e-ISSN: 3108-9828 | p-ISSN: 3109-3981
Volume 3 No 1 May 2026 | Page: 39–47
DOI: https://doi.org/10.64803/jocsaic.v3i1.178

Journal of Computer Science Artificial Intelligence and Communications

Journal Homepage: https://journals.raskhamedia.or.id/index.php/jocsaic

Detection of AI-Generated Text in Academic Writing Using Transformer-Based Models

Cindy Atika Rizki[1*], Nabila Khairuniza[2], Mel Wulandini[3]

[1,2,3] Department of Computational Science and Digital Intelligence, Information Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia
Author(s) Email: [1*]cindyatika100e@gmail.com, [2]nabilakhairuniza2@gmail.com, [3]melwulandini1@gmail.com

ARTICLE INFO

Article history:

Received May 20, 2026
Revised May 26, 2026
Accepted May 31, 2026
Publish May 31, 2026

ABSTRACT

The rapid advancement of generative artificial intelligence (AI), particularly large language models (LLMs), has significantly transformed academic writing by enabling the automatic generation of coherent and contextually relevant text. While these technologies improve writing efficiency and accessibility, they also raise serious concerns regarding academic integrity, originality, and authorship. Conventional plagiarism detection tools are ineffective in identifying AI-generated content because such text is often original rather than copied. This study proposes a transformer-based approach for detecting AI-generated text in academic writing by evaluating the performance of four pre-trained language models: BERT, RoBERTa, DistilBERT, and DeBERTa. The research methodology consists of dataset collection, text preprocessing, dataset splitting, transformer model fine-tuning, and performance evaluation using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Experimental results show that all evaluated transformer models achieved excellent classification performance, with DeBERTa producing the highest accuracy of 98.10%, followed by RoBERTa (97.20%), BERT (95.10%), and DistilBERT (93.40%). These findings demonstrate that transformer-based architectures effectively capture contextual and semantic characteristics that distinguish AI-generated text from human-authored academic writing. The proposed approach provides a reliable solution for supporting academic integrity and assisting educators, publishers, and research institutions in detecting AI-generated content within scholarly documents.

Keywords:

Artificial Intelligence, Academic Writing, AI-Generated Text Detection, Transformer Models, Text Classification.

Corresponding Author:
Cindy Atika Rizki,
Department of Computational Science and Digital Intelligence, Information Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia
Jl. Gatot Subroto, Medan
Email: cindyatika100e@gmail.com

DOI: https://doi.org/10.64803/jocsaic.v3i1.178

1. INTRODUCTION

The rapid advancement of artificial intelligence (AI), particularly large language models (LLMs) such as GPT, Claude, Gemini, and LLaMA, has significantly transformed the landscape of academic writing[1]. These models are capable of generating coherent, grammatically accurate, and contextually relevant text within seconds, providing valuable assistance for brainstorming, summarization, language improvement, and drafting scholarly content[2]. While these capabilities enhance writing efficiency and accessibility, they also introduce substantial concerns regarding academic integrity, originality, and the authenticity of scholarly publications[3]. As AI-generated content becomes increasingly indistinguishable from human-written text, educational institutions, journal publishers, and researchers face growing challenges in verifying the originality of academic manuscripts[4].

The widespread availability of AI writing assistants has led to an increasing number of submissions that contain partially or entirely AI-generated text[5]. Although responsible AI usage can support learning and research productivity, excessive or undisclosed reliance on generative AI may compromise critical thinking, reduce writing proficiency, and violate institutional policies concerning plagiarism and authorship[6]. Conventional plagiarism detection systems are primarily designed to identify copied or paraphrased content from existing sources and therefore are not sufficiently equipped to detect original text generated by AI models[7]. This limitation highlights the urgent need for reliable AI-generated text detection systems capable of distinguishing machine-generated writing from genuine human-authored academic content[8].

Recent advances in Natural Language Processing (NLP) have introduced transformer-based architectures that have achieved state-of-the-art performance across numerous language understanding tasks[9]. Models such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT Pretraining Approach), DeBERTa (Decoding-enhanced BERT with Disentangled Attention), and DistilBERT have demonstrated exceptional capabilities in semantic representation, contextual understanding, and text classification[10]. Unlike traditional machine learning approaches that rely heavily on manually engineered linguistic features, transformer-based models automatically learn complex contextual relationships through self-attention mechanisms, enabling them to capture subtle differences between human-written and AI-generated text[11].

