OPEN ACCESS
ISSN 2356-5462
https://journals.telkomuniversity.ac.id/ijoict

Intl. Journal on ICT
Vol. 12, No. 1, Jun 2026. pp. 66-77
doi.org/10.21108/ijoict.v12i1.11223

Text Classification Using Large Language Models to Detect AI and Human-Generated Sentences

Arfan Rahman Yudiantoro[1*], Prati Hutari Gani[2], Donni Richasdy[3]

[1,2,3] *School of Computing, Telkom University*
Jl. Telekomunikasi No. 1 Terusan Buah Batu, Bandung, Jawa Barat, Indonesia,40257

[*]arfanry @student.telkomuniversity.ac.id

Abstract

As Artificial Intelligence (AI) evolves quickly, especially with Large Language Models (LLMs) like ChatGPT, GPT-4, and Gemini, machines are becoming more adept at producing text that closely mimics human writing. While this capability provides significant benefits across various domains, it also raises concerns regarding content authenticity and information integrity. A major challenge is to automatically and accurately differentiate between AI-generated text and text written by humans.This research introduces a method for detecting AI-generated text in the Indonesian language, utilizing the IndoBERT model. The research process involves gathering datasets of human-written and AI-generated texts, preprocessing the text, extracting contextual features with IndoBERT embeddings, training a binary classification model, and assessing the model's performance through accuracy, precision, recall, and F1-score metrics. This study utilized a balanced dataset of 18,750 text samples, comprising 9,375 human-written and 9,375 AI-generated texts.The experimental findings indicate that the IndoBERT model we proposed reached an accuracy of 99.44%, a precision of 99.14%, a recall of 99.76%, and an F1-score of 99.45%. In addition, IndoBERT surpassed multiple baseline models, such as Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). These findings indicate that Transformer-based architectures are highly effective in capturing contextual and semantic characteristics of Indonesian text, making them suitable for AI-generated text detection and supporting content authenticity verification in digital environments.

Keywords: AI-generated text detection, IndoBERT, Large Language Models, Text Classification, Natural Language Processing.

I. INTRODUCTION

The rapid development of Artificial Intelligence (AI), particularly Large Language Models (LLMs), has transformed the way humans interact with language and information. Models such as ChatGPT, GPT-4, and Gemini are capable of generating text that closely resembles human writing in terms of structure, grammar, and reasoning flow [1]. This capability makes the distinction between human-written text and AI-generated text increasingly difficult to identify [2]. LLMs have been increasingly used in many areas such as article generation, marketing content creation, report writing and automated conversation in customer service and digital platforms [3], [4]. On the one hand, this technology boosts efficiency and productivity. On the other hand, the ease of instant text generation can lead to problems such as misinformation dissemination, fake news generation and questionable content authenticity [2], [5]. Multiple approaches have been suggested to detect AI-generated text. The early approaches were primarily based on statistical techniques such as perplexity analysis, word distribution analysis and simple linguistic features [6], [7]. However, recent studies have shown that statistical-

Received on 03 Jun 2026. Revised on 23 Jun 2026. Accepted and Published on 27 Jun 2026.

based methods are often insufficient for contemporary generative text, especially for advanced LLMs with high semantic complexity and human-like writing styles [1], [8].

Recent advances of Natural Language Processing (NLP) [3], [4] have demonstrated that deep learning and Transformer-based methods are more effective solutions for AI-generated text detection tasks. These models can identify semantic and contextual relationships in text better than traditional machine learning methods. Existing works have also demonstrated the effectiveness of pretrained language models such as BERT and RoBERTa on the task of AI-generated text classification [9]. Recently, IndoBERT has shown promising results on several NLP tasks such as sentiment analysis, text classification and misinformation detection for Bahasa Indonesia [10], [11], [12], [13]. In addition, Pitriani et al. [14] showed that IndoBERT can be used for detecting AI-generated text in an automated essay assessment system. To aid in essay evaluation, their research combined semantic similarity scoring with the detection of AI-generated text. The ability of IndoBERT to grasp the linguistic and semantic features of Indonesian text is underscored by these findings. Unlike the work of Pitriani et al. [14], which concentrated on educational assessment scenarios, this study looks at detecting AI-generated text as a separate binary classification task.Numerous earlier studies have investigated the identification of AI-generated text through different methods, yet the majority of existing research concentrates on English-language datasets and global contexts [1], [6], [15]. Conversely, this study focuses on identifying whether text in Bahasa Indonesia is generated by AI or written by humans, utilizing IndoBERT, a pretrained language model tailored for Indonesian text. Furthermore, the proposed approach is evaluated against several baseline methods, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM), to provide a comprehensive assessment of its effectiveness.

