing the product of the second pre-training phase on a further enriched Indonesian corpus, which offers richer representations of Indonesian text compared to the first-phase variant. For downstream classification, a task-specific head is appended on top of the pre-trained encoder and fine-tuned end-to-end, as described below.

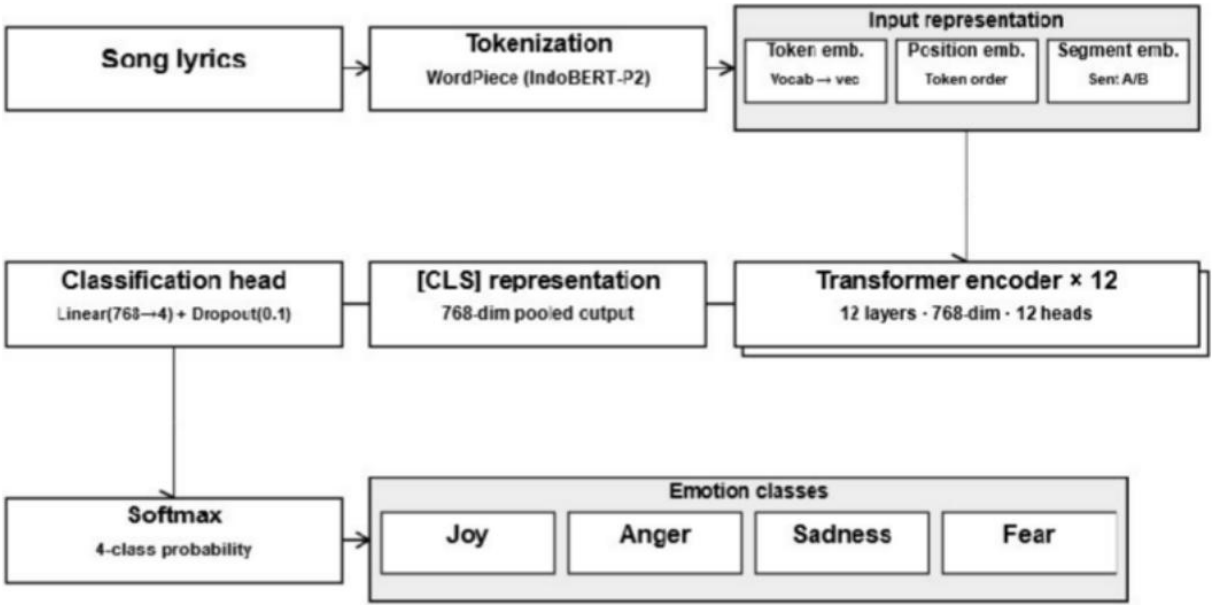

Figure 2. IndoBERT Architecture.

As illustrated in Figure 2, each lyric was tokenized using the pre-trained tokenizer associated with IndoBERT-base-p2, producing input IDs and attention masks. A classification token [CLS] was prepended to serve as the sequence-level representation, and a separator token [SEP] was appended as the boundary marker. The resulting [CLS] embedding $h\ CLS$ was passed through a linear projection layer followed by a Softmax activation function to produce a probability distribution over the four emotion classes [20]. Formally, the classification output is defined as:

$$\hat{y} = softmax(W \cdot h[CLS] + b) \quad (1)$$

where W is the weight matrix of the linear layer, b is the bias vector, and $\hat{y}$ represents the predicted probability distribution over the emotion classes.

Fine-tuning was conducted using Cross-Entropy Loss as the objective function, which measures the discrepancy between the predicted probability distribution and the true class labels [21]. The loss is defined as:

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \quad (2)$$

where N is the number of training samples, C is the number of emotion classes (i.e., four),and $y_{i,c}$ the ground-truth label indicator, and $\hat{y}_{i,c}$ is the predicted probability for sample i belonging to c. Minimizing this loss drives the model to assign higher probabilities to the correct emotion class.

Model parameters were updated using the AdamW optimizer [22]. which extends the Adam algorithm with decoupled weight decay regularization to mitigate overfitting. Unlike standard Adam, AdamW separates the weight decay term from the gradient-based update, directly penalizing large parameter values rather than incorporating them into the adaptive learning rate mechanism. This decoupling has been shown to improve generalization performance, particularly in transformer-based models fine-tuned on downstream tasks.

All experiments used a fixed batch size of 16, a maximum sequence length of 256 tokens, and a maximum of 10 training epochs. To identify the optimal model configuration, a series of experiments was performed by systematically varying two key hyperparameters: learning rate and dropout rate. Early stopping with a patience of 3 was applied based on validation loss, whereby training was halted when no improvement was observed over three consecutive epochs, and the model checkpoint with the lowest validation loss was saved for final evaluation.

