Analyzing Structural and Stylistic Fingerprints in AI Text Classification

Spencer Davis
Linguistics 581, Section 001, Fall 2025
Submitted to Professor Dougal
Department of Linguistics
Brigham Young University
dspencem@byu.edu

Abstract

This study examines whether structural and stylistic features can reliably distinguish AI-generated text from human writing without relying on semantic content. Using three publicly available datasets, we extracted sixteen measurable features related to sentence structure, readability, vocabulary diversity, part-of-speech distributions, punctuation patterns, and sentiment tone. These features were used to train Logistic Regression models that identify the traits most predictive of authorship. The models achieved high accuracy on document-length datasets, reaching approximately 88 percent, while performance declined on single-sentence data due to limited structural information. Analysis of feature weights showed that AI-generated text typically exhibits greater mechanical complexity but lower vocabulary diversity, while human writing reflects longer and more variable sentence structures. These findings provide evidence that AI systems produce consistent structural patterns that remain identifiable across different sources and topics. This abstract was generated with the assistance of ChatGPT (OpenAI, 2025).

1 Introduction

Large Language Models are constantly improving the ability to mimick human vocabulary and semantics, which complicates detection for humans and classifiers. In fields like academia, verifying originality is critical. Current detection methods often focus on *what* the text says rather than *how* the model writes it. This can make classification models brittle and prone to fail across different domains. For example, a classifier trained on academic essays performs poorly on creative writing because the vocabulary changes entirely. This study shifts the focus from content to form. We investigated two research questions: *(1) Which stylistic features are most indicative of AI authorship?* and *(2) Can structural and stylometric features detect AI text independently of semantic content?* We hypothesize that while an AI can easily change its topic, it cannot easily mask its underlying generation patterns. By isolating these structural fingerprints, we aim to create a detection method that is effective across domains.

2 Literature Review

Research shows that humans cannot reliably tell the difference between AI-written and human-written text. For example, (Soni and Wade, 2023) found that human reviewers could distinguish between real summaries and those generated by ChatGPT about 50% of the time, the same as random chance. However, computer algorithms can successfully tell them apart by looking for specific patterns. AI models leave behind a fingerprint that machines can distinguish, even if people cannot. This project explores that idea by analyzing specific stylistic and structural features to detect AI text.

2.1 Stylistic Features

Several studies have identified specific writing styles common in AI text. (Mitrović et al., 2023) used explainable AI tools to show that ChatGPT tends to be ”polite, without specific details, using fancy and atypical vocabulary, impersonal, and typically it does not express feelings.” Our model accounts for this by calculating sentiment polarity and the density of pronouns to see if the text feels personal rather than robotic. (Georgiou, 2025) analyzed the specific types of words used by AI versus humans. They found that AI shows “greater usage of nouns and coordinating conjunctions,” while humans use more pronouns and auxiliaries. To account for this, we included features in our code to count the density of different parts of speech, including nouns, verbs, and adjectives.

2.2 Structural and Complexity Features

Beyond word choice, sentence structure is a strong indicator of AI authorship. (Jaashan and Bin-Hady, 2025) compared different AI models and found that they all have unique and distinguishable patterns in sentence length and vocabulary variety. Because AI text is often more uniform than human text, Standard Deviation of Sentence Length is an important feature. Based on this research, our metrics for variance in human text should be high, while AI text stays consistently uniform.

Readability is another factor. (Mindner et al., 2023) found that combining perplexity and readability scores can create an accurate classifier. (Shah et al., 2023) also highlighted metrics like the Gunning Fog Index and Flesch Reading Ease as being unique in AI text. (Rujeedawa et al., 2025) achieved good results by simply counting punctuation and measuring text length. We calculate these readability scores and punctuation counts to measure the text's complexity and structure.

3 Methodology

3.1 Research Design

We focused our research on isolating stylistic features to test the accuracy of a model based only on these metrics. We designed a supervised machine learning pipeline that relies on numeric features calculated from the raw text and replicated this experiment across three datasets to verify pattern consistency.

3.2 Data Collection

To capture different text lengths and models, we utilized data from three publicly available sources:

- **Dataset A**: The "AI vs Human Text" dataset (Gerami, 2023), which served as the primary baseline. It contains approximately 487,000 samples.
- **Dataset B**: The "Human vs AI Sentences" dataset (Hassan, 2023). This contains roughly 19,000 shorter text samples to test the model on shorter inputs.
- **Dataset C**: The "LLM Detect AI Generated Text" dataset (Thite, 2023), consisting of about 27,000 samples, was used as a secondary validation set.