This study makes three key contributions. Initially, it examines how well a fine-tuned IndoBERT model can differentiate between AI-generated and human-written text in Bahasa Indonesia. Secondly, it offers a comparative assessment of IndoBERT against various traditional machine learning and deep learning baseline models. Third, it offers a thorough performance evaluation with accuracy, precision, recall, and F1-score metrics to determine the appropriateness of various classification methods for detecting Indonesian AI-generated text.

II. LITERATURE REVIEW

Wu et al. [1] reviewed statistical, machine learning and deep learning approaches for AI-generated text detection and found that LLM-based approaches generally provide more stable performance, due to their better ability of capturing semantic relationships and contextual patterns of language than traditional statistical approaches. Several works have pointed out the limitations of the traditional AI-generated text detection techniques. Sadasivan et al. [2] showed that many conventional detectors are susceptible to paraphrasing and text manipulation, thus leading to significant performance degradation in evaluating the outputs from modern LLMs. Likewise, Sanchez-Medina [7] suggested a machine learning method to distinguish AI-generated text from human-written text. The method was moderately successful, but the study noted that traditional machine learning techniques are often not enough to understand the complex semantic structures in modern generative text. Krishna et al. [8] also pointed out that paraphrasing techniques can defeat many existing AI detectors, which highlights the need for more sophisticated semantic-aware techniques.

To overcome these limitations, recent works have increasingly used deep learning and Transformer based architectures due to their better contextual understanding capabilities. Deep learning-based text classification methods have been very effective in capturing contextual and semantic information in textual data [3], [4]. Petrillo et al. [5] studied fine-tuned Transformer-based models with Explainable Artificial Intelligence (XAI) techniques to classify AI-generated and human-written sentences. Their findings showed that Transformer-based approaches achieve strong classification performance and improve model interpretability by highlighting important tokens that impact the classification decision. Pre-trained Transformer models such as BERT and RoBERTa have performed well in the classification of AI-generated text. Fine-tuning pretrained language models can improve the classification performance significantly compared to training the models from scratch, as shown by Yadagiri et al. [9].

In addition, Mobin and Islam [15] proposed the LuxVeri framework which uses an Inverse Perplexity Weighted Ensemble strategy to combine multiple LLMs, thus improving the robustness of multilingual AI-

generated text detection. Their results indicated that Transformer-based ensemble methods can enhance detection performance across various languages and writing styles.

For the Indonesian language, IndoBERT has become one of the best-performing Transformer-based language models for NLP tasks. IndoBERTweet, a pretrained language model for Indonesian social media text, was proposed by Koto et al. [10]. IndoNLU, a benchmark and resource collection for evaluating Indonesian Natural Language Understanding tasks, was introduced by Wilie et al. [11]. IndoBERT has been demonstrated to be effective in sentiment analysis [13], misinformation detection [12], and AI-generated text detection in Bahasa Indonesia [14] in prior works. Table I summarizes representative studies on AI-generated text detection using traditional machine learning, deep learning and large language model based approaches.

TABLE I
COMPARISON OF PREVIOUS AI-GENERATED TEXT DETECTION STUDIES

Study	Method	Dataset	Accuracy	Recall	Precision	F1-Score
Mobil et al	Bert	COLING 2025 Workshop on Detecting AI-Generated Content	-	-	-	75%
Yadagiri et al.	TF-IDF	40,000 inquiries alongside corresponding responses from ChatGPT and humans.	99%	-	-	-
Sanchez et al.	Random Forest	ChatGPT generated	-	84%	84%	-
Wilie et al.	Bert	generated by AI - ChatGPT-generated and written by humans.	98%	96%	1.0%	98%

Table I provides a summary of the results of several prior studies on AI-generated text detection. From the comparison, we can see that generally Transformer-based approaches outperform traditional machine learning techniques, due to their ability in capturing contextual and semantic information. In addition, research specifically on Indonesian-language AI-generated text detection is still scarce. Therefore, this study aims to examine the performance of fine-tuned IndoBERT to classify Indonesian AI-generated text and human-written text.