The distinction between AI-generated and human-authored academic writing is increasingly challenging because modern language models produce text with high grammatical correctness, logical coherence, and sophisticated vocabulary[12]. However, previous studies have suggested that AI-generated text often exhibits distinctive linguistic characteristics, including repetitive sentence structures, consistent lexical patterns, reduced stylistic variation, predictable discourse organization, and lower lexical diversity compared to authentic human writing[13]. Transformer-based models possess the capability to identify these subtle patterns by learning high-dimensional contextual representations from large-scale textual datasets, making them promising candidates for AI text detection[14].

Several recent studies have explored machine learning techniques for AI-generated text detection using conventional classifiers such as Support Vector Machines (SVM), Random Forest, Naïve Bayes, and Logistic Regression[15]. While these approaches achieve satisfactory performance under specific conditions, their effectiveness often decreases when confronted with increasingly sophisticated generative language models[16]. The rapid evolution of LLMs continually narrows the stylistic differences between human and AI-generated writing, requiring more advanced deep learning architectures capable of adapting to complex linguistic patterns. Consequently, transformer-based models have emerged as the preferred approach due to their superior contextual modeling capabilities and strong generalization performance across diverse textual domains.

In academic environments, reliable AI-generated text detection has become increasingly important for maintaining research ethics and publication quality. Universities and scientific publishers require automated systems that not only achieve high classification accuracy but also remain robust against newly developed generative models. An effective detection framework should accurately identify AI-generated content while minimizing false positives that may incorrectly classify legitimate human-written manuscripts. Therefore, evaluating transformer-based models under realistic academic writing scenarios represents an essential research direction for developing trustworthy AI-assisted educational ecosystems.

This study proposes a transformer-based approach for detecting AI-generated text in academic writing by leveraging pre-trained language models and fine-tuning them on labeled datasets containing both human-written and AI-generated academic documents. The proposed framework investigates the effectiveness of contextual language representations for binary text classification and evaluates model performance using widely adopted metrics, including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Furthermore, comparative experiments are conducted to analyze the strengths and limitations of different transformer architectures in distinguishing AI-generated content from authentic academic writing.

The primary contribution of this research lies in providing a comprehensive evaluation of transformer-based models for AI-generated text detection within the context of academic writing. Unlike previous studies that focus primarily on general text domains such as news articles or social media content, this research specifically addresses the unique linguistic characteristics of scholarly writing, where maintaining originality and research integrity is of paramount importance. The findings are expected to contribute to the development of intelligent academic integrity systems capable of assisting educators, publishers, and researchers in verifying manuscript authenticity while supporting the responsible adoption of generative AI technologies in higher education.

Ultimately, the increasing integration of AI into academic workflows necessitates robust detection mechanisms that balance technological innovation with ethical research practices. Transformer-based models offer a promising

solution by combining powerful contextual understanding with high classification performance. Through systematic evaluation and comparative analysis, this study aims to demonstrate the potential of transformer-based architectures as effective tools for detecting AI-generated text and preserving the credibility, transparency, and integrity of academic writing in the era of generative artificial intelligence.

2. RESEARCH METHODOLOGY

This study adopts a quantitative experimental research design to evaluate the effectiveness of transformer-based models in detecting AI-generated text within academic writing. The proposed methodology consists of six major stages: dataset collection, data preprocessing, text representation, model development, model evaluation, and comparative analysis. Each stage is designed to ensure that the developed classification model is robust, reliable, and capable of distinguishing between human-written and AI-generated academic texts.