2.5. Evaluation

Model performance was assessed on the held-out test set using standard classification metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score, consistent with evaluation frameworks commonly adopted in multi-class classification studies [23]. Given the multi-class nature of the task, both macro-averaged and weighted averaged variants of these metrics were computed to account for class imbalance across the four emotion categories. All evaluation procedures were implemented using the scikit-learn library in Python.

3. RESULTS AND DISCUSSIONS

3.1. Hyperparameter Experiment Results

A total of nine fine-tuning experiments were conducted by systematically varying two hyperparameters: learning rate (1×10^{-5}, 2×10^{-5}, and 3×10^{-5}) and dropout rate (0.1, 0.2, and 0.3), while keeping all other configurations fixed. The results of all experiments are summarized in Table 2.

Table 2. Hyperparameter experiment results.

Exp	Dropout	LR	Best epoch	Accuracy	Macro F1	Weighted F1
1	0.1	1×10^{-5}	4	70.87%	70.29%	70.38%
2	0.1	2×10^{-5}	4	72.82%	72.91%	72.57%
3	0.1	3×10^{-5}	4	75.73%	75.85%	76.02%
4	0.2	1×10^{-5}	4	73.79%	74.61%	74.12%
5	0.2	2×10^{-5}	4	73.79%	73.75%	73.56%
6	0.2	3×10^{-5}	2	67.96%	69.17%	67.78%
7	0.3	1×10^{-5}	4	70.87%	71.08%	70.77%
8	0.3	2×10^{-5}	4	72.82%	72.73%	72.63%
9	0.3	3×10^{-5}	4	71.84%	72.08%	71.58%

Across all experiments, LR 3×10^{-5} demonstrated the most variable behavior: it yielded the highest individual performance in Experiment 3 (Macro F1 = 75.85%) while also producing the lowest in Experiment 6 (Macro F1 = 69.17%), indicating strong sensitivity to dropout configuration at this learning rate. LR 2×10^{-5} produced the most consistent results across all dropout settings, with macro F1 scores ranging narrowly between 72.73% and 73.75%, reflecting stable convergence regardless of regularization strength. LR 1×10^{-5} exhibited the most stable convergence behavior, with early stopping consistently triggered at epoch 4 across all dropout settings, yet yielded the lowest average macro F1 of the three learning rates. Regarding dropout, experiments with dropout 0.1 achieved the highest average macro F1 of 73.02%, primarily driven by the strong performance of Experiment 3,

while dropout 0.3 yielded the lowest average of 71.96%, suggesting that overly strong regularization may constrain the model's capacity to acquire sufficient task-specific representations from the current dataset.

The best configuration was identified as Experiment 3 (dropout = 0.1, LR = 3×10^{-5}), which achieved 75.73% accuracy and 75.85% macro-averaged F1-score. This configuration is selected as the final model for subsequent evaluation.

3.2. Best Model Evaluation

The best model (experiment 3) was evaluated on the held-out test set. The per-class classification report is presented in Table 3.

Table 3. Per-class classification results of the best model (Dropout = 0.1, LR = 3×10^{-5}).

Class	Precision	Recall	F1-Score	Support
Joy	85.71%	80.00%	82.76%	30
Anger	100.00%	61.90%	76.47%	21
Sadness	64.10%	80.65%	71.43%	31
Fear	69.57%	76.19%	72.73%	21
Macro avg	79.85%	74.69%	75.85%	103
Weighted avg	78.83%	75.73%	76.02%	103

The model achieved its highest F1-score on the joy class (82.76%), followed by anger (76.47%), fear (72.73%), and sadness (71.43%). The anger class attained a perfect precision of 100%, indicating that all instances predicted as anger were correct, though its recall of 61.90% suggests that a substantial portion of anger lyrics were misclassified into other categories, particularly sadness. The joy class demonstrated the most balanced performance, achieving both the highest F1-score and reasonably strong precision (85.71%) and recall (80.00%). The sadness class exhibited the lowest precision (64.10%) alongside a comparatively high recall (80.65%), indicating a systematic tendency of the model to over-predict this class, as evidenced by false positives drawn from joy, anger, and fear samples in the confusion matrix.