III. RESEARCH METHOD

As shown in Figure 1, this study puts forward a workflow for identifying whether text in Bahasa Indonesia is generated by AI or written by a human. The suggested method relies on Transformer-based text classification techniques that have been commonly employed in earlier research. Specifically, this study employs IndoBERT as a pretrained language model to create contextual text representations and execute binary classification. It all starts with gathering and labeling data; then comes preprocessing, which includes text normalization, eliminating unnecessary characters, and tokenization. To aid in the model's development and assessment, the dataset is split into training, validation, and testing sets. Our proposed approach moves away from the usual dependence on statistical features like TF-IDF or Bag-of-Words and instead employs contextual embeddings created by IndoBERT.

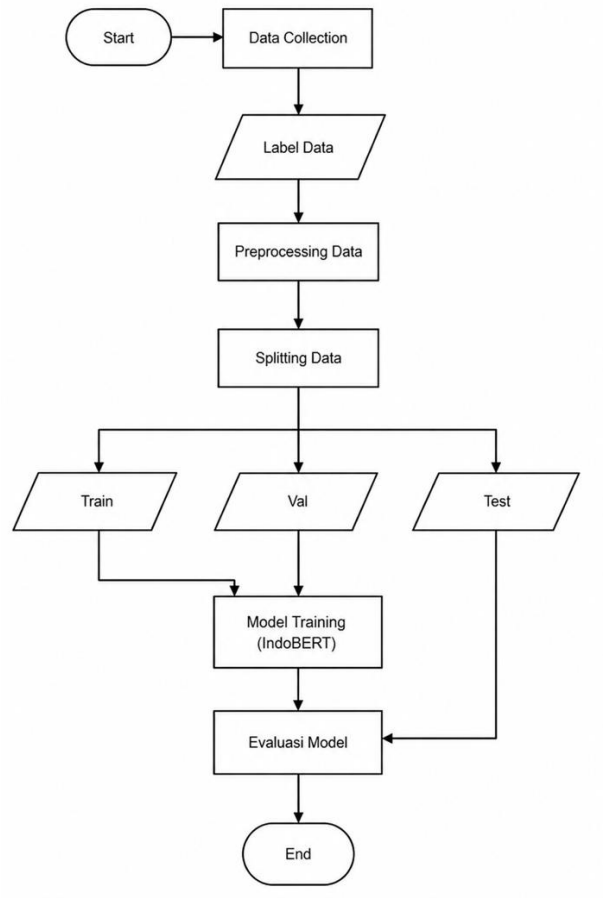

Fig. 1. Flowchart Design

This allows the model to improve its understanding of semantic relationships and contextual subtleties, which are crucial for distinguishing between AI-generated text and human writing. To enhance its ability to detect Indonesian AI-generated text, the method was fine-tuned several times. This involves utilizing a balanced dataset, creating a separate validation set for model selection, and employing optimization strategies like learning-rate warmup, dropout regularization, and early stopping. To improve training stability, reduce overfitting, and boost model generalization, these strategies were put into action.A significant achievement of this approach is the adaptation of IndoBERT for the specific task of identifying AI-generated text in Bahasa Indonesia. The goal of this approach is to outperform the standard machine learning models that serve as baselines in this study by combining contextual language representations with training optimization methods.

A. Data Collection

The study used two categories of Indonesian text: human-written text and AI-generated text. The dataset was taken from the FreedomIntelligence/evol-instruct-indonesian dataset on Hugging Face. This dataset consists of instruction-response pairs in Indonesian. Each text instance is labelled with a source label indicating whether it was generated by a human (human) or by an AI model (gpt). Human-written text samples were extracted from records labelled human. AI-generated text samples were extracted from records labelled gpt. During preprocessing, only the textual content contained in the value field was kept, while metadata and formatting information were removed. For the classification task each extracted text was treated as an independent sample. For this work, we have taken the original labels given by the FreedomIntelligence/evol-instruct-indonesian dataset as the ground truth. human Create a story about a guy who buys a new car, and then everything goes wrong for him in relation to the new car. The dataset annotation scheme treated human-labeled

texts as representatives of human written text, and gpt-labeled text as representatives of AI generated text. However, it should be noted that the human class is mostly composed of instruction-like texts rather than natural human-authored documents, which might influence the learned characteristics of the classifier. To enhance data quality, duplicate records, null values and empty texts were checked and removed where necessary. The dataset was shuffled randomly, and then split into training, validation and test sets. This was done to mitigate any ordering bias and to have a balanced amount of human and AI generated texts in all subsets. In this paper, we utilize a dataset comprising 20,000 text samples, with 10,000 authored by humans and 10,000 produced by AI.The data distribution used in this study, after balancing and pre-processing, is shown in Table II. To ensure a fair evaluation and reduce the risk of data leakage, the dataset was randomly shuffled before it was divided into training, validation, and testing subsets. Consistent with the experimental design of this study, 70% of the data (14,000 samples) was allocated for training, 15% (3,000 samples) for validation, and the remaining 15% (3,000 samples) for testing. The IndoBERT model was fine-tuned with the training set; the validation set was used for hyperparameter tuning and monitoring model performance during training, while the testing set was reserved for the final evaluation of the learning process.

