Scientific Journal of Informatics
Vol. 12, No. 3, August 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Comparative Analysis of High School Student and AI-Generated Essays Using IndoBERT and Linguistic Features

Muhammad Harits Shofwan Adani[1], Alqis Rausanfita[2*], Tanzilal Mustaqim[3]
[1,2,3]Department of Informatics, Telkom University Surabaya, Indonesia

Abstract.
Purpose: The purpose of this study is to address the growing challenge of distinguishing between essays written by humans and essays generated by AI, particularly in the context of high school education in Indonesia. This study aims to analyze the semantic and linguistic differences between student-written and ChatGPT-generated in Indonesian language.
Methods: The study employs an IndoBERT-based semantic model trained with triplet loss to generate paragraph-level embeddings, allowing the measurement of semantic similarity within and between essay classes. Additionally, linguistic features such as lexical diversity, word count, modal usage, and stopword ratio were extracted to capture stylistic and structural differences. These three key features are combined and used as input to a neural network classifier.
Result: The IndoBERT-based semantic model successfully grouped student-written and ChatGPT-generated essays into distinct clusters. The similarity scores within student essays ranged from 0.7 to 0.9, while the similarity between classes was mostly negative with a few outliers, reflecting the cosine similarity metric used in this study, which has a range of -1 to 1. The classification model showed a 90.55% accuracy and an AUC of 0.9999 when evaluated on the independent test set defined in the Data Preparation stage. These results suggest that student-written and ChatGPT-generated essays form distinct semantic clusters. Students' essays show more linguistic diversity, while ChatGPT essays show consistency in the coherence and formality aspects of the essays.
Novelty: This study provides empirical insights of semantic similarities and linguistic features to differentiate between human and AI-generated essays in the Indonesian language. It contributes to supporting academic integrity efforts and highlighting the need for further research across different writing models and contexts.

Keywords: ChatGPT, Essay, IndoBERT, Linguistic features, Semantic similarity, Text classification
Received June 2025 / **Revised** September 2025 / **Accepted** September 2025

INTRODUCTION

Natural Language Processing (NLP) combines computational linguistics and machine learning to enable machines to understand, interpret, and generate human language [1]. With major advances such as transformer-based models like ChatGPT developed by OpenAI, it can generate coherent and contextually relevant text across multiple domains [2], and has been trained with an extensive multilingual corpus, so it can generate essays, feedback, and scientific texts that closely resemble human writing [3]–[5]. This technology is starting to be known in the context of education, making it more accessible to students and educators [6]. This study focuses on ChatGPT because it is widely adopted, publicly accessible and free for generate text such as essays, and remains the most prominent large language model compared to emerging competitors such as Google Bard (Gemini) and Alibaba's Tongyi Qianwen [7], [8].

With the increasing use of AI tools such as ChatGPT are increasingly used in educational settings, pedagogical issues also arise, especially in high school education. In senior high schools (SMA) in Indonesia, essay writing not only as a medium to assess students' language skills but also a tool for fostering critical and analytical thinking skills [9], [10]. The use of AI-generated texts in student assignments raises several challenges such as dependence on AI, reduced motivation for independent thinking, and increased risk of plagiarism [11]–[13]. These concerns highlight the need for deeper analysis to understand how student-written essays differ from AI-generated essays in terms of linguistic and structural quality. To address this, methods are needed that can capture semantic and lexical features within the text.

*Corresponding author.
Email addresses: mhrtss@student.telkomuniversity.ac.id (Adani), alqisfita@telkomuniversity.ac.id (Rausanfita)*, tanzilal@telkomuniversity.ac.id (Mustaqim)
DOI: 10.15294/sji.v12i3.27732

Semantic model offers an approach that allows us to detect subtle differences in coherence, variability, and expressiveness between human and AI writing. Various AI models have been studied to better understand and analyze text, with Bidirectional Encoder Representations from Transformers (BERT) [14], a transformer-based architecture that aims to capture the meaning of words from two directions. Its Indonesian Variant, IndoBERT [15], [16], has been trained on a large Indonesian corpus and performed well on various NLP tasks such as sentiment analysis [17], [18] and text classification [19], [20]. This makes IndoBERT suitable for analyzing Indonesian essay texts.

Various approaches have been developed to detect AI-generated text in academic contexts [21], [22]. Online detection tools and various models show mixed accuracy rates in identifying ChatGPT-generated academic text [21]–[25], highlighting the need for more reliable methods. Studes have shown that prompt-based approaches can effectively distinguish human-written and AI-generated texts [26]. Other studies also highlight the potential of fine-tuned transformer models, such as BERT, to perform well in educational contexts [27]. In addition, a hybrid approach that combines BERT with Siamese Bi-LSTM have shown strong results in detecting English academic datasets [28]. Although these methods demonstrate high performance in English academic texts, its application to Indonesian essays, especially at the high school level, remains underexplored.

The use of linguistic features, such as syntactic complexity and lexical variation, is also a focus in detecting differences between student and ChatGPT-generated texts [29], [30]. A large-scale study conducted by Herbold et al. systematically compared human-written and ChatGPT-generated argumentative student essays, evaluated by a large number of expert teachers. Their findings revealed that although AI-generated essays were often rated higher in terms of quality, they exhibited distinct linguistic patterns compared to human-written texts [31].

Many studies have shown that contrastive learning used for fine-tuning BERT model [32]–[34]. Furthermore, the application of triplet loss effectively learns discriminative embeddings by minimizing distances within similar samples and maximizing those between dissimilar ones, achieving strong results in text classification [35], [36]. Thus, combining contrastive and metric learning approaches can improve the separation between human and AI-generated texts. For classification, neural networks have been successfully applied in Indonesian text classification [28]–[33]. Model performance will be evaluated using standard metrics accuracy, precision, recall, F1-score, and Area under the Curve (AUC) and Receiver Operating Characteristic (ROC) to assess detection capability across confidence levels [28]–[37]. This comprehensive evaluation will determine the effectiveness of the integrated approach for identifying AI-generated essays in the Indonesian high school context.

While numerous studies have explored the differences between AI-generated text and human-written text using transformer models and linguistic features, research focusing on Indonesian high school essays is limited. As AI writing tools such as ChatGPT become more accessible to Indonesian students, there is a need to analyze this issue to maintain academic integrity and encourage independent critical thinking. Without timely attention, students risk a decline in writing and critical thinking skills, while schools and policymakers may face blind spots in regulating the responsible use of AI in education.