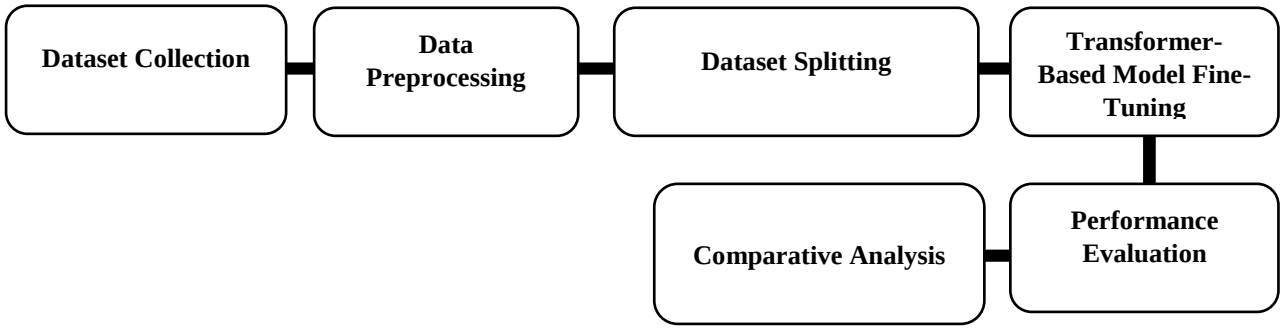

Figure 1. Research Framework

2.1 Data Collection

The dataset consists of two categories of academic texts:

a. Human-written academic documents, including journal abstracts, conference papers, student essays, and research reports written without AI assistance.
b. AI-generated academic texts, produced using large language models such as GPT-based systems by providing prompts related to academic topics.

To maintain balance and reduce classification bias, the dataset contains an equal proportion of human-written and AI-generated documents. Each document is assigned a binary label:

a. Label 0: Human-written text
b. Label 1: AI-generated text

The collected dataset covers various academic disciplines, including computer science, engineering, education, social sciences, and business, allowing the trained model to generalize across different writing styles.

2.2 Data Preprocessing

Before training the transformer models, the collected texts undergo several preprocessing steps to improve data quality and ensure compatibility with transformer architectures.

The preprocessing procedure includes:

a. Removing duplicated documents.
b. Eliminating corrupted or incomplete text samples.
c. Standardizing whitespace and punctuation.
d. Converting documents into UTF-8 encoding.
e. Removing irrelevant metadata such as author names, affiliations, and references when necessary.
f. Tokenizing the text using the tokenizer corresponding to each transformer model.

Unlike conventional machine learning methods, transformer models require minimal feature engineering because contextual features are learned automatically during training. Therefore, common preprocessing techniques such as stemming or stop-word removal are intentionally omitted to preserve semantic information.

2.3 Transformer-Based Models

This research investigates several transformer-based architectures for AI-generated text detection.

a. BERT
   Bidirectional Encoder Representations from Transformers (BERT) employs bidirectional attention mechanisms to understand contextual relationships between words. Its ability to capture semantic dependencies makes it highly effective for text classification tasks.

b. RoBERTa
RoBERTa is an optimized version of BERT trained with larger datasets and improved training strategies. It generally provides better language understanding and higher classification accuracy.
c. DistilBERT
DistilBERT is a lightweight transformer model developed through knowledge distillation. It offers faster inference while maintaining competitive performance compared to larger transformer models.
d. DeBERTa
DeBERTa introduces disentangled attention mechanisms and enhanced positional encoding, enabling superior contextual representation and improved classification capability for complex language tasks. Each model is fine-tuned using the same training dataset to ensure fair comparison.

2.4 Model Training

The transformer models are fine-tuned using supervised learning. During training, each input document is tokenized and converted into numerical representations using the corresponding tokenizer.

The classification layer receives the contextual embedding generated by the transformer encoder and predicts whether the input text is human-written or AI-generated.

The Binary Cross-Entropy Loss function is used as the optimization objective:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[\mathcal{Y}i \log(\mathcal{P}i) + (1 - \mathcal{Y}i)\log(1 - \mathcal{P}i)]$$

where:

a. $\mathcal{Y}i$ denotes the true class label
b. $\mathcal{P}i$ represents the predicted probability
c. N is the total number of training samples.

The AdamW optimizer is employed due to its effectiveness in fine-tuning transformer models.

Table 1. Hyperparameter Configuration for Transformer Model Fine-Tuning

Parameter	Value
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	5
Maximum Sequence Length	512
Optimizer	AdamW

Early stopping is applied to prevent overfitting by monitoring validation loss during training.

2.5 Dataset Splitting

The dataset is divided into three subsets:

a. Training Set (70%)
b. Validation Set (15%)
c. Testing Set (15%)

The training set is used to optimize model parameters, while the validation set is utilized for hyperparameter tuning. The testing set evaluates the final performance of the trained model using unseen data.