TABLE II
DATA DISTRIBUTION

Dataset Type	Source	Number of Data
Human	News articles, blogs, educational content, and Indonesian social media,and FreedomIntelligence/evol-instruct-indonesian dataset	10.000
AI	ChatGPT, GPT-4, Gemini, and FreedomIntelligence/evol-instruct-indonesian dataset	10.000
Total		20.000

B. Preprocessing Data

The dataset is preprocessed in a number of stages prior to the training process to improve data quality and compatibility with the IndoBERT model. The preprocessing consists of the following steps: text extraction, text cleaning, case normalization, tokenization, padding and truncation.The text extraction step is performed to retain only the main text content and remove the metadata and formatting information from the raw dataset. Next, text cleaning is performed to remove URLs, special characters, punctuation symbols and unnecessary white spaces that can affect the classification process. After that, all the text is converted to lowercase format to maintain consistency across the dataset. Table III shows examples of sentences both before and after they have been processed.

TABLE III
EXAMPLE OF TEXT BEFORE AND AFTER PREPROCESSING

Raw Text	After Preprocessing
{'from':'gpt','value':'Teks ini termasuk ke dalam kategori argumentatif. Hal ini dikarenakan terdapat sebuah pernyataan...'}	Teks ini termasuk ke dalam kategori argumentatif
{'from':'human','value':'Apa saja fitur yang harus dimiliki sebuah aplikasi untuk membuatnya mudah	Apa saja fitur yang harus dimiliki sebuah aplikasi untuk membuatnya mudah

Then the cleaned text is tokenized with IndoBERT tokenizer to convert the sentences into token representation that is compatible to the Transformer architecture. Padding and truncation methods are also used to standardize sequence length for model input processing. These preprocessing steps are necessary to enhance the quality of text representation and to optimize the performance of the proposed classification model.

C. Splitting Data

After the pre-processing stage, the dataset was further divided into training, validation and test sets to evaluate the performance of proposed model. The split was made randomly, with an equal number of human and AI text samples. The dataset was split into 70% training data, 15% validation data, and 15% test data. The IndoBERT classification model was trained using the training dataset and the validation dataset was used to monitor the performance of the model and to optimize the hyperparameters setting during the training process. During this time the test dataset was used to evaluate the ultimate performance of the model in distinguishing AI generated text from human written text. This data splitting strategy is frequently used in Natural Language Processing (NLP) tasks in order to ensure that the proposed model can generalize well to unseen data and to reduce the risk of overfitting during training. The dataset splitting can be seen in Table IV.

TABLE IV
SPLITTING DATASET DISTRIBUTION

Split Dataset	Human	AI	Total
Train	6,562	6,563	13,125
Test	1,407	1,406	2,813
Validation	1,406	1,406	2,812
Total	9,375	9,375	18,750

D. Model IndoBERT

The core of this research is the model training process, where IndoBERT serves as the base model and is fine-tuned for text classification tasks [10], [11]. The architecture is IndoBERT with a classification layer on top to distinguish between two classes i.e. AI generated text and human written text. The model is then trained on the prepared training data set, and its performance is monitored on the validation data set.

The optimal classification performance is achieved by tuning several training parameters such as learning rate, batch size, MAX_LENGTH, DROPOUT, PATIENCE, and number of epochs based on the validation results. We use the AdamW optimizer for training as it does well for Transformer based architectures and can increase the stability of optimization [3], [4]. Furthermore, an early stopping technique is applied to avoid overfitting during the fine-tuning process. The IndoBERT model is expected to learn linguistic features that differentiate between AI-generated text and human-written text, such as sentence complexity, grammatical structure, semantic pattern, and word usage in Bahasa Indonesia [1], [5], through the training process. The training parameter shown in Table V.