This study addresses this gap by combining IndoBERT-based semantic embeddings with linguistic features tailored to the characteristics of Indonesian students' writing. Our method has three main features. It starts with the use of IndoBERT and triplet loss for semantic clustering analysis. Next, it combines linguistic features, including the use of modal verbs and stopword ratios, which are adapted to the structure of the Indonesian language. Lastly, is the classification of texts in high school essays, to show the impact of AI writing skills related to student learning and the important role of regulation of AI use in the educational context. This research expected to contribute to methodological advances and practical insights that support academic integrity in the Indonesia's education system. Future research is expected to explore a wider range of linguistic features, evaluate various current AI text generation tools, and combine diverse datasets to assess the impact of AI across different educational contexts.

METHODS

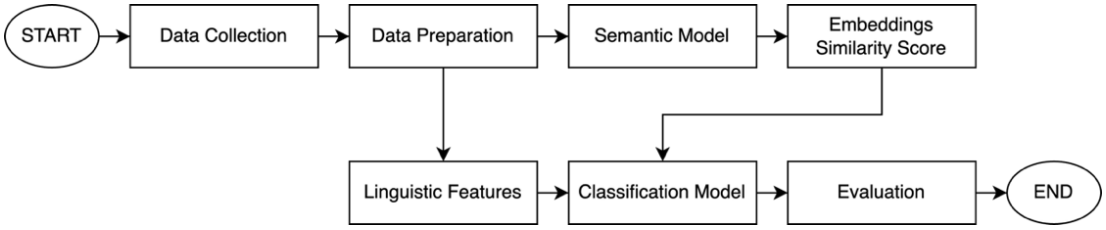

Figure 1. Research method flowchart

This section outlines the research methodology used in this study with an overview of the workflow illustrated in Figure 1, starting with data collection from high school students and ChatGPT-generated essays. The process includes contrastive pairs generation and training using triplet loss on IndoBERT. The next stages involve feature extraction, includes embeddings, similarity scores, and linguistic features, ending with a binary classification task. Performance is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. All analytical processes in this study were conducted using Python version 3.12.7. To support data transparency, all datasets and Python code used in this study have been made publicly available on the OSF platform (https://osf.io/q5zht/).

Data collection

The dataset in this study uses student essay data obtained from a national competition called *Olimpiade Seni dan Bahasa Indonesia* (OSEBI), which is part of the Science and Culture Festival (Festival Sains dan Budaya) organized by Eduversal Foundation. The event serves as a platform for Indonesian youth to express their creativity through visual arts, performing arts, and literary arts. The dataset consists of 20 essays written by high school students (SMA/SMK/MA) who were selected as finalists in the essay writing category of the 2024-2025 competition

Furthermore, the AI-generated essay sets were produced using Python scripts with OpenAI API, the ChatGPT-4o model was chosen as one of the models that can be used on the ChatGPT free plan [38]. To ensure comparability, the prompts were designed to replicate the competition guidelines, instructing ChatGPT to write as if it were a high school student in a national essay contest. Each essay began with a title and continued with 20-30 paragraphs of 3-6 sentences, written in Indonesian following PUEBI and in a natural, communicative style. The prompts also reflected the competition requirements, including the designated theme, target essay length of 3-5 pages, and evaluation criteria such as originality, creativity, relevance, coherence, and engagement.

All procedures adhered to ethical research practices. Student essays were obtained only from publicly available finalist submissions, ensuring no private or sensitive data were used, while the AI-generated texts were created solely for research purposes. This study respects the originality of student work, maintains confidentiality, and complies with academic integrity principles in the responsible use of generative AI.

Data preparation

After the data collection process, the next step is to prepare the essays for input to the IndoBERT semantic similarity model and classification model. This stage aims to standardize the text written by students and that generated by ChatGPT to ensure consistent format and compatibility. The subset of essays used in this study is paragraphs, with each paragraph treated as a single data entry. In case of class imbalance, random undersampling [39]–[41] is applied to the training subset to maintain a balanced dataset and data not selected from the undersampling process is included in the test set. The dataset was then divided into training, validation, and test subsets with a ratio of 75:5:20. This ratio was chosen to maximize the amount of data used for training so the model could learn effectively, while still maintaining sufficient data for validation and allocating an adequate portion for testing. Applying this split consistently across both models ensured comparability of results and fairness in evaluation.

Following the dataset split, the text was prepared for linguistic feature extraction and tokenization using the IndoBERT base model with the BertTokenizer. Special tokens [CLS] and [SEP] are added to mark input boundaries, and the token [PAD] is used to equalize the input length. Then, all tokens are converted into

ID Tokens and paired with an Attention Mask to distinguish actual tokens from padding during processing [42], [43]. This study does not apply traditional text preprocessing, as IndoBERT has been pre-trained on a variety of language styles including formal and slang, and previous studies have shown that BERT models maintain high accuracy without preprocessing, with minimal performance differences across domains [44]–[46].

Semantic model

The proposed semantic model utilizes the pre-trained IndoBERT model to measure the semantic similarity between student essays and ChatGPT generated essays with a summary of the model configuration and training settings presented in Table 1. To preserve the pre-trained language representations, the BERT layers are frozen, and the [CLS] tokens are extracted and passed through multiple dense and dropout layers with L2 normalization to produce standardized embeddings.

IndoBERT was selected as the base model because it has consistently demonstrated strong performance across Indonesian NLP benchmarks. In particular, the variant introduced by Wilie et al. [15] trained on a broad and diverse corpus, was considered more generalizable for essay analysis compared to the IndoBERT version by Koto et al. [16] which was primarily trained on Twitter data. This broader coverage makes the chosen model more suitable for academic contexts such as high school essay writing.

Table 1. Summary of semantic model

Component	Details
Pre-trained Model	IndoBERT (indobenchmark/indobert-base-p2)
Sequence Length	256
Hyperparameters	Learning rate: 2e-5; epoch: 10; batch: 16; LR patience: 2; seed: 42
Ratio	75:5:20

The model is trained using a triplet loss function, where student essays paired with other student essays serve as positive samples, while ChatGPT essays serve as negative samples (and vice versa). By fine-tuning with the Adam optimizer, the model minimized the distance between positive pairs and maximized the distance between negative pairs, enabling it to effectively learn the semantic separation between student-written and AI-generated texts. This Training was conducted using Python 3.12 under a Linux environment (WSL2 on Windows 11 Pro) using TensorFlow 2.19, and training was performed on a workstation equipped with an NVIDIA GeForce RTX 4060 Ti GPU, a 13th Gen Intel® Core™ i7-13700F processor, and 32 GB of RAM.