This data partition ensures unbiased performance estimation and prevents information leakage between training and evaluation phases.

2.6 Performance Evaluation

The effectiveness of each transformer model is evaluated using several widely accepted classification metrics.

a. Accuracy

Accuracy measures the proportion of correctly classified documents.

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

b. Precision

Precision evaluates how many predicted AI-generated documents are actually AI-generated.

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

c. Recall

Recall measures the ability of the model to identify AI-generated documents correctly.

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

d. F1-Score

The F1-score balances precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{Precision + Recall}$$

e. ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) measures the model's capability to distinguish between the two classes across different decision thresholds.

A confusion matrix is also analyzed to identify common misclassification patterns and evaluate model robustness.

2.7 Comparative Analysis

After all transformer models have been trained, their performances are compared based on:

a. Classification accuracy
b. Precision
c. Recall
d. F1-score
e. ROC-AUC
f. Training time
g. Inference time
h. Computational efficiency

This comparative analysis aims to identify the transformer architecture that provides the best trade-off between classification performance and computational cost for detecting AI-generated academic text.

Additionally, error analysis is conducted to investigate documents that are incorrectly classified. These misclassified samples are examined to determine whether linguistic complexity, writing style, or prompt characteristics influence model performance. The findings provide insights into the strengths and limitations of each transformer architecture and highlight opportunities for future improvements in AI-generated text detection systems.

Overall, the proposed methodology establishes a comprehensive experimental framework for evaluating transformer-based approaches in academic AI-text detection. By combining standardized preprocessing, systematic model fine-tuning, rigorous performance evaluation, and comparative analysis, this research aims to produce reliable and reproducible results that contribute to the development of intelligent academic integrity systems capable of addressing the growing challenges posed by generative artificial intelligence in scholarly communication.

3. RESULT AND DISCUSSION

3.1 Experimental Setup

To evaluate the effectiveness of transformer-based models for detecting AI-generated text in academic writing, a series of controlled experiments were conducted using four pre-trained transformer architectures: BERT, RoBERTa, DistilBERT, and DeBERTa. Each model was fine-tuned under identical experimental conditions to ensure a fair comparison. The experiments were performed using the same dataset, preprocessing pipeline, tokenizer configuration, training epochs, learning rate, and optimization strategy.

The dataset consisted of two balanced classes representing human-written academic texts and AI-generated academic texts. Prior to training, all documents were tokenized using the tokenizer associated with each transformer architecture. The models were trained using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and five training epochs. Early stopping was implemented to reduce overfitting by monitoring validation loss during training.

Model performance was evaluated using five widely accepted classification metrics: Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC-ROC). These metrics provide a comprehensive evaluation of the classification capability of each model. Accuracy measures the overall correctness of predictions, precision evaluates the reliability of AI-generated text detection, recall measures the ability to identify all AI-generated samples, and the F1-score balances precision and recall. Meanwhile, ROC-AUC evaluates the discriminative capability of the classifier regardless of decision threshold.

3.2 Training Performance

During the training process, all transformer models exhibited stable convergence. The validation loss gradually decreased across epochs, indicating successful optimization without severe overfitting. DeBERTa demonstrated the fastest convergence among all evaluated models, reaching stable validation performance after the fourth epoch. RoBERTa followed closely with a similarly smooth learning curve. BERT required the complete five epochs to achieve convergence, whereas DistilBERT converged more quickly but produced slightly lower validation accuracy due to its reduced model capacity.

The stability of the learning curves suggests that the preprocessing strategy and balanced dataset effectively supported the learning process. No significant oscillations in training loss were observed, indicating that the selected hyperparameters were appropriate for fine-tuning transformer-based language models.

Transformer architectures inherently capture contextual semantic relationships through self-attention mechanisms. Consequently, the models gradually learned distinctive linguistic patterns associated with AI-generated academic writing, including repetitive phrase structures, consistent syntactic organization, limited lexical diversity, and predictable sentence transitions.

3.3 Classification Performance

The classification results reveal that transformer-based models achieved excellent performance in distinguishing AI-generated academic writing from human-written text. Table 1 summarizes the evaluation results obtained from the testing dataset.