TABLE V
TRAINING PARAMETER CONFIGURATION

Parameter	Tested Values
Learning Rate	3e-5
Bath Size	32
Max_length	128
Drop Out	0,4
Patience	2
Epoch	2

E. Model Evaluation

The evaluation process in this study utilized several commonly used classification metrics, namely accuracy, precision, recall, and F1-score [11], [12]. The overall correctness of predictions produced by the model was measured using accuracy. Precision was the accuracy of the model classifying AI-generated text, recall evaluated the model's ability to identify all relevant AI-generated samples. Furthermore, the F1-score was employed to provide a balanced evaluation between precision and recall. These metrics were selected because they are widely used in text classification and AI-generated text detection research [1], [5].

IV. RESULTS AND DISCUSSION

A. Hyper tuning Parameter

During the training process, the IndoBERT model was trained using 822 training steps and 82 warmup steps. The warmup strategy was applied to stabilize the learning rate during the early stage of training, allowing the optimization process to proceed more smoothly. The training and validation losses obtained during the training process are presented in Table VI.

TABLE VI
TRAINING AND VALIDATION LOSS PER EPOCH

Epoch	Training Loss	Validation Loss
1	0.1567	0.0675
2	0.0348	0.0432

As seen in Table VI, both training loss and validation loss decreased from the first epoch to the second epoch. The training loss decreased from 0.1567 to 0.0348 and the validation loss decreased from 0.0675 to 0.0432. This downward trend means that the model was able to learn patterns in the dataset during the training process. The reduction of both loss values indicates that the chosen hyperparameter setting (3e-5 learning rate, 32 batch size, 0.4 dropout rate, and the use of an early stopping mechanism) contributed to the stability of the training process. Furthermore, no increase in the validation loss was observed during the second epoch, which is considered an initial indicator of no severe overfitting during training. However, these results should be interpreted with caution as the evaluation was performed on data from the same overall distribution as the training data. Therefore, a decrease in the loss values does not provide enough evidence that the model has a good generalization ability when used on data from different sources or distributions. The fast convergence observed in the two epochs is due to the pretrained status of IndoBERT, which has already learned rich representations of Indonesian from large-scale corpora. Therefore, the fine-tuning was mainly aimed at the model's capacity to distinguish AI-generated texts from human-written ones. After the training process was finished, the model was tested on a held-out test set of Indonesian texts written by humans and generated by AI. We evaluated model performance with four common classification metrics: accuracy, precision, recall and F1-score.

B. Model Evaluation

The evaluation results of the proposed IndoBERT model are presented in Table VII. The proposed IndoBERT model achieved the accuracy, precision, recall, and F1-score of 99.44%, 99.14%, 99.76%, and 99.45%, respectively, as shown in Table VII. These results indicate that the model correctly classified most of the samples in the testing dataset. The high precision score indicates that only a few of the texts written by humans were misclassified as AI generated texts. The high recall score suggests that most of the AI-generated texts were detected by the model. This combination of metrics yielded a high F1-score, indicating a good balance between classification accuracy and detection ability.The performance obtained in this study indicates that IndoBERT has a great potential to be employed for AI-generated text detection in Indonesian.

TABLE VII
MODEL EVALUATION RESULT

Accuracy	Precision	Recall	F1-Score
0.9944	0.9914	0.9976	0.9945

This ability can be attributed to the Transformer-based architecture of IndoBERT which can effectively learn contextual and semantic relations between words than traditional feature-based machine learning approaches. The very high performance obtained in this study must, however, be interpreted with caution. The training and testing data sets were generated from the same data collection process, which could cause similarities of linguistic characteristics among the splits of data sets. The similarities may contribute to the high performance obtained during evaluation. Moreover, the model has not yet been tested in out-of-distribution conditions, e.g. paraphrased texts, texts from different domains or texts generated by language models not included in the data collection. Therefore, more evaluation is needed to see if the model can maintain its performance in more challenging real-world conditions.

To reduce the risk of direct data leakage, duplicate and overlap checks were performed on the training, validation, and testing datasets. Considering these limitations, the results of this study can be considered as preliminary evidence that IndoBERT can effectively distinguish AI-generated texts from human-written texts in the dataset used in this research. More diverse datasets and more challenging evaluation settings are needed in future work to better evaluate the robustness and generalizability of the proposed approach.

C. Robustness Evaluation Using Paraphrased Texts

To further evaluate the robustness of the proposed IndoBERT model, an additional experiment was performed over the paraphrased texts. Previous research has demonstrated that AI text detector performance can degrade when text is paraphrased while preserving original meaning. Thus, we conducted a preliminary robustness assessment by applying paraphrasing to the human-written and AI-generated text samples and assessed the model's predictions. The objective of this experiment was to verify that the model was able to preserve its classification ability when the texts' lexical and syntactic structures are changed without significant changes to their semantic content. The results are given in Table VIII.