Embeddings and similarity score

After tokenizing each paragraph, embeddings are generated using the trained IndoBERT-based semantic model. These embeddings are then normalized using L2-normalization, and the semantic similarity score between paragraphs is calculated using cosine similarity for student-student pairs, student-chatgpt pairs, and vice versa. The paragraph-level embeddings and similarity scores were directly fed into the classification model without dimensionality reduction, serving as semantic representations that capture differences in coherence and meaning between human- and AI-generated texts.

Linguistic features

The linguistic feature extraction in this study is partially adapted from Mizumoto et al. [29] study with adjustments to suit Indonesian language. These features are selected to capture various aspects of writing structure, including lexical diversity, syntactic fluency, and semantic properties. All features used in this study are summarized in Table 2. The explanation of the parameters used is as follows:

(1) Lexical Diversity (LD): Measures the diversity of vocabulary used in essays. High LD indicates a more diverse use of vocabulary.
(2) Word Count: Represents the number of words in the essay. This measure reflects the writer's fluency and ability to explore their ideas.
(3) Unique Word Count: Indicates the number of different words used. A higher count indicates greater lexical variety.
(4) Modal Count: Measures the frequency of modal verbs (e.g., *dapat*, *harus*, *bisa*) using an Indonesian POS tagger, these words reflect the author's reasoning and argumentation.

(5) Stopword Ratio: Calculates the proportion of stopwords to total words. This feature can help differentiate between human and AI writing, as AI-generated text may use more function words to maintain fluency.
(6) Average Sentence Length: Reflects the average number of words per sentence. Longer sentences may imply more complex sentence construction.
(7) Sentence Length Variation: Captures the standard deviation in sentence lengths. Human writing typically exhibits more natural variation.
(8) Punctuation Ratio: Represents the proportion of punctuation to total words. Proper use of punctuation often correlates with coherent and well-structured writing.

Table 2. Parameters used for linguistic analysis of essays

Construct	Measure		Abbreviation
Lexical Diversity	1.	Lexical Diversity	LD
Fluency	2.	Total words in the essay	word_count
	3.	Total unique words*	unique_count
Semantic Property	4.	Modals	modal_count
	5.	Stopwords ratio*	stopword_ratio
Syntactic Fluency	6.	Average sentence length*	avg_sent_len
	7.	Sentence length variation*	sent_len_var
Writing Mechanics	8.	Punctuation Ratio*	punct_ratio

*Measures were not included in the study by Mizumoto et al. [29]

Modals (MD) were identified using the Indonesian part-of-speech (POS) tagger developed by Dinakaramani et al. [47], and a list of stopwords based on literature referring to the study by Fadillah Z. Tala, as applied in the research of Kurniawan et al. [48]. These linguistic features are analyzed to complement the semantic information by capturing stylistic and structural patterns that may differ between human-written and AI-generated essays, thereby enhancing the overall effectiveness of the classification model.

Classification model

The classification model integrates three input features, semantic embeddings from IndoBERT, similarity scores between student and ChatGPT essays, and the linguistic features described earlier. Each input enters through a dedicated layer according to its dimension, with 256 units for IndoBERT embeddings, 2 units for similarity scores, and 8 units for linguistic features. Each of these inputs is first processed through a small dense layer to normalize its scale, and the three outputs are then concatenated into a single combined representation. A sigmoid activation function is used in the output layer, as the task is binary classification (student-written or ChatGPT-generated).

Table 3. Summary of classification model

Component	Details
Input Shape	Text embeddings; similarity scores, linguistic features
Hyperparameters	Loss: binary cross entropy; optimizer: Adam; learning rate: 1e-4; epoch: 30; batch: 16; LR patience: 2
Ratio	75:5:20

The training setup is summarized in Table 3, with binary cross-entropy as the loss function and the Adam optimizer with a learning rate of 1e-4. Training runs for 30 epochs with early stopping applied using a learning rate scheduler with a patience of 3 to prevent overfitting. The dataset split of 75% training, 5% validation, and 20% testing was the same as defined earlier in Data Preparation, ensuring consistency across all modeling stages. The learning rate and batch size (16) were adjusted to produce stable convergence while accommodating the computational limits of the available GPU. This classification model framework is designed to combine embeddings, semantic similarities, and linguistic features to enhance the model's ability to effectively distinguish between student-written and ChatGPT-generated essays.

Evaluation

To evaluate the proposed classification model, we use widely used evaluation metrics, namely AUC-ROC, sensitivity-specificity tradeoff, accuracy, precision, recall, F1 score. AUC-ROC measure the classification performance of the model [49], [50] by representing the probability that a randomly selected ChatGPT-generated sample will receive a higher score than a randomly selected student-written sample. A sensitivity-specificity tradeoff analysis is performed to identify the optimal classification threshold by evaluating the

true positive ratio (sensitivity) and true negative ratio (specificity), This aims to reduce false positives and false negatives for accurate and fair differentiation between student-written and ChatGPT-generated essays [26].

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

$$F1\text{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

The evaluation metrics used in this study are Accuracy, Precision, Recall, and F1-score (1) - (4). These metrics provide an evaluation of the model's performance. Accuracy reflects the overall accuracy of the model's predictions in both classes. Precision measures the proportion of correct positive predictions among all instances classified as positive, indicating how reliable the model is in predicting student essays or ChatGPT-generated text. Recall measures the model's ability to correctly identify all true positive cases, ensuring that relevant examples are not overlooked. The F1-score, as the harmonic mean of precision and recall, provides a balanced and useful metric when dealing with class imbalance [51].

RESULTS AND DISCUSSIONS

This section discusses the results of the analysis between student-written essays and those created by ChatGPT in terms of semantic similarity, linguistic aspects used, and metrics produced in the model classification. All these findings are the subject of discussion that will be discussed at the end of the section.