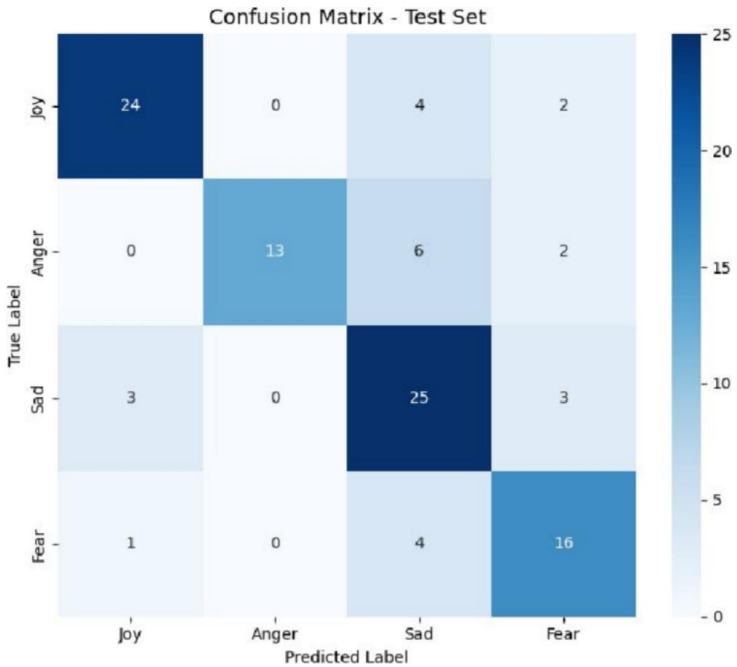

Figure 3. Confusion matrix of the best model (Dropout = 0.3, LR = 2×10^{-5}).

The confusion matrix of the best model is presented in Figure 3. Of the 30 joy samples, 24 were correctly classified, with 4 misclassified as sadness and 2 as fear. The anger class achieved 13 correct predictions out of 21, with 6 misclassified as sadness and 2 as fear, which accounts for its low recall despite perfect precision. The sadness class correctly identified 25 out of 31 samples, with the remaining misclassifications involving 3 predicted as joy and 3 as fear; however, its low precision stems from false positives originating from all other emotion classes being predicted as sadness. The fear class correctly identified 16 out of 21 samples, with 1 misclassified as joy and 4 as sadness. These misclassification patterns confirm that sadness is the most difficult boundary to enforce, attracting false positives from all other emotion classes.

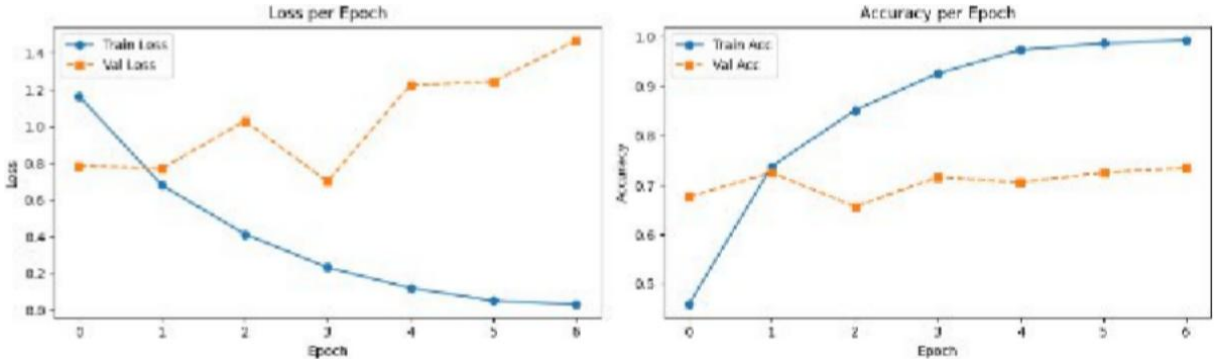

Figure 4. Training and validation loss history of the best model (Dropout = 0.3, LR = 2×10^{-5}).

The training history of the best model is illustrated in Figure 4. Training loss decreased steadily across all seven epochs, while validation loss exhibited a non-monotone trajectory, reaching its minimum at epoch 4 (0.7009), after which it increased consistently through epochs 5, 6, and 7, indicating the onset of overfitting. Early stopping was triggered at epoch 7, and the model checkpoint from epoch 4 was retained for evaluation. The divergence between training and validation loss observed from epoch 5 onward is characteristic of a model beginning to memorize training patterns, further validating the role of early stopping in preserving generalization performance. The training accuracy continued to climb beyond epoch 4, reaching 99.27% at the stopped epoch, while validation accuracy plateaued between 70% – 73%, reinforcing the importance of selecting the best checkpoint based on validation loss rather than training accuracy.

