Human or Machine? Analyzing Text Authorship Using AI Models

Abstract

With the rapid evolution of artificial intelligence, language models can now generate text that is fluent, coherent, and grammatically sound. However, this advancement has created a growing concern, as AI-generated content is often used without verification, even in critical or confidential documents, where accuracy and reliability are essential. This paper proposes a machine learning–based system to detect AI-generated text and differentiate it from human-written content. The model evaluates various linguistic and statistical features, including lexical diversity, Flesch reading ease, Gunning fog index, grammar error rate, passive voice ratio, predictability score, and burstiness. It provides a percentage-based likelihood indicating AI authorship. Experimental findings reveal that although AI-generated text closely resembles human writing, measurable stylistic and structural gaps persist, enabling reliable detection and reducing risks associated with AI misuse in sensitive contexts.

1. Introduction

The rapid evolution of Generative Pre-trained Transformers (GPTs), such as GPT-3 and subsequent models, has transformed the landscape of natural language generation. These large-scale language models are capable of producing highly coherent, contextually relevant, and stylistically consistent text that closely resembles human writing. While their applications span a wide range of domains—including education, journalism, business communication, and creative content generation—they have simultaneously introduced significant ethical and academic challenges. The boundary between authentic human authorship and algorithmically generated text has become increasingly blurred, giving rise to issues of plagiarism, authorship authenticity, misinformation, and intellectual property ambiguity.

Traditional plagiarism detection tools, which primarily depend on string matching, database comparison, and lexical similarity, have proven inadequate in this emerging context. Unlike conventional plagiarized content that can be traced back to original human sources, AI-generated text is entirely synthetic, exhibiting no direct overlap with existing material. This inability to detect non-replicated yet artificially generated content has created vulnerabilities within academic and professional ecosystems, where the credibility and originality of written work are paramount.

As AI systems become more sophisticated, there is an urgent need for advanced detection frameworks capable of identifying the subtle linguistic, semantic, and structural differences between human and machine-generated text. To address this growing challenge, the proposed research focuses on developing an intelligent text detection system that leverages Natural Language Processing (NLP), machine learning, and deep transformer architectures. The system is designed to extract linguistic, semantic, and contextual features to assess writing style, coherence patterns, and vocabulary richness. It will utilize transformer-based embeddings such as BERT, RoBERTa, and LLaMA to capture deep contextual relationships and integrate them with machine learning classifiers for binary classification between AI- and human-generated content.

This approach bridges the existing gap between ethical awareness and technical implementation by providing a robust, data-driven, and adaptive solution. Through comprehensive analysis of modern AI language outputs, the model aims to contribute toward maintaining the integrity of digital communication, ensuring transparency in authorship, and reinforcing trust in academic and professional writing standards.

2. Related Work

Dehouche [1] explored the ethical implications of GPT-3's ability to generate text indistinguishable from human writing, focusing on issues of plagiarism and authorship authenticity. His work effectively highlights the growing academic concern about AI-assisted writing and the potential erosion of intellectual integrity. However, the study remains theoretical and does not propose any technical framework for identifying AI-generated content. While the discussion raises awareness of the moral risks, it lacks an actionable approach to mitigate them. In contrast, my work builds upon this concern by developing a **computational system** that directly addresses the detection problem through linguistic analysis and model-based classification, thus moving from ethical discourse to practical implementation.

Similarly, the study *Detection of AI-Generated Text Using Machine Learning Techniques* [2] presents early experimental models using **Logistic Regression**, **Support Vector Machines (SVMs)**, and basic neural networks to classify text based on statistical and linguistic cues such as word frequency, sentence complexity, and vocabulary diversity. The authors successfully demonstrate that such features can partially distinguish AI from human writing, establishing a strong baseline for AI text detection. However, the research is limited by its reliance on outdated datasets and the absence of modern contextual models, leading to lower adaptability against newer transformer-based text generators. To overcome these limitations, my approach incorporates **updated datasets**, **semantic-rich embeddings from transformer models**, and **hybrid learning techniques**, enabling the system to detect even the latest and context-aware AI-generated text with higher precision and robustness.

