Style-First Authorship Verification for Academic Integrity in the Generative AI Era (Student Abstract)

Jun Jang[1], **Thai Le**[2], **Bo Wang**[3]
[1]Oxford High School, MS, USA
[2]Indiana University, IN, USA
[3]University of Mississippi, MS, USA
jun.green.jang@gmail.com, tle@iu.edu, hbw@olemiss.edu

Abstract

With the rise of Generative AI (Gen AI), academic dishonesty in classrooms has skyrocketed, yet the existing solutions for detecting such dishonesty often fall short. Traditional "AI detectors" merely analyze one text at a time, failing to account for students' previous writings, which risks false positives. Meanwhile, many authorship verification (AV) models fail to analyze the nuances in writing styles that truly distinguish authorship. To fill this existing gap, we propose an AV framework that combines token-level stylometric features (e.g., POS tag patterns) with handcrafted stylistic features (e.g., sentence structure variation) to construct a comprehensive feature set. Using both benchmark corpora and real-world high school student essays, we trained multiple machine learning binary classifiers on these hybrid vectors. Our initial experiments show that this framework outperforms the standard token-only baselines by over 25%, while offering interpretable, style-based insights. Looking ahead, we plan to extend this work with large language models and multi-agent approaches to further enhance robustness and adaptability. These preliminary results highlight the importance of nuanced stylistic features and suggest that a holistic AV system can provide educators with more reliable and transparent detection tools.

Code — https://github.com/jjang-07/academic-authorship-verification

Introduction

The rise of generative AI has raised concerns about academic integrity. Studies report that a majority of high school students admit to cheating on assignments and writing using such tools (Edutopia (Online) 2024). While Gen AI offers various learning benefits, empirical studies show that an overreliance on these tools erodes creativity and critical thinking, which are crucial for the future of innovation (Gerlich 2025).

There are currently two detection approaches: standard "AI detectors" and Authorship Verification (AV) models. Standard "AI detectors" operate in a one-shot manner, classifying a single essay in isolation as either human or AI-written (Yang et al. 2025). Without referencing a student's past work, they are prone to false positives due to their ignorance of one's writing style. In contrast, *authorship verification* (AV) compares a student's new text against their prior submissions, focusing on whether the style matches. Prior AV methods often rely heavily on embeddings or token-level models. While effective in some contexts, these approaches fail to capture the nuanced stylistic patterns that truly distinguish authorship and remain uninterpretable for educators.

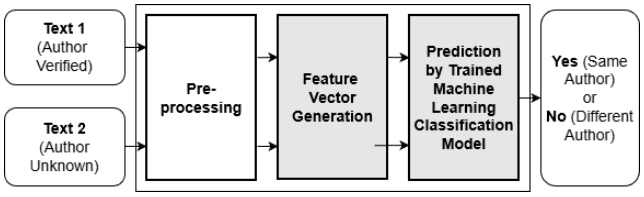

Figure 1: Authorship verification workflow

We aim to address this gap with an interpretable, style-first AV framework. The system compares two input texts, one with a verified author and another of uncertain origin, and preprocesses them to clean noisy information. Each text is then presented as a comprehensive feature vector that combines token-level baselines with handcrafted stylistic features that we designed. Our trained machine learning classifier evaluates these vectors and outputs the probability that the two texts were written by the same author. To ensure transparency, Shapley value analysis highlights which features drive each decision, giving educators interpretable insights. Figure 1 visualizes the overall flow of the authorship verification process. Preliminary code is available on the GitHub Repository.

Methodology

We view *writing style* as a collection of measurable features, which are individual attributes that capture aspects of an author's text. Our AV framework consists of three parts: data collection and preprocessing, feature development and extraction, and classification using machine learning. Figure 2 visualizes the 3 major components.

Data and Preprocessing: For preliminary training and testing, we used the Reuters_50_50 dataset (Liu 2006), which was previously applied in prior AV research. The dataset consists of 5,000 unique texts between 100 authors (50 per

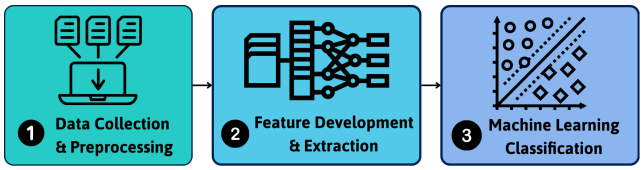

Figure 2: Methodology Overview

Model	AUC	c@1	$F_{0.5}$	F1	Brier	Overall
Logistic Regression (Handcrafted)	0.593	0.563	0.569	0.590	0.755	0.614
Gradient Boosting (Handcrafted)	0.627	0.591	0.594	0.612	0.761	0.637
Logistic Regression (Full Set)	0.711	0.651	0.650	0.653	0.782	0.689
Gradient Boosting (Full Set)	0.745	0.674	0.680	0.647	0.790	0.707
Wan (2024) Logistic Regression	0.565	0.525	0.548	0.613	0.557	0.562