Datasets

The student-written essay dataset consists of several columns, including *teks_esai*, *nama*, *asal_sekolah*, and *tahun*. Meanwhile, ChatGPT-generated essays were obtained using the OpenAI API with the ChatGPT-4o model through the chat completion endpoint using Python. Requests were made in line with the competition guidelines, and the chat completions were run twice to accommodate the instructions for the 2024 and 2025 essay competitions, as detailed earlier in the Data Collection section. To ensure representativeness, the number of AI-generated essays was matched to the 20 finalist essays written by students (10 from each year). Topics, essay length, and writing style were controlled by following the official competition guidelines, and the generated outputs were aligned with the minimum, maximum, and average paragraph counts observed in the student essays. This design ensured that topic variation, style, and length between student and AI essays were balanced, making the two sets directly comparable. Raw responses were cleaned by removing unnecessary markers such as ### and **. The cleaned results from both years were then combined into single dataset.

This study uses paragraph-level data to enable more detailed extraction and analysis of linguistic features. The final dataset consists of 327 essays written by students and 325 essays created by ChatGPT. To address class imbalance, random undersampling was applied during the splitting of the dataset with a split ratio of 75% training, 5% validation, and 20% testing, the number of texts generated in each subset is summarized in Table 4.

Table 4. Summary of dataset splitting

Class	Training Set	Validation Set	Test Set
Student	223	15	60
ChatGPT	223	15	67

This data splitting strategy ensures that the model has sufficient data for training, sufficient data for validation, and a representative test set to evaluate performance. Each paragraph in the dataset is tokenized

using the indobenchmark/indobert-base-p2 pre-trained model, which generates token outputs, input IDs, and attention masks, which are required to train the IndoBERT-based model.

In addition to semantic modeling, the dataset was further processed for linguistic feature extraction. This analysis is performed to assess the stylistic and structural characteristics of the language, allowing a deeper comparison between the student-written and the ChatGPT-generated essays at the level of language usage.

Similarity score comparison

The comparison of semantic similarity scores shown in Figure 2, categorized by source (student or ChatGPT) on test set. The left plot shows the similarity between student-student pairs and student-ChatGPT pairs, while the right plot shows the ChatGPT-student and ChatGPT-ChatGPT comparisons.

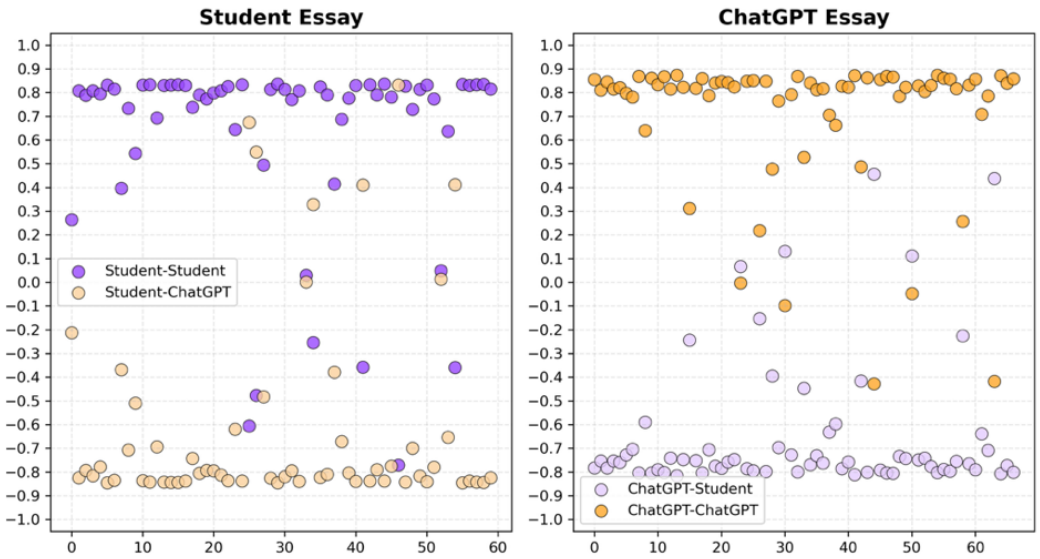

Figure 2. Semantic similarity scores between student and ChatGPT essays on the test set

The analysis shows that student pairs consistently demonstrate higher similarity scores, mostly clustered in the range of 0.7–0.9, which represents the lower bound of semantic similarity within the same category despite variations in style or vocabulary. A few outliers appeared, including cross-category pairs with unexpectedly high similarity. In contrast, student-ChatGPT pairs mostly ranged between –0.7 and –0.9, highlighting a clear semantic gap between human and AI-generated essays.

The same pattern is also seen in the right plot, where ChatGPT-ChatGPT pairs also show high similarity values, while ChatGPT-student pairs are still widely scattered with mostly negative values. These results confirm that student essays and ChatGPT form distinct semantic clusters and that the trained IndoBERT-based semantic model effectively captures the separation between the two writing styles. This clear separation supports the suitability of semantic similarity as a discriminative feature for the classification task in this study.

Linguistic features analysis

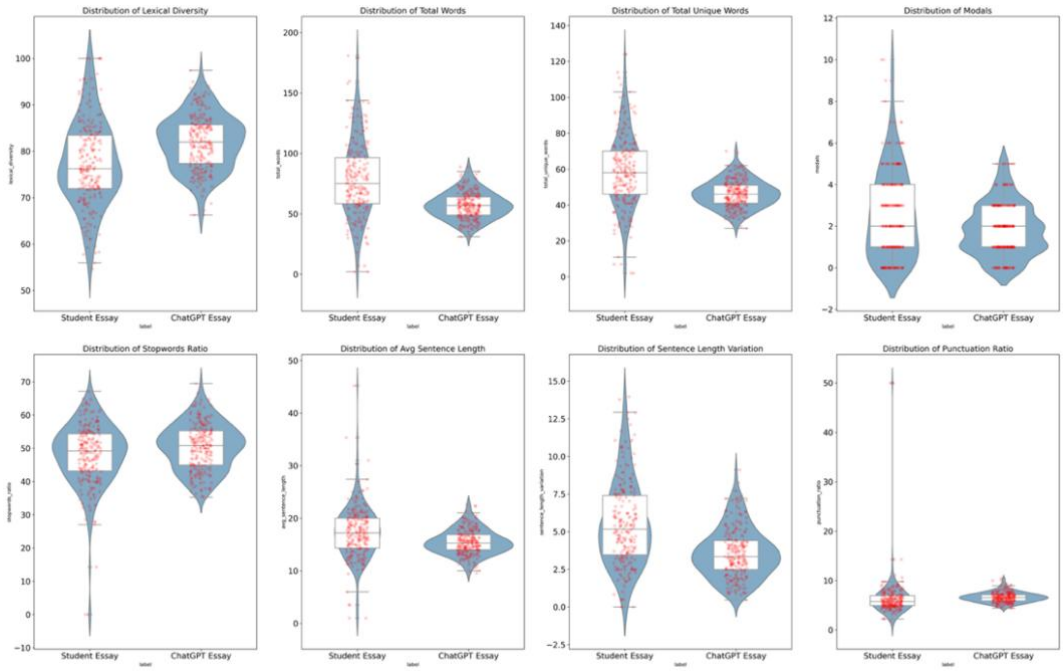

Figure 3. Visualization of the distribution of linguistic features on the test set for student and ChatGPT essays