In a recent study, [3] the authors investigated the detection of ChatGPT-generated text within short restaurant reviews by applying a transformer-based classifier and comparing it with a perplexity-driven approach. Their method effectively demonstrates that machine learning models can outperform traditional perplexity measures in distinguishing AI-produced content, particularly highlighting linguistic traits such as formality, politeness, and lack of emotional depth. While the work offers valuable insights into short-text detection and contributes explainability through SHAP analysis, its scope is limited to a single domain, short input length, and a narrow focus on ChatGPT-generated samples. In contrast, the proposed system expands beyond domain-specific constraints by integrating linguistic, semantic, and transformer-based embeddings to detect AI-generated text across diverse writing styles and modern LLMs. This broader, hybrid methodology addresses the limitations of domain restriction and model specificity, resulting in a more generalizable and robust detection framework.

Researchers such as Ma et al. [4] investigated the growing threat of AI-generated scientific text, emphasizing issues like misinformation, hallucinated references, and the difficulty of distinguishing AI-

written abstracts from authentic scholarly work. Their study is strong in its multi-level analysis—covering syntax, semantics, and pragmatics—and in exploring both feature-based and fine-tuned transformer detection methods. However, the work is narrowly focused on scientific abstracts and relies heavily on domain-specific datasets, which limits its generalizability to broader text types. In contrast, the proposed system incorporates a wider set of linguistic, semantic, and readability-based features and leverages modern transformer embeddings to detect AI-generated content across diverse domains, thereby addressing the scalability and adaptability gaps present in their approach.

The authors [5] present a supervised approach for detecting ChatGPT-generated essays by creating a topic-aligned dataset and training XGBoost classifiers using TF-IDF and handcrafted linguistic features. Their work is notable for offering one of the earliest structured datasets for ChatGPT-vs-human essays and for demonstrating that even traditional machine-learning models can achieve strong accuracy in distinguishing AI-generated text. However, the system is limited by its reliance on shallow lexical features, model-specific patterns, and domain-restricted data, making it less effective against modern LLMs and varied writing contexts. In contrast, the proposed framework integrates transformer-based semantic embeddings with richer linguistic and contextual features, enabling detection beyond surface-level patterns and improving robustness across domains and generative models. This broader, feature-enhanced design directly addresses the scalability and generalization weaknesses present in their methodology.The authors [6] introduce BART, a denoising autoencoder that combines bidirectional and autoregressive transformers, capable of handling a wide range of NLP tasks through flexible noising and sequence-to-sequence reconstruction. Its strength lies in generalizing prior models like BERT and GPT, achieving state-of-the-art performance in text generation, comprehension, summarization, and dialogue tasks. However, while BART excels in generating and understanding text, it is not explicitly designed for detecting AI-generated content or distinguishing it from human-written text. In contrast, the proposed framework leverages transformer-based embeddings along with detailed linguistic and semantic feature analysis to classify AI- versus human-generated content, thereby addressing the detection challenge that BART and similar generation-focused models do not target.

The authors present [7] SciBERT, a BERT-based language model pretrained on a large corpus of scientific publications, aiming to improve performance on NLP tasks within scientific domains. By using an in-domain vocabulary and extensive pretraining on scientific text, SciBERT achieves state-of-the-art results in tasks requiring domain-specific understanding. Its strength lies in providing rich contextual embeddings tailored for scientific content, which enhances downstream applications such as knowledge extraction and document classification. However, SciBERT is primarily focused on representation learning and task performance; it does not address the detection of AI-generated text or distinguish human-written from machine-generated scientific content. In contrast, the proposed framework explicitly targets this detection problem by combining linguistic, semantic, and transformer-based features to classify AI- versus human-authored texts, thereby filling the gap left by general purpose scientific language models like SciBERT.

Recent studies [8] highlight the emergence of automatically generated scientific content, including fully or partially AI-authored

papers, raising both technical and ethical concerns for publishers (Cabanac et al., 2021; Noorden, 2021; Labbé & Labbé, 2012). Experiments such as SCIgen demonstrated that nonsensical, machine-

generated papers could pass peer review, while “paper mills” and paraphrasing tools like SpinBot further contribute to fraudulent scientific manuscripts. The strength of this line of research lies in identifying the prevalence and impact of AI-generated scientific content, and in organizing competitions to develop detection methods. However, existing efforts are largely limited to specific datasets or rule-based detectors and often fail to generalize to new models or evolving AI-generated content. In contrast, the proposed framework leverages linguistic, semantic, and transformer-based features to create a more robust, adaptive detection system capable of identifying AI-generated text across multiple domains and scientific writing styles, thereby overcoming the limitations of prior methods.