For all three datasets, we performed data cleaning using Python's Pandas library, removing duplicate entries and dropping rows with missing values. We kept only the raw text data for the analysis. We intentionally bypassed standard NLP normalization steps, like lower-casing or stripping punctuation, because those elements could be signals for stylistic analysis.

3.3 Analysis Methods

The analysis consisted of two stages: feature extraction and predictive modeling.

Feature Extraction: We calculated 16 numeric features divided into three stylistic categories using the TextStat and NLTK libraries on the raw data:

1. **Structure**: Average Sentence Length, Average Word Length, Character Count, and Standard Deviation of Sentence Length (to measure structural variance).
2. **Readability**: Gunning Fog Index, Flesch Reading Ease, Simpson's Index, and Herdan's C (to quantify vocabulary richness and text difficulty).
3. **Morphology and Tone**: Densities of Nouns, Adjectives, Adverbs, Pronouns, and Auxiliary Verbs; Punctuation Ratio; and VADER-based Neutrality Scores and Sentiment Polarity.

Modeling: We trained a Logistic Regression model for each dataset because of interpretability. The model coefficients provide a direct measure of feature importance, showing which stylistic traits are most predictive. To ensure equal comparisons between features of different scales, we standardized all inputs using `StandardScaler`. We used an 80/20 train-test split with a fixed random seed for reproducibility and configured the model with balanced class weights to mitigate any class imbalances.

4 Results

We evaluated the Logistic Regression model to test both accuracy and universality. The results indicate that stylistic features alone are sufficient to distinguish AI text with high accuracy in document-length text.

4.1 Model Performance

Table 1 summarizes the classification metrics. On the primary baseline (Dataset A) and the validation set (Dataset C), the model achieved nearly

Dataset	Accuracy	$F1_H$	$F1_{AI}$
A (Baseline)	0.883	0.902	0.853
B (Sentences)	0.700	0.707	0.694
C (Validation)	0.888	0.903	0.867

Table 1: Comparison of model performance across the three datasets.

Feature	Coeff.	Prediction
Avg Sent Length	-3.84	Human
Gunning Fog	+3.66	AI
Sent Len Std Dev	-1.86	Human
Simpson's Index	+0.95	AI
Aux Verb Density	-0.91	Human
Adjective Density	+0.61	AI
Pronoun Density	+0.56	AI
Flesch Score	-0.53	Human

Table 2: Top feature coefficients for Dataset A. Positive values indicate AI-generated text; negative values indicate human text.

identical accuracy rates of approximately 88%. This consistency suggests the stylistic fingerprint is robust across large-scale corpora. Performance dropped to 70% on Dataset B, which consists of individual sentences. This was expected, as structural features like sentence length variance require a larger sample of text to calculate effectively.

4.2 Feature Importance Analysis

To identify the most significant markers, we analyzed the feature coefficients from Dataset A. Table 2 shows the features with the highest absolute influence.

1. **AI Predictors:** The strongest predictors for AI text were the **Gunning Fog Index** (+3.66) and **Simpson's Index** (+0.95). This indicates that AI text tends to be mechanically complex yet restricted in its vocabulary.

2. **Human Predictors:** The strongest predictors for human text were **Average Sentence Length** (-3.84) and **Sentence Length Standard Deviation** (-1.86). This confirms that human writing is characterized by variability, whereas AI writing is more uniform.

These patterns held consistent across Dataset C as well, where Gunning Fog (+3.60) and Average Word Length (+2.39) remained the strongest AI predictors.

5 Discussion

Given the results, there is strong evidence that AI text leaves behind a stylistic fingerprint. Without analyzing a semantic word, we achieved 88% accuracy on document-level datasets. Structural fingerprints are clear indicators of authorship, with AI being more rigid and mechanically complex, but showing less vocabulary diversity than human writing.

5.1 Structural Rigidity of AI

The most significant finding from our analysis is the contrast in sentence structure. The negative coefficient for **Sentence Length Standard Deviation** (-1.86) confirms that human writing tends to naturally vary, alternating between short and long complex clauses. On the other hand, AI models stay consistent. The lack of variance is likely a result of the model optimizing for the next most probable word, which smooths out the natural irregularities of human thought. Research by (Jaashan and Bin-Hady, 2025) documents this rigidity, and our findings confirm it.