The distribution of the eight linguistic features extracted from the test set is visualized in Figure 3, displaying the differences between the student-written essays and ChatGPT-generated essays. Student essays show greater variability in lexical diversity, word count, unique words, and sentence length variation, indicating a more diverse and less uniform writing style Lexical diversity and unique word counts were generally higher, reflecting a broader vocabulary range. In terms of modality, student essays also tend to use more modal verbs, indicating a stronger tendency towards reasoning and argumentative expression, which is consistent with human writing patterns. By contrast, ChatGPT-generated essays show a higher stopwords ratio and more uniform sentence structures, suggesting the model's reliance on filler words and its tendency toward concise, standardized phrasing. Punctuation use was also more varied among students, with a few outliers, reflecting creative use of mechanics to shape text flow and emphasis.

These findings support Mizumoto et al. [29], who demonstrated that lexical diversity, word counts, and modal usage are effective markers for distinguishing AI from human-generated texts. At the same time, this study extends their framework by including additional measures such as stopword ratio, average sentence length, sentence length variation, and punctuation ratio, which were not part of their analysis. The results therefore not only reinforce prior evidence that human essays tend to be more diverse and variable, but also highlight new linguistic cues that further separate student writing from ChatGPT output. Collectively, these patterns strengthen the theoretical basis for using linguistic features as discriminative markers and justify their integration into the classification model.

Classification evaluation

To evaluate the performance of the classification model, various metrics and visualizations are used to more easily understand the evaluation results. These include ROC curves, sensitivity-specificity tradeoff plots, and confusion matrix. Each visualization offers complementary insights into the effectiveness and reliability of the model in distinguishing between student-written and ChatGPT-generated essays.

The ROC curve shown in Figure 4, evaluates the model's ability to distinguish between two classes at various classification thresholds by plotting the true positive rate against the false positive rate. The model achieved an AUC of 0.9999 on the combined training and validation sets, which indicates near-perfect discriminative performance. While this very high value indicates strong internal consistency, it also raises the possibility of overfitting, particularly given the relatively small size of the test set. Therefore, the result should be interpreted with caution, and further evaluation on larger and more diverse datasets is necessary

to confirm the model's robustness. Next step is to analyze the tradeoffs between sensitivity and specificity to determine optimal balanced threshold.

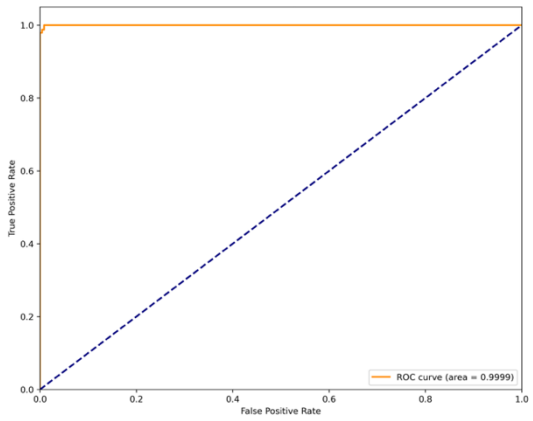

Figure 4. ROC curve on the combined training and validation sets

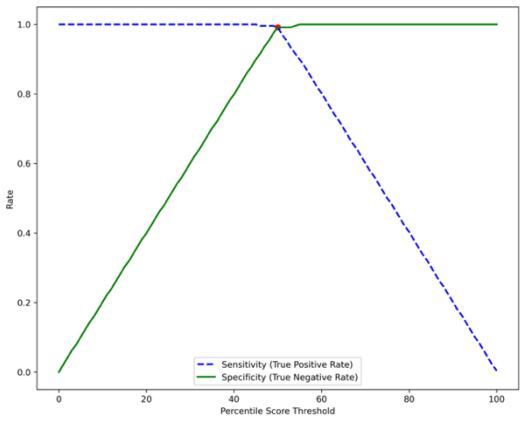

Figure 5. Tradeoff between sensitivity and specificity on the training and validation sets

To identify the most appropriate threshold for decision making, a sensitivity-specificity tradeoff analysis was performed and visualized in Figure 5. This curve depicts the inverse relationship between the true positive rate (sensitivity) and the true negative rate (specificity) as the classification threshold is varied. The optimal point where sensitivity and specificity are equal is found at the 50th percentile with a threshold value of 0.8604, reaching a rate of 0.9916 for both metrics. This threshold was chosen to balance false positive and false negative rates, ensuring fairer classification between student-written and ChatGPT-generated essays. This helps maintain consistency in identifying key differences between the two, balancing false positive and false negative results.

