Detection of AI-Generated Text Using Linguistic Features and Machine Learning for Preserving Academic Integrity

Saleem Raja Abdul Samad[1], Basant Kumar[2]

[1] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia.

[2]Modern College of Business and Science, Muscat, Sultanate of Oman

Email ID : asaleemrajasec@gmail.com

Abstract

Advancement of Large Language Models (LLMs) has reshaped how information is produced and consumed, enabling highly coherent, human-like text across diverse applications such as academic research, tutoring, summarization, and code generation. While these advancements enhance productivity and learning efficiency, they also raise pressing ethical concerns, including misinformation, plagiarism, and academic integrity violations. Current detection methods often rely on model-specific parameters or hidden states, which limits their adaptability to new and evolving architectures. Similarly, linguistic and embedding-based approaches, though effective, are computationally intensive, reducing scalability and hindering real-time deployment. Proprietary detection tools further complicate accessibility, as they are costly and frequently lack accuracy against rapidly advancing models. To address these challenges, a lightweight hybrid framework integrating linguistic and machine learning techniques is proposed. By combining stylometric, syntactic, and pragmatic features, this approach aims to improve generalization across domains while maintaining computational efficiency. Evaluated on multi-domain datasets with adversarial testing, the framework demonstrates resilience against paraphrasing and evolving LLMs. Such solutions are crucial for ensuring trustworthy AI integration, balancing the benefits of LLMs with the need for ethical safeguards and scalable detection mechanisms.

Keywords: Large Language Models, AI-generated text detection, misinformation, academic integrity, hybrid framework, stylometric features, adversarial robustness, scalability, generalization, ethical concerns.

Introduction

The rapid evolution of artificial intelligence particularly large language models (LLMs) has reshaped how information is generated, accessed, and communicated. These systems now possess the capability to generate remarkably coherent and contextually accurate text, mirroring human-authored content with impressive consistency. Applications of AI-generated text are expanding rapidly, including aiding in academic research, streamlining information gathering and summarization, personalizing tutoring experiences, enabling problem-solving, and even assisting with code generation. These advancements significantly enhance the teaching and learning process, boosting skills in areas like critical thinking and writing. AI also plays a vital role in research. It assists with gathering relevant information from vast datasets, conducting data analysis, formulating hypotheses, generating code, and drafting proposals. Industries like medicine, law, and the creative industries are also seeing significant influence from AI-generated text. This transformative shift offers significant advantages, enhancing the teaching and learning process by enabling critical thinking and writing skills. However, the increasing prevalence of AI-generated text also raises crucial concerns about misinformation, the authenticity of academic work, and the potential for plagiarism due to lack of proper source attribution. These issues demand careful consideration of ethical implications settings and broader public use.

Existing research work follows one of two approaches to classify AI-generated text. The first approach is based on model-dependent methods, which rely on access to a specific model's internal signals such as its parameters, gradients, or hidden states to detect whether a piece of text was generated by that model. The other approach employs linguistic features or embedding models in combination with machine learning techniques to distinguish AI-generated text from human-written content. However, both of these approaches are computationally complex. Model-specific techniques lack generalizability, and extracting linguistic features is a challenging task that requires deep expertise in linguistics and AI to identify the most suitable and broadly applicable features. Moreover, Licensed AI-detection software solutions offer partial support and are costly, proprietary, and lack full reliability, especially when confronted with rapidly evolving generative models. Therefore, it is crucial to develop computationally feasible methods that can generalize across different large language model generated texts to effectively distinguish between AI-generated and human-written content. Such advancements will strengthen trust and promote confident integration of AI technologies in both academic and business domains.

Literature Review:

Rujeedawa et al. (2025) conducted one of the largest-scale investigations into the detection of AI-generated texts, analyzing a dataset of 483,000 essays. The study focused on linguistic and stylistic features such as text length, punctuation frequency, vocabulary richness, readability indices (including the Gunning Fog Index and Flesch Reading Ease), and sentiment polarity. Using these features, the authors trained multiple classifiers and found that the Random Forest model achieved the highest performance, with an accuracy of 82.6% on a custom test set. However, the study revealed a critical limitation: detectors trained on outputs from ChatGPT-3.5 struggled significantly when applied to more advanced models, underscoring the fragility of feature-based approaches in keeping pace with evolving language models.

Ma et al. (2023) approached the problem from a scientific writing perspective, examining differences between AI- and human-authored scientific abstracts. Their framework integrated syntax, semantics, and pragmatics to capture deeper linguistic distinctions. The study found that AI-generated texts often lacked novel insights and exhibited inconsistencies in knowledge representation, which made them distinguishable from human writing. Among the models tested, RoBERTa achieved an F1-score of 88.3%, outperforming logistic regression and even surpassing human evaluators in detection accuracy. Importantly, the authors noted that as syntactic differences between AI and human texts diminish, semantic and pragmatic cues become increasingly vital. Despite these strengths, the study highlighted vulnerabilities: linguistic detectors often fail when applied across different large language models, domains, or languages, and they can be easily deceived by paraphrasing or minor edits.