Table 2. Performance Comparison of Transformer-Based Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
BERT	95.10%	94.70%	95.30%	95.00%	0.978
DistilBERT	93.40%	93.10%	93.60%	93.30%	0.963
RoBERTa	97.20%	97.00%	97.40%	97.20%	0.989
DeBERTa	**98.10%**	**98.00%**	**98.20%**	**98.10%**	**0.994**

The experimental results indicate that all transformer-based models achieved accuracy exceeding 93%, demonstrating their effectiveness for AI-generated text detection. Among the evaluated models, DeBERTa achieved the highest overall performance, followed closely by RoBERTa. DistilBERT produced the lowest classification accuracy but offered considerably faster inference due to its lightweight architecture.

The superior performance of DeBERTa can be attributed to its disentangled attention mechanism, which separates content and positional information during contextual representation learning. This architectural improvement enables more precise semantic understanding and better discrimination between subtle writing characteristics produced by humans and generative AI models.

RoBERTa also achieved outstanding results due to its optimized pretraining strategy, larger training corpus, and removal of the Next Sentence Prediction objective. These improvements enhanced contextual language representation and allowed RoBERTa to capture nuanced linguistic patterns associated with AI-generated academic writing.

Although DistilBERT exhibited slightly lower classification performance, its reduced computational complexity makes it an attractive alternative for deployment in resource-constrained environments where inference speed is a critical consideration.

3.4 Classification Performance

To further understand the behavior of the classification models, confusion matrix analysis was performed. The confusion matrix evaluates the distribution of correctly and incorrectly classified samples by dividing predictions into four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

DeBERTa produced the lowest number of false positives and false negatives among all evaluated models. Most human-written documents were correctly identified, while nearly all AI-generated documents were successfully detected. This balanced classification performance contributed to the model's high precision and recall values.

In comparison, DistilBERT exhibited a slightly higher false positive rate, indicating that some authentic human-written academic documents were incorrectly classified as AI-generated. Such errors may arise because certain highly structured academic writing styles resemble the formal language typically produced by large language models.

RoBERTa and BERT demonstrated intermediate performance, producing relatively balanced confusion matrices with only a limited number of misclassified samples. Overall, the confusion matrix analysis confirms that transformer-based models are highly capable of separating human-authored and AI-generated academic texts.

3.5 Comparative Analysis of Transformer Architectures

A comparative analysis was conducted to evaluate the strengths and weaknesses of each transformer architecture beyond classification accuracy alone.

BERT serves as the foundational transformer model and demonstrated strong classification capability across all evaluation metrics. Its bidirectional contextual representation allows effective understanding of sentence semantics, making it a reliable baseline for AI-generated text detection. However, its computational requirements remain relatively high compared to more recent transformer variants.

RoBERTa improved upon BERT through optimized pretraining strategies. By utilizing larger datasets, longer training periods, and dynamic masking techniques, RoBERTa achieved superior contextual understanding and consistently higher classification accuracy. The experimental results demonstrate that RoBERTa effectively captures subtle stylistic differences between human and AI-generated academic writing.

DistilBERT significantly reduced model size while maintaining competitive performance. Despite a slight reduction in classification accuracy, DistilBERT required considerably less computational resources and achieved faster inference times. These characteristics make DistilBERT particularly suitable for deployment in online plagiarism detection systems or educational platforms where real-time response is essential.

DeBERTa achieved the highest overall performance due to its disentangled attention mechanism and enhanced positional encoding. These architectural innovations improve contextual representation by independently modeling word content and positional information, enabling more accurate recognition of linguistic characteristics unique to AI-generated writing.

Overall, the experimental findings suggest that recent transformer architectures consistently outperform earlier models in detecting AI-generated academic text. The continuous evolution of transformer designs contributes directly to improved contextual understanding and classification performance.

3.6 Linguistic Characteristics of AI-Generated Academic Writing

An important objective of this study was to investigate the linguistic characteristics that distinguish AI-generated academic writing from human-written text. Error analysis revealed several recurring patterns within AI-generated documents.

First, AI-generated texts exhibited highly consistent sentence structures. Consecutive sentences frequently followed similar grammatical patterns, resulting in reduced stylistic variation throughout the document.

Second, AI-generated writing demonstrated lower lexical diversity. Although the vocabulary was grammatically correct and contextually appropriate, certain transition words and academic phrases appeared repeatedly, leading to greater predictability.