TABLE VIII
ROBUSTNESS EVALUATION ON PARAPHRASED TEXTS

Sample Type	Text	True Label	Predicted	Result
Human Orginal	Jelaskann sumber energi utama bagi tumbuhann	Human	Human	True
Human Paraphrased	fitur-fitur esensial apa saja yang wajib ada pada aplikasi ponsel yang sukses	Human	AI	False
AI Original	tabel excel dengan 5 tips untuk meningkatkan layanan	AI	AI	True
AI Paraphrased	esai ini dapat digolongkan sebagai teks argumentatif	AI	AI	True

D. Classifier Model Comparison

This experiment looks at how well IndoBERT performs compared to several classification models like Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and LSTM are all methods used in data analysis and machine learning. All models were assessed by looking at accuracy, precision, recall, and F1-score metrics [11], [12]. Conventional machine learning. Models that use TF-IDF features have a hard time grasping the context and meaning behind the relationships in text.[8]. At the same time, deep learning methods such as LSTM and IndoBERT can understand more complicated language patterns from Indonesian text data [2],[3].

The results in Table IX show that IndoBERT had the best performance with accuracy 0.9944 and F1-score 0.9945. LSTM was also able to show promising results with an accuracy of 0.9900. Our method performed better than the traditional models with Naive Bayes having the lowest accuracy of 0.9093 followed by Logistic regression and SVM with 0.9502 and 0.9517 accuracy respectively. IndoBERT's improved performance can

be attributed to its transformer-based architecture that uses a bidirectional attention mechanism to capture the contextual information from both left and right contexts simultaneously [2], [10]. Different from the models based on TF-IDF, which mostly rely on the frequency of words, IndoBERT learns the semantic relationship between words and sentences, so that it can better distinguish the subtle linguistic features of AI-generated text and human-written text. IndoBERT is able to model long-range dependencies more effectively than LSTM, and take advantage of the knowledge gained from large-scale Indonesian language pre-training, which leads to richer text representations and better classification performance [1], [3].These results indicate that Transformer-based models like IndoBERT are more suitable for detecting AI-generated text in Indonesian because of their ability to capture contextual information and semantic structure [1], [4], [10].

TABLE IX
COMPARISON OF MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.9093	0.8966	0.9253	0.9107
Logistic Regression	0.9502	0.9396	0.9623	0.9508
SVM	0.9517	0.9441	0.9602	0.9520
LSTM	0.9900	0.9893	0.9908	0.9900
IndoBERT	0.9944	0.9914	0.9976	0.9945

E. Error Analysis

To further investigate the classification behavior of the proposed IndoBERT model, an error analysis was conducted using the confusion matrix and representative misclassified samples. The confusion matrix provides a detailed overview of the model predictions and highlights the types of classification errors produced during testing.

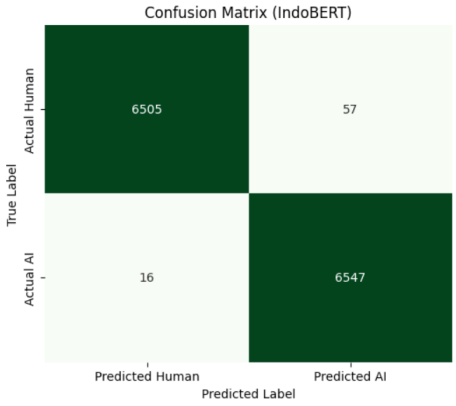

Fig. 2. Confusion Matrix IndoBERT

As shown in Fig. 2, the model correctly classified 6,505 human-written texts and 6,547 AI-generated texts. Meanwhile, 57 human-written texts were incorrectly classified as AI-generated texts (false positives), whereas 16 AI-generated texts were incorrectly classified as human-written texts (false negatives).The confusion matrix indicates that the number of correctly classified samples was substantially higher than the number of misclassified samples, demonstrating the effectiveness of the proposed model on the experimental dataset. However, the existence of false positives and false negatives suggests that certain linguistic and structural characteristics were shared between human-written and AI-generated texts, making the classification task more challenging. The higher number of false positives compared to false negatives indicates that some human-written texts exhibited characteristics commonly associated with AI-generated content. These texts often contained highly structured instructions, formal language, technical descriptions, or programming-related tasks. Consequently, the model occasionally classifies such samples as AI-generated texts. Conversely, the relatively

small number of false negatives suggests that most AI-generated texts were successfully detected. The AI-generated texts that were misclassified as human-written generally contained coherent explanations, natural sentence structures, and document-like organization that closely resembled human-authored writing. To better understand these classification errors, several representative misclassified samples are presented in Table X.