Recent work [9] on detecting machine-generated scientific text emphasizes the challenges posed by realistic AI-generated submissions, which can be highly fluent and contextually coherent (Cabanac & Labbé, 2021). Frameworks such as SynSciPass have been developed to label text by the type of generative technology used, enabling more nuanced detection beyond simple binary classification and improving robustness under domain shifts. The strength of this research lies in its systematic dataset creation and attention to model attribution, which helps mitigate misflagging of legitimate manuscripts. However, even with advanced datasets, existing models still struggle to generalize across diverse scientific domains and evolving AI models. In comparison, the proposed approach integrates transformer-based embeddings, semantic and linguistic feature analysis, and hybrid classification techniques to achieve more accurate, adaptable, and domain-agnostic detection of AI-generated scientific content, effectively addressing the limitations of prior work.

Recent work [10] by Hayawi et al. investigates detection across multiple genres (essays, code, stories) using a dataset of human- vs LLM-generated text and machine-learning classifiers. Their strength lies in covering a diverse range of text types, demonstrating genre-specific detection capabilities. However, their dataset remains relatively small and lacks deep linguistic-semantic analysis. In comparison, the proposed framework enriches this by leveraging advanced transformer embeddings and detailed linguistic features (e.g., burstiness and predictability) to improve robustness across multiple genres and larger datasets.

Another line of research [11] by Adilazuarda provides a comparative survey of detection approaches, including shallow models, fine-tuned language models, and multilingual detectors, highlighting relative performance trade-offs. The strength of this work is its broad methodological coverage, revealing which strategies succeed under different conditions. Its limitation is the absence of feature-level interpretability and deeper linguistic analysis. By contrast, the proposed system combines interpretability via linguistic/statistical features with powerful transformer embeddings, enabling both high accuracy and explainability.

Weber-Wulff et al. [12] tested a suite of 12 off-the-shelf commercial and open-source AI text detectors in realistic use-case scenarios, revealing significant false negatives and sensitivity to paraphrasing. Their strength lies in providing a real-world benchmark for existing detection tools. Its drawback is that they do not propose new detection algorithms or linguistic insights. The proposed system complements this by offering a principled, feature-based detection model tailored to avoid the blind spots of existing detectors.

Elkhatat et al. [13] evaluated several commercial AI-text detection tools, including GPTZero and Copyleaks, and observed substantial inconsistency in detection, especially for newer or adapted LLMs.

The strength of their work is in empirically assessing real-world detection tools across valid use-cases. But since they do not develop a custom detection model, coverage and adaptability remain limited. The proposed framework addresses this by building a dedicated detection system based on transformer embeddings and linguistic features, ensuring adaptability beyond today's existing commercial tools.

3. Methodology :

A. Overview of Existing Detection Approaches

Several approaches have been proposed to detect AI-generated text, each with distinct strengths and limitations. Rule-based or heuristic methods, such as perplexity scores, word-frequency rules, and stylistic heuristics, provide simplicity and computational efficiency; however, they fail to capture the contextual and semantic nuances of modern AI-generated text. Traditional machine learning, or feature-based methods, leverage classifiers like Logistic Regression, Support Vector Machines (SVMs), Random Forests, and XGBoost using manually engineered features including word frequency, sentence complexity, part-of-speech distributions, and vocabulary richness. While effective to some extent, these methods often suffer when applied to domain-specific datasets or transformer-generated content. More recent approaches rely on deep learning and transformer-based models, such as BERT, SciBERT, and GPT embeddings, which capture semantic meaning, contextual flow, and coherence. Despite their advanced capabilities, these models require large datasets, high computational resources, and often underperform when used in isolation without additional linguistic or statistical features. Collectively, these studies demonstrate that no single method provides both robust detection and interpretability across multiple domains.