5.2 Complexity vs. Diversity

We found a contradiction when comparing complexity metrics. AI text had a higher **Gunning Fog Index** (+3.66), indicating it uses more complex words and sentence structures. However, it also had a higher **Simpson's Index** (+0.95), which indicates lower vocabulary diversity. This suggests that while AI uses sophisticated words to appear authoritative, it recycles them more frequently within a document than a human would. This leads to repetitive ideas and words, which our analysis captured.

5.3 Tone and Mechanics

The positive weights for **Adjective Density** (+0.61) and **Pronoun Density** (+0.56) are consistent with (Mitrović et al., 2023) and the "helpful assistant" persona. AI text tends to be descriptive and plain. In contrast, the negative weight for **Auxiliary Verb Density** (-0.91) suggests humans use more functional and active language compared to AI's affirmative style.

5.4 Impact of Text Length

The model's performance drop to 70% on Dataset B (single sentences) highlights the dependency of

this method on text length. Structural features, especially **Sentence Length Standard Deviation**, require a sufficient sample of sentences to calculate a meaningful signal. When the text is too short, the fingerprint becomes too faint to detect reliably. While our stylistic analysis is sufficient for essays or articles, it becomes less effective on short-form content.

6 Conclusion

6.1 Main Findings

This study set out to answer two research questions. First, *can stylistic features detect AI text independently?* Our results show the answer is yes. We achieved 88.3% accuracy using only 16 stylistic features. Second, *which features are most indicative?* We found the clearest signals of AI authorship are a high **Gunning Fog Index** and a high **Simpson's Index**. The clearest signals of human authorship are a high **Sentence Length Standard Deviation** and longer **Average Sentence Length**.

6.2 Limitations

A significant limitation in any text classification is text length. Stylistic features, particularly variance metrics like **Sentence Length Standard Deviation**, require a sufficient sample of text to provide a reliable signal. Our 70% accuracy result on Dataset B confirms that the fingerprint is less consistent on sentence-level inputs. This aligns with findings from (Rujeedawa et al., 2025), who noted that models struggle with shorter texts because they "lack the necessary linguistic and stylistic features that longer texts provide."

6.3 Future Research

Future research should focus on the durability of these fingerprints against prompts specifically designed to mask them. Exploring the resilience of stylistic features against intentional manipulation is an important next step. Additionally, this method must be continually validated against new AI models and hybrid human-AI texts to ensure its long-term effectiveness.

References

Georgios P. Georgiou. 2025. Differentiating between human-written and AI-generated texts using automatically extracted linguistic features. *Information*, 16(11):979.

Shayan Gerami. 2023. AI vs human text. https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text.

Shahxeeb Hassan. 2023. Human vs ai sentences. https://huggingface.co/datasets/shahxeebhassan/human_vs_ai_sentences.

Hasan M. S. Jaashan and Wagdi Rashad Ali Bin-Hady. 2025. Stylometric analysis of AI-generated texts: a comparative study of ChatGPT and DeepSeek. *Cogent Arts & Humanities*, 12(1):2553162.

L. Mindner, T. Schlippe, and K. Schaaff. 2023. Classification of human- and AI-generated texts: Investigating features for ChatGPT. In *Artificial Intelligence in Education Technologies: Proceedings of AIET 2023*, Singapore. Springer.

S. Mitrović, D. Andreoletti, and O. Ayoub. 2023. ChatGPT or human? detect and explain: Explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv preprint arXiv:2301.13852.

OpenAI. 2025. Chatgpt. https://chat.openai.com.

Muhammad Irfaan Hossen Rujeedawa, Sameerchand Pudaruth, and Vusumuzi Malele. 2025. Unmasking AI-generated texts using linguistic and stylistic features. *International Journal of Advanced Computer Science and Applications*, 16(3):213–221.

Aditya Shah, Prateek Ranka, Urmi Dedhia, Shruti Prasad, Siddhi Muni, and Kiran Bhowmick. 2023. Detecting and unmasking AI-generated texts through explainable artificial intelligence using stylistic features. *International Journal of Advanced Computer Science and Applications*, 14(10):1043–1053.

M. Soni and V. Wade. 2023. Comparing abstractive summaries generated by ChatGPT to real summaries through blinded reviewers and text classification algorithms. arXiv preprint arXiv:2303.17650.

Sunil Thite. 2023. Llm detect ai generated text dataset. https://www.kaggle.com/datasets/sunilthite/llm-detect-ai-generated-text-dataset.