TABLE X
REPRESENTATIVE MISCLASSIFIED TEXT SAMPLES

Text	True Label	Predicted
bagaimana cara menggunakan c untuk menghasilkan resep pembuka yang unik dan kreatif yang menggabungkan apel, bawang bombay, bawang putih, dan minyak zaitun dengan cara yang baru dan menarik?nnuntuk menciptakan resep pembuka yang unik dan kreatif,	AI	Human
berikut contoh pesan dalam format json yang dapat kita kirim untuk meminta penyelesaian tugas dengan sopan dan	Human	AI
siapa orang termuda dalam tabel?n2. siapa lakilaki tertua dalam tabel?n3.	Human	AI
nama aktivitas jenis tujuan materi langkahlangkah keamanan durasi perkiraan kodesumber daya.	AI	Human

The examples presented in Table X provide additional insight into the classification errors made by the proposed model. Texts authored by humans that were incorrectly classified as AI-generated usually featured highly structured instructions, technical jargon, and a focus on tasks. Texts produced by AI that were wrongly categorized as human-written showed traits that are usually linked to content created by humans. The recipe text uses creative language and has coherent explanations, whereas the activity-planning text resembles a document with clearly structured information. Thanks to these characteristics, the replies generated seem natural and fit the context, which raises the challenge of telling them apart from authentic human-written texts. It seems from the observations in Table X that the model might depend on stylistic and structural cues found in the training data, in addition to semantic information. As a result, it becomes harder to classify prompts as human-written or AI-generated when human-written prompts closely resemble AI-generated instructions or when AI responses are crafted in a natural, human-like style.

F. Threats to Validity

The main contribution of this study is the evaluation of a fine-tuned IndoBERT model for distinguishing AI-generated and human-written Indonesian text and its comparison with several traditional and deep learning baseline models. Despite the promising results, several threats to validity should be considered.From a construct validity perspective, the dataset was derived from the FreedomIntelligence/evol-instruct-indonesian dataset, where human-labeled texts mainly consist of user instructions or prompts, while AI-labeled texts consist of language model responses. Therefore, the model may partially learn differences between prompts and responses rather than solely distinguishing human-written and AI-generated text. From an internal validity perspective, the training, validation, and testing sets were obtained from the same overall data collection process. Although the data were randomly shuffled and separated before training, similarities in linguistic patterns across dataset splits may have contributed to the high classification performance. From an external validity perspective, the dataset was collected from a single source and may not fully represent the diversity of Indonesian texts found in real-world applications. Furthermore, the model was not extensively evaluated on out-of-distribution scenarios such as paraphrased texts, cross-domain datasets, or outputs generated by unseen language models.

This limitation is particularly important because previous studies have shown that AI-generated text detectors may experience performance degradation when confronted with paraphrased or domain-shifted inputs [8], [15]. In addition, the proposed approach involves a trade-off between performance and computational cost. While IndoBERT provides good contextual understanding and achieves high classification performance, it requires greater computational resources and offers less interpretability than traditional machine learning methods Therefore, the validity of the findings is limited to the dataset and experimental settings used in this study. Future research should investigate cross-domain generalization, robustness against paraphrasing techniques, and performance on outputs generated by newer Large Language Models.

F. Limitations

There are several limitations to this study. The dataset is initially created from the FreedomIntelligence/evol-instruct-indonesian dataset, where the human-labeled texts are predominantly user instructions or prompts, and the AI-labeled texts are language model responses. This may enable the model to learn some structural differences between prompts and responses, rather than just learning to distinguish human written from AI generated text. Secondly, the dataset was collected from one source, which might not fully represent the diversity of Indonesian texts that appear in real world applications. Third, although we conducted a preliminary robustness assessment, the model was not rigorously tested on external datasets, cross-domain texts, or outputs generated by unseen language models. So, the results should be interpreted in the context of the dataset and evaluation settings used in this study.

V. CONCLUSION

The experimental results revealed that the selected hyperparameter configuration improved the stability and performance of the IndoBERT model during training. The decreasing training and validation losses indicate that the model was able to learn linguistic patterns from the dataset without significant overfitting. In comparison with Naïve Bayes, Logistic Regression, SVM, and LSTM, the fine-tuned IndoBERT model achieved superior performance, demonstrating the effectiveness of Transformer-based architectures in capturing contextual and semantic information within Indonesian text.