Masih et al. (2025) expanded the comparative landscape by evaluating machine learning, deep learning (LSTM/GRU), and transformer-based models on a balanced dataset of 20,000 samples, consisting of human-written texts and outputs from ChatGPT-3.5 and ChatGPT-4. Their results confirmed the superiority of transformer models, with RoBERTa achieving the highest accuracy at 96.1%. Beyond raw performance, the study emphasized robustness through calibration, pruning, and explainability techniques such as LIME and SHAP, which provided interpretable insights into the linguistic distinctions between human and AI texts. Nevertheless, the authors acknowledged two key

limitations: the balanced dataset may not reflect real-world distributions, and the scope was restricted to ChatGPT-3.5 and 4, leaving questions about generalizability to newer or alternative models.

Mahmoud Ragab et al. (2025) introduced a novel hybrid detection model combining Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU), optimized using the Spotted Hyena Algorithm (SHO). This approach, termed CHWAIG-DLSHO, was designed to classify human-written versus AI-generated sentences with high precision. The methodology incorporated preprocessing steps, Latent Dirichlet Allocation (LDA)-based embeddings, and extensive hyperparameter tuning. The model achieved an impressive accuracy of 99.17%, outperforming existing detection techniques. Despite its success, the authors acknowledged that the hybrid CNN-GRU model, coupled with SHO optimization, may demand significant computational resources, which could hinder its deployment in real-time applications.

Nuruzzaman et al. (2024) conducted one of the largest-scale detection studies, leveraging a dataset of 1.6 million samples to classify human versus AI-generated text. Their experiments compared deep learning models (Bi-GRU and LSTM) with traditional machine learning approaches (Decision Tree and Gradient Boosting). Results showed that Bi-GRU achieved 99.69% accuracy and LSTM achieved 99.68%, both with perfect F1-scores of 100%, significantly outperforming traditional classifiers. The study revealed distinct linguistic patterns between human and AI texts and demonstrated practical applicability by deploying a detection system via a Streamlit app. However, the reliance on deep learning models introduces high computational costs for training and deployment, and the use of a private dataset raises concerns about generalizability across different genres, domains, and languages.

Nuzhat Noor Islam Prova (2024) proposed a detection framework that integrates both machine learning and deep learning approaches to distinguish AI-generated text from human writing. The study evaluated models such as XGBoost (XGB), Support Vector Machines (SVM), and BERT, with BERT achieving the highest accuracy of 0.93, outperforming traditional machine learning techniques. The findings underscored the effectiveness of contextual and linguistic analysis in AI text detection, highlighting the strength of transformer-based models in capturing nuanced differences. Nonetheless, the study faced limitations due to its relatively small dataset of 3,000 samples, which restricts the robustness of conclusions. Additionally, BERT's high computational requirements for training and inference may limit its scalability and practical deployment in resource-constrained environments.

Maktabdar et al. (2025) investigated the detection and classification of ChatGPT-generated content using deep transformer models. The study evaluated a range of machine learning and deep learning approaches, ultimately finding that a RoBERTa-based model achieved near-perfect performance, with an F1-score of 0.992 and accuracy of 0.991. This work not only established a strong baseline for AI text detection but also contributed a publicly available dataset to support future research. Despite its impressive results, the study's scope was limited to ChatGPT outputs, raising concerns about generalizability to other AI text generators. Furthermore, the authors acknowledged that transformer-based models require substantial computational resources, and their effectiveness can be undermined by paraphrasing or text manipulation.

Yadagiri et al. (2024) proposed a classification system for distinguishing human-written from AI-generated text using linguistic and structural features derived from the HC3-English dataset. Their experiments demonstrated that transformer-based models, particularly RoBERTa, achieved exceptionally high accuracy, with results reaching 99.73%, underscoring the effectiveness of deep

learning in preserving content integrity. The study highlighted the role of linguistic markers such as part-of-speech tags and readability measures in detection. However, the authors cautioned that such extremely high accuracy may indicate potential overfitting to the dataset, limiting the model's robustness in real-world scenarios. Additionally, reliance on specific linguistic and structural features may fail to capture the evolving stylistic and contextual strategies of newer AI text generators.

In the field of AI-generated text detection, several critical research gaps remain that limit the effectiveness, scalability, and adaptability of current approaches. Addressing these gaps is essential for developing robust detection systems that can operate across diverse contexts and withstand adversarial challenges.