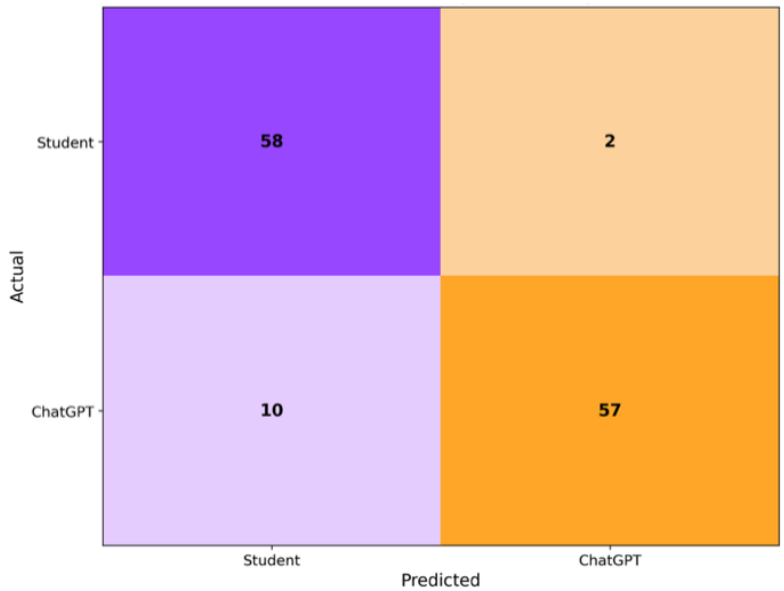

Figure 6. Confusion matrix on the test set

Further testing of the model's performance is presented in the confusion matrix in Figure 6, which summarizes the classification results on the test set held. The model correctly identified 58 out of 60 student essays and 57 out of 67 ChatGPT essays, resulting in a total accuracy of 90.55%. Also, this model achieved a precision at 96.61%, recall rate at 85.07%, and an F1-score at 90.48% in detecting ChatGPT-generated essays. These results also reflect high precision and recall rates for both classes, which reinforces the effectiveness of combining semantic similarity and linguistic features in the classification process.

Although there were a few misclassifications, where 2 student essays were predicted as ChatGPT and 10 ChatGPT essays were classified as student writing. These errors can be explained by overlapping linguistic and semantic characteristics. Outliers in semantic similarity also contributed, as some essays clustered closer to the opposite category in the embedding space. Student essays that are very formal with limited personal perspectives or repetitive sentence structures tend to produce feature vectors similar to ChatGPT output, leading to classification errors. In contrast, ChatGPT essays that use a broader vocabulary or more varied syntax sometimes align with human-like patterns. This suggests that semantic outliers together with linguistic features such as lexical diversity, modals, and stopwords played a key role in shaping the observed misclassification patterns.

These findings suggest that while the model performs well overall, it can be challenged by essays that blur stylistic boundaries between human and machine-generated texts. The ability of ChatGPT-4o to produce text that closely resembles human writing indicates the importance of incorporating contextual and content level features into AI detection models in future research.

Case analysis of misclassified essays

Table 5 summarizes each essay's classification confidence and its average semantic similarity score to the predicted class. To better understand the model, this section analyzes four essay samples that were misclassified by the classification model, two written by students but predicted as ChatGPT and two written by ChatGPT but classified as student essays.

The first student essay was classified as ChatGPT with high confidence level at 0.9303 and an average similarity score at 0.8309. This essay displays a skilled style, high semantic cohesion, and formal structure, a writing style that typically resembles ChatGPT-generated text. Similar pattern is showed in the second misclassified student essay, which, despite having a more personal tone, still maintains a structured format and limited lexical variation. These characteristics may have influenced the model to interpret the writing as AI-generated text, resulting in a confidence level of 0.8845 and a similarity score of 0.5487.

Table 5. Examples of misclassified essays

Essay Texts	Model Classification Confidence	Avg. Similarity
di era digital dan globalisasi, usaha menjadi sangat penting. teknologi memudahkan kita belajar dan berkembang tanpa batas geografis, tetapi juga meningkatkan persaingan. karena itu, usaha yang konsisten dan kerja keras diperlukan untuk memanfaatkan peluang digital dan mencapai keberhasilan. globalisasi turut menuntut tenaga kerja yang kompeten dan adaptif, sehingga peningkatan kualitas diri melalui pendidikan, keterampilan digital, dan kerja keras menjadi kunci sukses menghadapi tantangan global.	0.9303 (as ChatGPT)	0.8309 (to ChatGPT)
ada banyak bentuk usaha yang bisa kita lakukan dalam mengolah hidup, salah satunya adalah melalui pendidikan. pendidikan adalah ladang subur yang menghasilkan manfaat berkelanjutan. melalui usaha dalam belajar, kita dapat mengembangkan potensi, meningkatkan keterampilan, dan membuka peluang hidup yang lebih luas.	0.8845 (as ChatGPT)	0.5487 (to ChatGPT)
setiap orang memulai hidupnya dengan kanvas kosong yang perlahan akan diwarnai oleh berbagai peristiwa dan pengalaman. warna-warna yang kita pilih, pola dan bentuk yang tercipta, semuanya adalah manifestasi dari pilihan dan tindakan kita. ada kalanya kita dihadapkan pada warna-warna suram, mungkin akibat kesedihan atau kegagalan, namun kita punya kuasa untuk memilih, apakah warna suram itu akan menguasai kanvas kita atau kita akan menambahkan warna lain yang lebih cerah? pilihan ini mempengaruhi bagaimana kita menghargai hidup kita sendiri.	0.8875 (as Student)	0.4376 (to Student)
kesadaran lingkungan adalah dimensi lain dalam usaha mengolah hidup. di tengah kerusakan ekologi yang kian nyata, pemuda dituntut untuk bergerak bersama, berpikir kritis dan bertindak nyata ancaman terhadap lingkungan alam. menyadari bahwa merekalah bagian dari alam yang lebih besar adalah awal dari pembentukan kesadaran ekologis. mereka didorong untuk berpartisipasi dalam kegiatan pelestarian alam, seperti penghijauan dan penanganan limbah, sehingga dapat meninggalkan bumi yang lebih baik bagi generasi mendatang. ketika setiap individu menerima tanggung jawab ekologis dan menjadikannya bagian dari hidup, maka kekhawatiran tentang masa depan yang suram berubah menjadi harapan kolektif.	0.8577 (as Student)	0.4556 (to Student)

Conversely, the third and fourth essays were generated by ChatGPT but misclassified as student writing. Although both essays are thematically rich, they use more varied sentence structures and stylistic expressions than typical AI output. The third essay, for example, uses metaphorical language and introspective narrative ("hidup seperti kanvas kosong"), potentially mimicking human creativity. It was classified as written by a student with a confidence level of 0.8875, despite an average similarity score of only 0.4376 with the student class. The fourth AI-generated essay also integrated socially conscious language and subtle rhetorical cues, resulting in a confidence score of 0.8577 as written by a student and a similarity score of 0.4556.