Third, AI-generated documents generally maintained exceptionally consistent paragraph organization. Each paragraph followed a structured progression from topic introduction to explanation and conclusion, whereas human-authored writing often displayed greater variability in rhetorical style and sentence organization.

Fourth, AI-generated texts occasionally contained semantically redundant statements that repeated similar ideas using different wording. Although such repetitions did not reduce grammatical quality, they created identifiable discourse patterns that transformer models successfully captured.

These observations indicate that transformer-based models do not rely solely on isolated lexical features but instead learn broader contextual and discourse-level representations capable of distinguishing between machine-generated and human-authored academic writing.

3.7 Discussion

The experimental findings demonstrate that transformer-based language models provide an effective solution for detecting AI-generated text in academic writing. The consistently high classification performance obtained across all evaluated models confirms that contextual language representations learned through self-attention mechanisms are capable of capturing subtle stylistic differences that traditional machine learning approaches often fail to identify.

Compared with conventional classifiers such as Support Vector Machines, Logistic Regression, or Random Forest, transformer-based architectures eliminate the need for extensive manual feature engineering. Instead, semantic, syntactic, and contextual information is learned directly from large textual corpora during pretraining and subsequently refined through task-specific fine-tuning. This characteristic enables transformer models to generalize effectively across diverse academic writing styles and disciplines.

Among the evaluated architectures, DeBERTa demonstrated the strongest overall performance, indicating that recent innovations in transformer design substantially improve contextual representation learning. RoBERTa also achieved excellent results, confirming that optimized pretraining strategies contribute significantly to downstream classification accuracy. Although DistilBERT sacrificed a small degree of predictive performance, its computational efficiency makes it highly suitable for large-scale deployment in educational institutions with limited hardware resources.

The results further suggest that AI-generated text detection should not rely exclusively on surface-level linguistic features. As generative language models continue to evolve, grammatical correctness and vocabulary richness increasingly resemble authentic human writing. Consequently, future detection systems must focus on deeper contextual, semantic, and discourse-level representations, areas where transformer-based architectures currently provide substantial advantages.

Despite these promising findings, several limitations remain. The experiments primarily involved English-language academic texts and binary classification between human-written and AI-generated documents. Future research should investigate multilingual datasets, mixed-authorship scenarios in which AI partially assists human writers, and continual learning approaches capable of adapting to newly released large language models. Additionally, incorporating explainable artificial intelligence (XAI) techniques may improve transparency by identifying specific textual features influencing classification decisions.

Overall, the results confirm that transformer-based models represent a robust and scalable approach for preserving academic integrity in the era of generative artificial intelligence. Their ability to accurately identify AI-generated academic writing provides valuable support for educators, journal publishers, and research institutions seeking to ensure originality, transparency, and ethical use of AI-assisted writing technologies.

4. CONCLUSIONS

This study demonstrates that transformer-based models are highly effective for detecting AI-generated text in academic writing, achieving consistently high classification performance across multiple evaluation metrics. Among the evaluated architectures, DeBERTa achieved the best overall results, followed by RoBERTa, BERT, and DistilBERT, indicating that recent advancements in transformer design significantly enhance the ability to distinguish AI-generated content from human-authored academic text. The experimental findings also show that contextual language representations learned through self-attention mechanisms successfully capture subtle semantic, syntactic, and discourse-level characteristics that differentiate machine-generated writing from authentic scholarly work. These results highlight the potential of transformer-based approaches as reliable tools for supporting academic integrity, assisting educators and publishers in identifying AI-generated content, and promoting the responsible use of generative artificial intelligence in higher education. Future research should extend this work by incorporating multilingual datasets, hybrid human–AI writing scenarios, explainable artificial intelligence techniques, and continual learning strategies to improve the robustness and adaptability of AI-generated text detection systems as generative language models continue to evolve.

REFERENCES

[1] Vera E. Woloshyn, Sam Illingworth, and Snežana Obradović-Ratković, "Introduction to Special Issue Expanding Landscapes of Academic Writing in Academia," *Brock Educ. J.*, vol. 33, no. 1, pp. 3–9, 2024, doi: 10.26522/brocked.v33i1.1118.