3.3. Discussion

The experimental results demonstrated that the interaction between learning rate and dropout rate jointly determined model performance, with neither hyperparameter exhibiting a consistently dominant effect in isolation. Across all dropout settings, LR 3×10^{-5} produced the most variable outcomes: it yielded the highest individual result in Experiment 3 (Macro F1 = 75.85%) while simultaneously producing the lowest in Experiment 6 (Macro F1 = 69.17%), indicating high sensitivity to the accompanying dropout configuration. This behavior reflects the dual nature of large learning rates during transformer fine-tuning: when paired with low dropout, the model converges aggressively toward strong task-specific representations; when paired with moderate dropout, the combined regularization pressure and large gradient updates destabilize training, impairing generalization.

LR 1×10^{-5} demonstrated the most stable convergence behavior, with early stopping consistently triggered at epoch 4 across all dropout settings, yet yielded the lowest average macro F1 of 71.99%. This suggests that while a smaller learning rate preserves the pre-trained weight structure and avoids overfitting, it may also limit the extent of task-specific adaptation achievable within the available training epochs. LR 2×10^{-5} produced the most consistent results across dropout configurations, with macro F1 scores ranging narrowly between 72.73% and 73.75%, reflecting a stable balance between convergence speed and generalization that is less dependent on the specific dropout setting.

Regarding dropout, experiments with dropout 0.1 achieved the highest average macro F1 of 73.02%, primarily driven by the strong performance of Experiment 3, while dropout 0.3 yielded the lowest average of 71.96%. This contrasts with the common expectation that stronger regularization benefits smaller datasets, suggesting that for this particular task and dataset size, a lower dropout rate preserves sufficient representational capacity for the model to capture the nuanced emotional semantics present in Indonesian song lyrics.

The best configuration (dropout = 0.1, LR = 3×10^{-5}) achieved 75.73% accuracy and 75.85% macro-averaged F1-score. Among the four emotion classes, joy was classified most reliably, attaining an F1-score of 82.76%, followed by anger (76.47%), fear (72.73%), and sadness (71.43%). The anger class achieved perfect precision of 100%, indicating that all instances predicted as anger were correct, though its recall of 61.90% reveals that a considerable portion of anger lyrics were absorbed into neighboring classes, particularly sadness. The sadness class presented the greatest classification challenge, exhibiting the lowest precision of 64.10% alongside a high recall of 80.65%, indicating a systematic tendency to over-predict this class, as reflected by false positives originating from joy, anger, and fear samples in the confusion matrix.

Overall, the results demonstrated that IndoBERT-base-p2, when fine-tuned with appropriate hyperparameter configuration, is capable of effectively capturing emotional semantics in Indonesian song lyrics, a domain characterized by rich figurative language and culturally specific emotional expression.

4. CONCLUSION

This study presented the development and evaluation of an emotion classification model for Indonesian song lyrics using fine-tuned IndoBERT-base-p2. The dataset comprised 1,025 labeled lyric entries collected from three sources, namely Kaggle, Genius, and KapanLagi, covering four emotion categories: joy, sadness, fear, and anger. A systematic preprocessing pipeline was applied, followed by a series of nine fine-tuning experiments systematically varying learning rate and dropout rate under early stopping based on validation loss.

The best-performing configuration (dropout = 0.1, learning rate = 3×10^{-5}) achieved 75.73% accuracy and 75.85% macro-averaged F1-score on the held-out test set, demonstrating that IndoBERT-base-p2 is a viable and effective approach for domain-specific emotion classification in Indonesian song lyrics. The results indicate that the interaction between learning rate and dropout rate jointly shapes fine-tuning outcomes, with neither hyperparameter exhibiting a consistently dominant effect in isolation. Among the four emotion classes, joy and anger were classified most reliably, attaining F1-scores of 82.76% and 76.47% respectively, while sadness remained the most challenging category, exhibiting the lowest precision of 64.10% alongside a recall of 80.65%, reflecting a systematic tendency of the model to over-predict this class.

Future research may explore the application of data augmentation techniques, such as back-translation and synonym substitution [24], to mitigate class imbalance between emotion categories. The extension of this framework to a broader set of emotion categories beyond the four classes examined in this study is also worth pursuing. Additionally, the adoption of large language model (LLM)-based approaches, such as prompt-based or instruction-tuned models [25, 26], may offer improved performance and flexibility for emotion classification in Indonesian lyric texts. The incorporation of explainability mechanisms, such as attention visualization, is also recommended to enhance the interpretability of model predictions [27].