The confusion matrix and error analysis further showed that the proposed model correctly classified the majority of samples, with only a limited number of misclassifications. Nevertheless, some human-written texts were incorrectly classified as AI-generated due to their highly structured and instruction-oriented writing style, while some AI-generated texts were classified as human-written because of their natural and coherent language patterns. These findings indicate that distinguishing between AI-generated and human-written text remains challenging when both classes exhibit similar linguistic characteristics. Despite the promising results, several limitations should be acknowledged. The dataset was derived from a single source consisting of instruction–response pairs, where human-labeled texts mainly represent prompts and AI-labeled texts represent generated responses. As a result, part of the classification performance may be influenced by structural differences between prompts and responses. In addition, the evaluation was conducted on data with a similar distribution and did not extensively assess performance on external datasets or more challenging out-of-distribution scenarios.

Future work should focus on evaluating the model using more diverse Indonesian text sources, cross-domain datasets, and outputs generated by different large language models. Further research may also investigate robustness against adversarial paraphrasing techniques, incorporate explainable AI methods to improve model interpretability, and explore ensemble-based approaches or newer Transformer architectures to further enhance the performance and robustness of AI-generated text detection in Bahasa Indonesia.

DATA AND COMPUTER PROGRAM AVAILABILITY

The dataset used in this research is publicly available through the Hugging Face platform. The AI-generated text dataset can be accessed from the following repository:
https://huggingface.co/datasets/FreedomIntelligence/evol-instruct-indonesian/tree/refs%2Fconvert%2Fparquet/default/train

ACKNOWLEDGMENT

The writer wants to express gratitude to the lecturers who have helped them.

REFERENCES

[1] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong, "A survey on LLM-generated text detection: Necessity, methods, and future directions," Computational Linguistics, vol. 51, no. 1, pp. 275–338, 2025.

[2] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-generated text be reliably detected?," arXiv preprint arXiv:2303.11156, 2023.

[3] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1–40, 2021.

[4] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 2, pp. 604–624, 2020.

[5] L. rrillo, F. Martinelli, A. Santone, and F. Mercaldo, "Toward the adoption of explainable pre-trained large language models for classifying human-written and AI-generated sentences," Electronics, vol. 13, no. 20, p. 4057, 2024.

[6] N. Prova, "Detecting AI generated text based on NLP and machine learning approaches," arXiv preprint arXiv:2404.10032, 2024.

[7] J. J. Sanchez-Medina, "Sentiment analysis and random forest to classify LLM versus human source applied to scientific texts," arXiv preprint arXiv:2404.08673, 2024.

[8] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," Advances in Neural Information Processing Systems, vol. 36, pp. 27469–27500, 2023.

[9] A. Yadagiri, L. Shree, S. Parween, A. Raj, S. Maurya, and P. Pakray, "Detecting AI-generated text with pre-trained models using linguistic features," in Proceedings of the 21st International Conference on Natural Language Processing (ICON), 2024, pp. 188–196.

[10] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," arXiv preprint arXiv:2109.04607, 2021.

[11] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," arXiv preprint arXiv:2009.05387, 2020.

[12] A. Wenang, K. A. Widianto, M. E. Nugraha, M. Rizki, and K. Dewi, "Implementasi Model Deep Learning IndoBERT dengan Antarmuka Aplikasi Mobile untuk Deteksi Berita Hoaks Berbahasa Indonesia," KONSTELASI: Konvergensi Teknologi dan Sistem Informasi, vol. 5, no. 2, 2025.

[13] D. K. Sumartha, "Implementasi IndoBERT untuk analisis sentimen opini publik terhadap kebijakan kenaikan UKT di era pemerintahan," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 3S1, 2025.

[14] P. Pitriani, D. S. A. Maylawati, and Y. A. Gerhana, "Deteksi generatif teks pada penilaian otomatis tes esai berbahasa Indonesia menggunakan IndoBERT," JEPIN (Jurnal Edukasi dan Penelitian Informatika), vol. 11, no. 2, pp. 305–313.

[15] M. K. Mobin and M. S. Islam, "LuxVeri at GenAI Detection Task 1: Inverse perplexity weighted ensemble for robust detection of AI-generated text across English and multilingual contexts," arXiv preprint arXiv:2501.11914, 2025.