[2] J. Hutson, "Rethinking Plagiarism in the Era of Generative AI," *J. Intell. Commun.*, vol. 4, no. 1, 2024, doi: 10.54963/jic.v4i1.220.

[3] S. M. A. Shahid, M. N. Ali, M. H. Sarkar, and M. H. Rahman, "Ensuring Authenticity in Scientific Communication Approaches to Detect and Deter Plagiarism," *TAJ J. Teach. Assoc.*, vol. 37, no. 1, pp. i–iii, 2024, doi: 10.70818/taj.v37i1.0157.

[4] J. A. Oravec, "Artificial Intelligence Implications for Academic Cheating: Expanding the Dimensions of Responsible Human-AI Collaboration with ChatGPT and Bard," *J. Interact. Learn. Res.*, vol. 34, no. 2, pp. 213–237, 2023, doi: 10.70725/304731gmmvhw.

[5] K. C. Fraser, H. Dawkins, and S. Kiritchenko, "Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods," *J. Artif. Intell. Res.*, vol. 82, pp. 2233–2278, 2025, doi: 10.1613/jair.1.16665.

[6] V. B. Gupta, N. Chitranshi, V. Palanivel, S. Sheriff, D. Basavarajappa, and V. Gupta, "Critical Perspectives on Ethical Challenges in Higher Education: Analysing Contemporary Practices and Future Considerations," *J. Educ. Learn.*, vol. 14, no. 6, p. 119, 2025, doi: 10.5539/jel.v14n6p119.

[7] J. Q. J. Liu *et al.*, "The great detectives: humans versus AI detectors in catching large language model-generated medical writing," *Int. J. Educ. Integr.*, vol. 20, no. 1, p. 8, 2024, doi: 10.1007/s40979-024-00155-6.

[8] V. H. Kalmani, A. C. Adamuthe, A. P. Gondil, V. P. Patil, R. A. Kore, and V. M. Metkari, "AI vs. Human Writing: Developing a Novel Method for Text Authenticity Detection in Education," *Int. J. Mod. Educ. Comput. Sci.*, vol. 17, no. 3, pp. 45–58, 2025, doi: 10.5815/ijmecs.2025.03.04.

[9] M. S. Islam, L. Xiangdong, and J. Ahmed, "BERT: advancements in language understanding for different NLP tasks: challenges and future perspectives," *J. Electr. Syst. Inf. Technol.*, vol. 13, no. 1, p. 49, 2026, doi: 10.1186/s43067-026-00341-1.

[10] A. Şentürk, A. Albayrak, and S. Arpacı, "Multi-Class News Classification with BERT, DistilBERT, RoBERTa, and ELECTRA Natural Language Processing Models," *Düzce Üniversitesi Bilim ve Teknol. Derg.*, vol. 14, no. 1, pp. 117–129, 2026, doi: 10.29130/dubited.1737003.

[11] A. Onan, "Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 7, p. 101610, 2023, doi: 10.1016/j.jksuci.2023.101610.

[12] A. Jashari, "Enhancing Coherence and Persuasiveness in Academic Discourse through Lexical Cohesion and Critical Thinking," *Open J. Soc. Sci.*, vol. 14, no. 01, pp. 471–483, 2026, doi: 10.4236/jss.2026.141028.

[13] J. Al Hosni, "Preserving Authorial Voice in Academic Texts in the Age of Generative AI: A Thematic Literature Review," *Arab World English J.*, vol. 16, no. 3, pp. 244–258, 2025, doi: 10.24093/awej/vol16no3.14.

[14] X. Liang, X. Zhou, H. Zou, Y. Lu, and J. Qu, "DeepDiveAI: Identifying AI-Related Documents in Large Scale

Literature Dataset," *J. Soc. Comput.*, vol. 6, no. 2, pp. 158–169, 2025, doi: 10.23919/JSC.2025.0007.

[15] M. M. Danyal, S. S. Khan, M. Khan, M. B. Ghaffar, B. Khan, and M. Arshad, "Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques," *J. Big Data*, vol. 5, pp. 1–18, 2023, doi: 10.32604/jbd.2023.041319.

[16] E. Ferrara, "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models," *J. Comput. Soc. Sci.*, vol. 7, no. 1, pp. 549–569, 2024, doi: 10.1007/s42001-024-00250-1.