Table 1: Performance comparison on Reuters_50_50 dataset.

author) with topical variation, which provides a strong baseline for AV. Additionally, we have also collected a small set of high-school essays across subjects such as history, English, and technology to directly mimic an academic setting. We have also identified additional formal datasets (e.g., arXiv abstracts, British Academic student writing, and Project Gutenberg novels) for future expansion to improve robustness.

Preprocessing included a variety of processes such as normalization to lowercase, tokenization, POS tagging, dependency parsing using NLTK (Bird, Klein, and Loper 2009) and spaCy (2025), and more. Finally, we generated balanced positive (same-author) and negative (different-author) text pairs by shuffling and iteratively combining within and across authors to ensure unbiased training for classification.

Features: We constructed a comprehensive feature vector that integrates two components: (1) token-level baselines drawn from prior work (Wan 2024; Weerasinghe, Singh, and Greenstadt 2021), including TF–IDF representations and common stylometric features, and (2) a novel set of handcrafted stylistic features designed to capture nuances in writing style. The latter include vocabulary distributions (via CEFR levels), sentence length and variance, sentence structure variation, perplexity, readability metrics, and adverbial placement patterns. By combining token-level and interpretable stylistic features, the vector captures both surface-level and deeper stylistic cues to truly distinguish authorship, while remaining transparent enough for educators to interpret through Shapley value analysis.

Machine Learning: The extracted feature vectors were used to train and evaluate 5 binary classifiers: Logistic Regression, Random Forest, Support Vector Machines, K-Nearest Neighbors, and Gradient Boosting. Because of the relatively small datasets, we applied k-fold cross-validation rather than a fixed train–test split, averaging results across folds to get a robust performance estimate. Hyperparameters were optimized with grid search, and model performance was evaluated using five metrics common in AV tasks: AUC, $c@1$, F0.5, F1, and Brier score. These metrics provide a balanced evaluation of our model and enables comparability with prior AV benchmarks.

Experimental Results & Discussion

Table 1 summarizes our initial performance on the Reuters_50_50 dataset across different feature sets and models. (Handcrafted) indicates our stylistic features alone, and (Full Set) refers to our comprehensive feature set that concatenates handcrafted and token-level features. Overall, incorporating stylistic features consistently improved performance across all metrics when compared to token-only baselines. Gradient Boosting with the full feature set achieved the strongest results (Overall = 0.707), while Logistic Regression also performed competitively (Overall = 0.689). In contrast, token-only baselines such as Wan (2024) underperformed, highlighting the limitations of relying exclusively on token-level signals. These findings suggest that nuanced stylistic features capture dimensions of authorship overlooked by token-based models, offering a more interpretable and robust approach to authorship verification.

Limitations & Future Work

While our initial results demonstrate the promise of style-first authorship verification, our approach is limited by dataset availability and the fact that no single feature works reliably across all contexts, meaning topic shifts can still undermine performance. In future work, we plan to expand our datasets to include larger, more diverse corpora and refine our feature set by incorporating additional stylistic attributes. Beyond handcrafted features, we are exploring the integration of large language models (LLMs), both through in-context learning and multi-agent frameworks where one model generates AV predictions and another provides corrective feedback grounded in our handcrafted features. These directions aim to improve robustness while preserving transparency in real-world educational contexts.

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Edutopia (Online). 2024. Fostering Students' Academic Integrity.

Gerlich, M. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1).

Liu, Z. 2006. Reuter_50_50 [Dataset].

spaCy (online). 2025. spaCy Library Architecture.

Wan, S. 2024. Transparent Authorship Verification with Machine Learning Models. https://math.mit.edu/research/highschool/primes/materials/2024/Wan.pdf.

Weerasinghe, J.; Singh, R.; and Greenstadt, R. 2021. Feature Vector Difference based Authorship Verification for Open-World Settings. In *CLEF (Working Notes)*, 2201–2207.

Yang, Z.; Feng, Z.; Huo, R.; Lin, H.; Zheng, H.; Nie, R.; and Chen, H. 2025. The Imitation Game revisited: A comprehensive survey on recent advances in AI-generated text detection. *Expert Systems with Applications*, 272: 126694.