ACKNOWLEDGMENTS

The authors express their gratitude to Universitas Islam Negeri Sultan Syarif Kasim Riau for its institutional support throughout this research. Appreciation is also extended to the Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, for providing the computational resources and research facilities utilized in this study. The authors also acknowledge the contribution of the expert validator from the Department of Psychology, Faculty of Psychology, Universitas Islam Negeri Sultan Syarif Kasim Riau, for the valuable input provided during the emotion label validation process.

REFERENCES

[1] Xu, L., Sun, Z., Wen, X., Huang, Z., Chao, C. J., & Xu, L. (2021). Using machine learning analysis to interpret the relationship between music emotion and lyric features. *PeerJ Computer Science*, **7**, e785.

[2] Setiadi, N. A. & Yuliati, Y. (2025). Dampak Musik pada Perubahan Suasana Hati. *Jurnal Pendidikan Indonesia*, **6**(4), 2151–2157.

[3] Andini, D. T., Khusaini, F., Arsila, S. P., & Mulyeni, S. (2026). Hubungan Antara Mendengarkan Musik dan Perubahan Mood. *WISSEN : Jurnal Ilmu Sosial dan Humaniora*, **4**(1), 42–53.

[4] Levy, A., Granot, R., & Peres, R. (2024). Lyrics do matter: how "coping songs" relate to well-being goals. The COVID pandemic case. *Frontiers in Psychology*, **15**, 1431741.

[5] Han, D., Kong, Y., Han, J., & Wang, G. (2022). A survey of music emotion recognition. *Frontiers of Computer Science*, **16**(6), 166335.

[6] Agrawal, Y., Shanker, R. G. R., & Alluri, V. (2021). Transformer-based approach towards music emotion recognition from lyrics. *European Conference on Information Retrieval*, 167–175.

[7] Shaday, E. N., Engel, V. J. L., & Heryanto, H. (2024). Application of the bidirectional long short-term memory method with comparison of Word2Vec, GloVe, and FastText for emotion classification in song lyrics. *Procedia Computer Science*, **245**, 137–146.

[8] Revathy, V. R., Pillai, A. S., & Daneshfar, F. (2023). LyEmoBERT: Classification of lyrics' emotion and recommendation using a pre-trained model. *Procedia Computer Science*, **218**, 1196–1208.

[9] Noveanto, M., Sastypratiwi, H., & Muhardi, H. (2022). Uji akurasi klasifikasi emosi pada lirik lagu bahasa indonesia. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, **10**(3), 311–318.

[10] Christian, W., Adamlu, D., Yu, A., & Suhartono, D. (2025). Leveraging IndoBERT and DistilBERT for Indonesian emotion classification in e-commerce reviews. *Procedia Computer Science*, **269**, 321–330.

[11] Shaw, C., LaCasse, P., & Champagne, L. (2025). Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT. *Social Network Analysis and Mining*, **15**(1), 22.

[12] Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). Hybrid models for emotion classification and sentiment analysis in Indonesian language. *Applied Computational Intelligence and Soft Computing*, **2024**(1), 2826773.

[13] Rahmi, N. A. & Defit, S. (2024). The Use of Hyperparameter Tuning in Model Classification: A Scientific Work Area Identification. *JOIV: International Journal on Informatics Visualization*, **8**(4), 2181–2188.

[14] Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. *Informatics*, **8**(4), 79.

[15] Yang, L. & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, **415**, 295–316.

[16] Pratama, A. B. (2023). *Indonesian Song Lyric Emotion Dataset*. Kaggle, URL: https://www.kaggle.com/datasets/bytadit/indo-song-emolyric-dataset.

[17] Khairani, U., Mutiawani, V., & Ahmadian, H. (2024). Pengaruh tahapan preprocessing terhadap model Indobert dan Indobertweet untuk mendeteksi emosi pada komentar akun berita Instagram. *Jurnal Teknologi Informasi dan Ilmu Komputer*, **11**(4), 887–894.

[18] Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857.

[19] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, **1**, 4171–4186).

[20] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning, 1(2), 1–800.

[21] Bishop, C. M. & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. *Information Science and Statistics*, **4**(4), 738.
[22] Loshchilov, I. & Hutter, F. (2017). Decoupled weight decay regularization. *ICLR 2019 Conference*, 1–19.
[23] Asyraffi, A., Okfalisa, O., Insani, F., Agustian, S., & Candra, R. M. (2025). Toddler nutritional status identification: Support vector machine (SVM) algorithm adoption. *Science, Technology, and Communication Journal*, **5**(2), 27–36.
[24] Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data*, **8**(1), 101
[25] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, **56**(2), 1–40.
[26] Liu, Z., Yang, K., Xie, Q., Zhang, T., & Ananiadou, S. (2024). Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5487–5496.
[27] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, **15**(2), 1–38.