B. Proposed Machine Learning Feature-Based Methodology

Motivated by the limitations of existing approaches, the proposed system adopts a hybrid machine learning methodology, emphasizing feature-based analysis. The framework integrates linguistic features (e.g., sentence length, lexical richness, readability scores), statistical features (e.g., word frequency, token variance, repetition patterns), syntactic features (e.g., part-of-speech tags, dependency patterns), and semantic embeddings derived from transformer models. These features are combined into a unified representation and fed into supervised classifiers, such as XGBoost and SVMs, to perform binary classification between AI-generated and human-authored text. This approach leverages the strengths of both engineered features and deep contextual embeddings, enabling accurate, domain-agnostic detection while maintaining computational efficiency.

C. Justification for Feature-Based Machine Learning Approach

The feature-based machine learning methodology was selected for several compelling reasons. First, it provides high generalizability across domains, addressing the shortcomings of transformer-only models that often fail on new or specialized text types, including academic essays, reports, and scientific

articles. Second, it enhances interpretability, allowing evaluators to understand classification decisions based on quantifiable features, such as sentence complexity, vocabulary usage, and syntactic patterns. Third, it offers robustness against evolving AI models, since linguistic and statistical cues tend to remain stable despite advances in generative models. Fourth, the literature demonstrates that hybrid models combining classifiers like XGBoost or SVMs with enriched feature sets consistently achieve strong detection performance. Finally, this approach balances detection accuracy with computational efficiency, while supporting multi-dimensional evaluation of authorship, grammar, and overall writing quality.

4. Data Collection and Preparation :

For this study, the publicly available dataset “AI vs Human Content Detection 1000+ Records (2025)” by Pratyush Puri on Kaggle was utilized [Source: Kaggle, publicly available]. The dataset comprises 1,367 text samples drawn from multiple domains, including academic papers, essays, creative writing, news articles, blog posts, and general articles. Each entry is labeled to indicate whether the text is AI-generated or human-written, providing supervised learning capability for classification tasks.

The dataset includes a rich set of features suitable for feature-based machine learning, covering linguistic, syntactic, semantic, and statistical properties. Key features include word count, character count, sentence count, lexical diversity, average sentence length, average word length, punctuation ratio, Flesch reading ease, Gunning fog index, grammar error count, passive voice ratio, predictability score, burstiness, and sentiment score. These features enable detailed quantitative and qualitative analysis of text, capturing patterns critical for AI-generated content detection.

Prior to model training, preprocessing steps were conducted to ensure data quality and consistency. Missing values were addressed in features such as Flesch reading ease, Gunning fog index, passive voice ratio, and sentiment score. Textual data were standardized to facilitate feature extraction, and the dataset was partitioned into training, validation, and testing subsets, ensuring balanced representation of AI-generated and human-written samples across each split.

This dataset was chosen due to its comprehensive coverage of multi-domain textual data, inclusion of diverse and informative features, and suitability for the proposed hybrid feature-based machine learning methodology. Its granularity supports robust feature extraction, effective classification, and generalizability across different types of written content.

Feature Name	Description	Type
Text	The raw text content of each sample	String
AI Label	Indicates whether the text is AI-generated (1) or human-written (0)	Categorical / Binary
Word Count	Total number of words in the text	Numeric

Character Count	Total number of characters in the text	Numeric
Sentence Count	Total number of sentences	Numeric
Lexical Diversity	Ratio of unique words to total words	Numeric
Average Sentence Length	Mean number of words per sentence	Numeric
Average Word Length	Mean number of characters per word	Numeric
Punctuation Ratio	Ratio of punctuation marks to total characters	Numeric
Flesch Reading Ease	Readability score based on sentence/word length	Numeric
Gunning Fog Index	Readability score estimating years of formal education required	Numeric
Grammar Error Count	Number of grammar mistakes detected	Numeric
Passive Voice Ratio	Proportion of sentences in passive voice	Numeric
Predictability Score	Measure of text predictability using language model	Numeric
Burstiness	Degree of variation in sentence length and structure	Numeric
Sentiment Score	Overall sentiment polarity of the text	Numeric

5. Data Preprocessing and Feature Extraction :

Prior to model training, the dataset underwent a series of preprocessing steps to ensure quality, consistency, and suitability for machine learning. Missing values in numerical features such as Flesch Reading Ease, Gunning Fog Index, Passive Voice Ratio, and Sentiment Score were addressed through feature-wise mean imputation. Text data were standardized by converting all characters to lowercase and removing extraneous whitespace, while punctuation and special characters were preserved where necessary to compute features such as Punctuation Ratio and Burstiness. The categorical AI Label was encoded into a binary format, with 1 representing AI-generated text and 0 representing human-written content.