One major limitation is dataset size. Most existing studies rely on relatively small datasets, which restricts the robustness and generalizability of detection models. Without large-scale, diverse corpora, models risk overfitting and may fail to perform well across different domains, genres, or languages. Expanding datasets to include varied text types and multilingual sources would significantly strengthen detection reliability.

Another gap lies in model-agnostic detection. Much of the research has focused narrowly on detecting ChatGPT-generated text, leaving outputs from other large language models (LLMs) underexplored. This creates a blind spot, as detection systems may not generalize across different architectures or training paradigms. Future work should prioritize methods that are adaptable to multiple LLMs, ensuring broader applicability.

Detection methods also face challenges with text length sensitivity. Short texts often lack distinctive linguistic cues, making it difficult for models to differentiate between human and AI-generated content. This is particularly problematic in social media, messaging apps, or microblogging platforms where brevity is common. Research into feature extraction techniques that remain effective in sparse contexts is needed.

A further vulnerability is adversarial manipulation. Many detection approaches are susceptible to paraphrasing, synonym substitution, or other adversarial strategies that mask AI-generated origins. This highlights the need for more resilient detection frameworks that can withstand intentional obfuscation and maintain accuracy under manipulation.

Additionally, most studies adopt a binary classification framework (human vs. AI), neglecting scenarios where texts are partially AI-generated or collaboratively written by humans and AI. Real-world applications often involve mixed authorship, and detection systems must evolve to handle these nuanced cases.

Finally, computational complexity presents a practical barrier. Deep learning models combined with optimization algorithms can be resource-intensive, limiting their feasibility for real-time or large-scale deployment. Research into lightweight, efficient detection methods is crucial to enable scalable applications without sacrificing accuracy.

Problem Statement

- AI-generated text from advanced LLMs is increasingly human-like, making reliable detection challenging.

- Existing detection methods are often model-specific, computationally intensive, and fail to generalize across diverse AI models.

Research Questions

- How can AI-generated text be reliably distinguished from human-written content across diverse LLMs?
- Which linguistic related features (Stylometric, Pragmatic, Syntactic)and machine learning methods are most computationally efficient and generalizable for detection?
- How can detection approaches remain accurate and robust against rapidly evolving AI models?

Proposed Method

- The proposed method uses lightweight hybrid linguistic and machine learning framework to detect AI-generated text, integrating stylometric, syntactic, and pragmatic features. It is evaluated on multi-domain datasets with adversarial testing to ensure robustness, generalization, and resilience against evolving large language models.

References

- M. I. H. Rujeedawa, S. Pudaruth, and V. Malele, "Unmasking AI-generated texts using linguistic and stylistic features," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 3, pp. 213-220, Mar. 2025, doi: 10.14569/IJACSA.2025.0160321

- L. Ma, Z. Wang, K. Wang, and J. Sun, "AI vs. Human -- Differentiation Analysis of Scientific Content Generation," arXiv preprint arXiv:2301.10416, 2023.

- Masih, A., Afzal, B., Firdoos, S. *et al.* Classifying human vs. AI text with machine learning and explainable transformer models. *Sci Rep* **15**, 43310 (2025). https://doi.org/10.1038/s41598-025-27377-z

- Mahmoud Ragab, Ehab Bahaudien Ashary, Faris Kateb, Abeer Hakeem, Rayan Mosli, Nasser N. Albogami, Sameer Nooh, Classification of human-written and AI-generated sentences using a hybrid CNN-GRU model optimized by the spotted hyena algorithm, Alexandria Engineering Journal, vol.126, 2025, pp.116-130. https://doi.org/10.1016/j.aej.2025.04.071.

- M. Nuruzzaman, A. Nadia, M. M. Billal and K. M. S. Nomani, "Human and AI Written Text Detection Using Deep Learning and Machine Learning," *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2024, pp. 539-544, doi: 10.1109/ICCIT64611.2024.11022524.

- Nuzhat Noor Islam Prova , *"Detecting AI Generated Text Based on NLP and Machine Learning Approaches,"* arXiv preprint arXiv:2404.10032, Apr. 2024. Available: https://arxiv.org/pdf/2404.10032.pdf

- Maktabdar Oghaz M, Babu Saheer L, Dhame K and Singaram G (2025) Detection and classification of ChatGPT-generated content using deep transformer models. Front. Artif. Intell. 8:1458707. doi: 10.3389/frai.2025.1458707

- 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 *Proc. 21st Int. Conf. Natural Language Processing (ICON)*, Chennai, India, Dec. 2024, pp. 188–196. Available: https://aclanthology.org/2024.icon-1.21.pdf