These findings highlight the differences between essays written by students and those generated by ChatGPT, specifically in context of semantic coherence and stylistic diversity. The model's classification performance with its distinctive semantic clustering and linguistic patterns, suggests that AI-generated text, while fluent, lacks the variety of expression typical of human writing. Classification errors, especially when ChatGPT imitates human writing style or when students use formal and impersonal styles, indicate the limitations of relying solely on surface features. Compared to studies in English, this work shows that IndoBERT embeddings with linguistic cues adapt well to the Indonesian language. Practically, the results can assist educators in essay evaluation, raise awareness of AI reliance in learning, and guide policy on academic integrity, while ongoing research focuses on larger datasets, additional AI models, and discourse-level features to strengthen robustness.

CONCLUSION

This study compares essays written by students and essays generated by ChatGPT in Indonesian language by analyzing their semantic similarities and linguistic features. The IndoBERT-based model with triplet loss revealed distinct semantic similarity clusters for each source, while linguistic features such as lexical diversity, sentence length variation, and modal usage highlighted further stylistic differences. All these features are combined for a classification model and achieved an accuracy of 90.55%.

However, this study focuses on a single AI model (ChatGPT-4o), further study could involve expanding the dataset, integrating other generative models, and exploring more advanced representations or richer language styles to enhance robustness and generalization across domains and languages.

REFERENCES

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[2] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Physical Syst.*, vol. 3, pp. 121–154, 2023, doi: 10.1016/j.iotcps.2023.04.003.

[3] A. R. Malik *et al.*, "Exploring Artificial Intelligence in Academic Essay: Higher Education Student's Perspective," *Int. J. Educ. Res. Open*, vol. 5, p. 100296, Dec. 2023, doi: 10.1016/j.ijedro.2023.100296.

[4] S. Lin and P. Crosthwaite, "The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback," *System*, vol. 127, p. 103529, Dec. 2024, doi: 10.1016/j.system.2024.103529.

[5] A. Kabir, S. Shah, A. Haddad, and D. M. S. Raper, "Introducing Our Custom GPT: An Example of the Potential Impact of Personalized GPT Builders on Scientific Writing," *World Neurosurg.*, vol. 193, pp. 461–468, Jan. 2025, doi: 10.1016/j.wneu.2024.10.041.

[6] S. P. T. Utami, A. Andayani, R. Winarni, and S. Sumarwati, "Utilization of artificial intelligence technology in an academic writing class: How do Indonesian students perceive?," *Contemp. Educ. Technol.*, vol. 15, no. 4, p. ep450, Oct. 2023, doi: 10.30935/cedtech/13419.

[7] A. Haleem, M. Javaid, and R. P. Singh, "An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil Trans. Benchmarks, Stand. Eval.*, vol. 2, no. 4, p. 100089, Oct. 2022, doi: 10.1016/j.tbench.2023.100089.

[8] B. C. Stahl and D. Eke, "The ethics of ChatGPT – Exploring the ethical issues of an emerging technology," *Int. J. Inf. Manage.*, vol. 74, p. 102700, Feb. 2024, doi: 10.1016/j.ijinfomgt.2023.102700.

[9] A. Pahrudin *et al.*, "The Effectiveness of Science, Technology, Engineering, and Mathematics-Inquiry Learning for 15-16 Years Old Students Based on K-13 Indonesian Curriculum: The Impact on the Critical Thinking Skills," *Eur. J. Educ. Res.*, vol. volume-10-, no. volume-10-issue-2-april-

2021, pp. 681–692, Apr. 2021, doi: 10.12973/eu-jer.10.2.681.
[10] Z. U. Irma, S. Kusairi, and L. Yuliati, "STREM PBL with E-Authentic Assessment: Its Impact to Students' Scientific Creativity on Static Fluid," *J. Pendidik. IPA Indones.*, vol. 12, no. 1, pp. 80–95, Mar. 2023, doi: 10.15294/jpii.v12i1.40214.
[11] A. Supriyono and T. Prihandono, "Dampak dan tantangan pemanfaatan ChatGPT dalam pembelajaran pada kurikulum merdeka: Tinjauan literatur sistematis," *J. Pendidik. dan Kebud.*, vol. 9, no. 2, pp. 134–152, 2024, doi: 10.24832/jpnk.v9i2.5214.
[12] Juanda and I. Afandi, "Assessing text comprehension proficiency: Indonesian higher education students vs ChatGPT," *XLinguae*, vol. 17, no. 1, pp. 49–68, Jan. 2024, doi: 10.18355/XL.2024.17.01.04.
[13] A. O. Ajlouni, F. Abd-Alkareem Wahba, and A. Salem Almahaireh, "Students' Attitudes Towards Using ChatGPT as a Learning Tool: The Case of the University of Jordan," *Int. J. Interact. Mob. Technol.*, vol. 17, no. 18, pp. 99–117, Sep. 2023, doi: 10.3991/ijim.v17i18.41753.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[15] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *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*, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.
[17] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/2826773.
[18] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Appl. Sci.*, vol. 15, no. 1, p. 172, Dec. 2024, doi: 10.3390/app15010172.
[19] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.
[20] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, p. 1071, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
[21] L. Rodrigues, F. Dwan Pereira, L. Cabral, D. Gašević, G. Ramalho, and R. Ferreira Mello, "Assessing the quality of automatic-generated short answers using GPT-4," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100248, Dec. 2024, doi: 10.1016/j.caeai.2024.100248.
[22] H. B. Essel, D. Vlachopoulos, A. B. Essuman, and J. O. Amankwa, "ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs)," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100198, Jun. 2024, doi: 10.1016/j.caeai.2023.100198.
[23] V. Bellini, F. Semeraro, J. Montomoli, M. Cascella, and E. Bignami, "Between human and AI: assessing the reliability of AI text detection tools," *Curr. Med. Res. Opin.*, vol. 40, no. 3, pp. 353–358, Mar. 2024, doi: 10.1080/03007995.2024.2310086.
[24] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," *Int. J. Educ. Integr.*, vol. 19, no. 1, p. 17, Sep. 2023, doi: 10.1007/s40979-023-00140-5.
[25] H. Hua and C.-J. Yao, "Investigating generative AI models and detection techniques: impacts of tokenization and dataset size on identification of AI-generated text," *Front. Artif. Intell.*, vol. 7, Nov. 2024, doi: 10.3389/frai.2024.1469197.
[26] R. An, Y. Yang, F. Yang, and S. Wang, "Use prompt to differentiate text generated by ChatGPT and humans," *Mach. Learn. with Appl.*, vol. 14, p. 100497, Dec. 2023, doi: 10.1016/j.mlwa.2023.100497.
[27] J. Campino, "Unleashing the transformers: NLP models detect AI writing in education," *J. Comput. Educ.*, vol. 12, no. 2, pp. 645–673, Jun. 2025, doi: 10.1007/s40692-024-00325-y.
[28] D. Viji and S. Revathy, "A hybrid approach of Weighted Fine-Tuned BERT extraction with deep