The dataset was partitioned into training, validation, and testing subsets, ensuring balanced representation of AI-generated and human-written samples across each split to facilitate effective supervised learning.

Feature extraction focused on capturing the linguistic, syntactic, semantic, and statistical properties of each text sample. Quantitative features such as Word Count, Character Count, Sentence Count, Lexical Diversity, Average Sentence Length, Average Word Length, and Punctuation Ratio were directly calculated from the text. Readability and complexity metrics, including Flesch Reading Ease and Gunning Fog Index, were derived to evaluate sentence structure and vocabulary usage. Grammar-focused features such as Grammar Error Count and Passive Voice Ratio were computed using rule-based natural language processing methods. Predictability and stylistic patterns were captured through Predictability Score and Burstiness, while Sentiment Score quantified the overall polarity of the text.

These features were then used as input to machine learning algorithms such as Logestic Regression, XGBoost and Support Vector Machines (SVMs), forming a feature-based classification pipeline for AI content detection. This approach leverages the quantitative, linguistic, and syntactic characteristics of text to achieve robust classification without relying on transformer-based embeddings.

6. Methodology: Machine Learning-Based Classification :

The proposed system adopts a feature-based machine learning approach to classify text as AI-generated or human-written. The extracted features from the dataset, encompassing linguistic, syntactic, semantic, and statistical properties, serve as the input for supervised learning algorithms. The primary models considered in this study are XGBoost and Support Vector Machines (SVMs), selected for their proven effectiveness in structured, feature-rich datasets and their balance of accuracy and computational efficiency.

A. Model Training : The dataset was divided into training (70%), validation (15%), and testing (15%) subsets to enable reliable evaluation of model performance and to prevent overfitting. Each model was trained on the combined feature set, using hyperparameter tuning via grid search to optimize parameters such as the number of estimators, learning rate, and maximum depth for XGBoost, and the kernel type, regularization parameter (C), and gamma for SVM. Features were scaled where necessary to ensure compatibility with SVM's distance-based optimization.

B. Evaluation Metrics : Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). These metrics provide a comprehensive assessment of the models' ability to distinguish AI-generated content from human-written text, accounting for both overall correctness and class-specific performance.

C. Rationale for Chosen Methodology : The feature-based ML approach was selected due to its robustness, interpretability, and adaptability across domains. Unlike transformer-based models, it requires moderate computational resources while allowing clear analysis of which features contribute to classification decisions. By leveraging linguistic, statistical, and syntactic cues, the approach captures stable patterns in text that remain effective even as AI-generated content evolves. Furthermore, hybrid use of XGBoost and SVM ensures both high accuracy and generalizability, while allowing for flexible integration with future feature enhancements.

7. References

1. Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, *21*, 17-23.
2. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.
3. Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.
4. Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). AI vs. Human--differentiation analysis of scientific content generation. *arXiv preprint arXiv:2301.10416*.
5. Shijaku, R., & Canhasi, E. (2023). Chatgpt generated text detection. *Publisher: Unpublished*.
6. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020, July). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871-7880).
7. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
8. Kashnitsky, Y., Herrmannova, D., De Waard, A., Tsatsaronis, G., Fennell, C. C., & Labbé, C. (2022, October). Overview of the DAGPap22 shared task on detecting automatically generated scientific papers. In *Proceedings of the Third Workshop on Scholarly Document Processing* (pp. 210-213).
9. Rosati, D. (2022). SynSciPass: detecting appropriate uses of scientific text generation. *arXiv preprint arXiv:2209.03742*.
10. Hayawi, K., Shahriar, S., & Mathew, S. S. (2024). The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. *Journal of Information Science*, 01655515241227531.
11. Adilazuarda, M. (2024, June). Beyond turing: A comparative analysis of approaches for detecting machine-generated text. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)* (pp. 1-12).
12. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1-39.
13. Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1-16.