Siamese Bi – LSTM model for semantic text similarity identification," *Multimed. Tools Appl.*, vol. 81, no. 5, pp. 6131–6157, Feb. 2022, doi: 10.1007/s11042-021-11771-6.
[29] A. Mizumoto, S. Yasuda, and Y. Tamura, "Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints," *Appl. Corpus Linguist.*, vol. 4, no. 3, p. 100106, Dec. 2024, doi: 10.1016/j.acorp.2024.100106.
[30] I. F. Emara, "A linguistic comparison between ChatGPT-generated and nonnative student-generated short story adaptations: a stylometric approach," *Smart Learn. Environ.*, vol. 12, no. 1, p. 36, May 2025, doi: 10.1186/s40561-025-00388-z.
[31] S. Herbold, A. Hautli-Janisz, U. Heuer, Z. Kikteva, and A. Trautsch, "A large-scale comparison of human-written versus ChatGPT-generated essays," *Sci. Rep.*, vol. 13, no. 1, p. 18617, Oct. 2023, doi: 10.1038/s41598-023-45644-9.
[32] S. Dehghan and M. F. Amasyali, "SupMPN: Supervised Multiple Positives and Negatives Contrastive Learning Model for Semantic Textual Similarity," *Appl. Sci.*, vol. 12, no. 19, p. 9659, Sep. 2022, doi: 10.3390/app12199659.
[33] T. Kim, K. M. Yoo, and S. Lee, "Self-Guided Contrastive Learning for BERT Sentence Representations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2528–2540. doi: 10.18653/v1/2021.acl-long.197.
[34] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6894–6910. doi: 10.18653/v1/2021.emnlp-main.552.
[35] Z. Ren, Q. Lan, Y. Zhang, and S. Wang, "Exploring simple triplet representation learning," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 1510–1521, Dec. 2024, doi: 10.1016/j.csbj.2024.04.004.
[36] K. Zhang *et al.*, "Description-Enhanced Label Embedding Contrastive Learning for Text Classification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 10, pp. 14889–14902, Oct. 2024, doi: 10.1109/TNNLS.2023.3282020.
[37] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, "Fine-Tuned IndoBERT Based Model and Data Augmentation for Indonesian Language Paraphrase Identification," *Rev. d'Intelligence Artif.*, vol. 37, no. 3, pp. 733–743, Jun. 2023, doi: 10.18280/ria.370322.
[38] L. Attal, E. Shvartz, N. Nakhoul, and D. Bahir, "Chat GPT 4o vs residents: French language evaluation in ophthalmology," *AJO Int.*, vol. 2, no. 1, p. 100104, Apr. 2025, doi: 10.1016/j.ajoint.2025.100104.
[39] K. SARAWAN, J. POLPINIJ, G. SOMPRASERTSRI, A. ROJARATH, and B. LUAPHOL, "MULTICLASS CLASSIFICATION APPROACH FOR DETECTING SOFTWARE BUG SEVERITY LEVEL FROM BUG REPORTS," *ICIC express Lett. Part B, Appl. an Int. J. Res. Surv.*, vol. 16, no. 5, pp. 567–576, 2025, [Online]. Available: https://ndlsearch.ndl.go.jp/books/R000000004-I034249118
[40] L. Alrajhi, A. Alamri, F. D. Pereira, A. I. Cristea, and E. H. T. Oliveira, "Solving the imbalanced data issue: automatic urgency detection for instructor assistance in MOOC discussion forums," *User Model. User-adapt. Interact.*, vol. 34, no. 3, pp. 797–852, Jul. 2024, doi: 10.1007/s11257-023-09381-y.
[41] R. Olusegun, T. Oladunni, H. Audu, Y. Houkpati, and S. Bengesi, "Text Mining and Emotion Classification on Monkeypox Twitter Dataset: A Deep Learning-Natural Language Processing (NLP) Approach," *IEEE Access*, vol. 11, pp. 49882–49894, 2023, doi: 10.1109/ACCESS.2023.3277868.
[42] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, p. 723, Jun. 2024, doi: 10.1057/s41599-024-03209-9.
[43] Y. Hu, J. Ding, Z. Dou, and H. Chang, "Short-Text Classification Detector: A Bert-Based Mental Approach," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/8660828.
[44] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, Sep. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.
[45] A. Kurniasih and L. P. Manik, "On the Role of Text Preprocessing in BERT Embedding-based DNNs for Classifying Informal Texts," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, 2022, doi: 10.14569/IJACSA.2022.01306109.
[46] E. Alzahrani and L. Jololian, "How Different Text-Preprocessing Techniques using the Bert Model

Affect the Gender Profiling of Authors," in *Advances in Machine Learning*, Sep. 2021, pp. 01–08. doi: 10.5121/csit.2021.111501.
[47] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," in *2014 International Conference on Asian Language Processing (IALP)*, Oct. 2014, pp. 66–69. doi: 10.1109/IALP.2014.6973519.
[48] A. A. Kurniawan, S. Madenda, S. Wirawan, and R. J. Suhatril, "Multidisciplinary classification for Indonesian scientific articles abstract using pre-trained BERT model," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 2, p. 331, Jul. 2023, doi: 10.26555/ijain.v9i2.1051.
[49] N. Borah, U. Baruah, M. Thylore Ramakrishna, V. V. Kumar, D. R. Dorai, and J. Rajkumar Annad, "Efficient Assamese Word Recognition for Societal Empowerment: A Comparative Feature-Based Analysis," *IEEE Access*, vol. 11, pp. 82302–82326, 2023, doi: 10.1109/ACCESS.2023.3301564.
[50] T. Spinde *et al.*, "Automated identification of bias inducing words in news articles using linguistic and context-oriented features," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102505, May 2021, doi: 10.1016/j.ipm.2021.102505.
[51] T. I. Sari, Z. N. Ardilla, N. Hayatin, and R. Maskat, "Abusive comment identification on Indonesian social media data using hybrid deep learning," *IAES Int. J. Artif. Intell.*, vol. 11, no. 3, p. 895, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp895-904.