DocInsight: A Semantic-AI Document Intelligence System

Shaurya Patil
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
shaurya.patil26@nmims.in

Vedant Kothari
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
vedant.kothari05@nmims.in

Tanishq Nabar
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
tanihsq.nabar137@nmims.in

Pranshu Chaniyara
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
pranshu.chaniyara152@nmims.in

Dr. Dhirendra Mishra
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering
Mumbai, India
dhirendra.mishra@nmims.edu

***Abstract*—Academic integrity is a cornerstone of academic life, yet current plagiarism detection tools like Turnitin are still hampered by surface-string matching. These tools are often unable to detect paraphrasing, misidentify domain-specific jargon, and are incapable of detecting stylistic incongruities indicating ghostwriting or AI-supported writing. To overcome such limitations, this paper introduces DocInsight, a semantic–stylometric document intelligence system aimed at ensuring accurate and interpretable authorship verification. The suggested method combines domain-adapted sentence embeddings with a two-stage retrieval and reranking pipeline for semantic similarity identification and uses a stylometric ensemble classifier to ensure writing-style abnormalities. Both semantic and stylometric evidences are combined through a calibrated model to produce sentence-level reports with highlighted matches and stylistic comments. Targets of evaluation include high precision in detection of paraphrases, better robustness to AI-generated text, and increased reviewer trust through clear evidence presentation. Through the integration of semantic embeddings, stylometric signals, and explainable reporting, DocInsight adds a next-generation solution to plagiarism detection and authorship verification, providing a more accurate alternative to current tools.**

***Index Terms*—Plagiarism Detection, Semantic Similarity, Stylometric Analysis, Authorship Verification, Sentence-BERT (SBERT), FAISS Retrieval, Cross-Encoder Reranker, Academic Integrity, AI-generated Text Detection**

I. Introduction

Academic integrity is the bedrock of higher learning and scholarship. Originality in academic writing is not only essential to upholding fairness and credibility but also to promoting authentic intellectual development. With the growing prominence of digital materials, paraphrasing engines, and large language models (LLMs), the challenges to maintaining academic honesty have multiplied. Scholars and students now have the ease of manipulating content—by paraphrasing, stylistic camouflage, or ghostwriting—making it more challenging for conventional plagiarism detection software to make correct evaluations.

Among these, Turnitin has become the most universally accepted solution among universities and research centers. Though good at catching verbatim copying, Turnitin is largely based on lexical similarity and n-gram overlap. This string-matching basis is fraught with major limitations. First, it finds it difficult to detect paraphrased or semantically equivalent content, usually permitting reworded plagiarism to escape detection [1], [7]. Second, it generates inflated similarity scores in domain-based environments, where technical terms and generic phrasing are wrongly identified as plagiarism [1]. Third, it supplies only coarse document-level similarity percentages without supplying sentence-level evidence, which restricts reviewer trust [7]. Last, it does not have strong authorship inconsistency detection mechanisms that are becoming important with the advent of AI-powered writing and contract cheating [5], [9].

Current NLP and computational linguistics advances present more encouraging options. Latent Semantic Scaling (LSS) and hierarchical semantic networks are semantic embedding techniques that provide strong cross-domain and cross-lingual semantic analysis capabilities, surpassing conventional string-matching methods [3], [4]. Concurrently, stylometric methods such as neural stylometry models [2], topic-debiased representation learning (TDRLM) [6], and attention-based similarity learning [10] provide deeper insights into style of writing, consistency of authorship, and anomalies that can indicate ghostwriting or AI-generated work [5]. Hybrid systems incorporating semantic similarity in combination with stylometric cues have demonstrated improved performance across various domains [2], [3], [6], but such systems are underresearched in actual scholarly use [9].

In response to these shortcomings, this paper introduces DocInsight, a semantic–stylometric document intelligence

platform. As opposed to standard string-matching systems, DocInsight incorporates semantic embeddings with stylometric anomaly detection for fine-grained, sentence-level analysis. Its design marries SBERT-based embeddings with FAISS search and cross-encoder reranking for semantic detection with a stylometric classifier that detects function-word usage, sentence length distributions, and topic-debiased style features. An evidence fusion engine aggregates such signals to create explainable, reviewer-friendly reports which capture semantic overlap and stylistic divergence.

By integrating state-of-the-art semantic and stylometric techniques into an open and interpretable system, DocInsight seeks to improve plagiarism detection and authorship attribution. Not only does this study aim to overcome the structural constraints of existing tools such as Turnitin, but it also adds to the general academic discussion on creating reliable, AI-robust integrity systems for research and education.

II. Literature Review

1. Limitations of existing plagiarism detection tools Plagiarism detection has traditionally been dominated by commercial tools like Turnitin, which basically work on string-matching and n-gram overlap. These methods excel on verbatim copying but lack in identifying semantic equivalence and stylistic manipulation. Emmanuel et al. [1] point out the way Turnitin overestimates domain-specific semantics, exaggerating similarity scores in certain contexts like agricultural extension writing. Likewise, Mphahlele and McKenna [7] demonstrate that Turnitin typically lacks sentence-level granularity, instead generating black-box similarity scores that undermine reviewer confidence. This surface-level reliance on overlap makes these tools susceptible to paraphrasing, ghostwriting, and AI-aided rewording, now more common in submitted academic work.

2. Advances in Stylometric Analysis

Stylometry — computational analysis of style of writing — offers an orthogonal solution to detecting plagiarism by recording linguistic and stylistic prints of writers. Ding et al. [2] illustrated that neural stylometric models, which are trained to learn distributed representations of style, are superior to conventional static features like n-grams and word counts. Grieve [9] also illustrated that stylometric signals need to be untangled from subject matter since register and topical variation can mislead authorship attribution.

Some new developments include topic-debiased representation learning (TDRLM) by Hu et al. [6], which boosts authorship verification by disentangling stylistic attributes from topical bias. The approach considerably enhances domain generalizability. Stylometry has also been applied to AI detection: Zaitsu and Jin [5] employed Japanese stylometric analysis to identify texts written by humans and by ChatGPT, showing the promise of style-based approaches in detecting AI-augmented writing. These developments render stylometry a potent instrument for not just plagiarism detection, but also authorship determination and integrity assurance in the age of generative AI.

3. Advances in Semantic Analysis

In addition to surface string matching, semantic embeddings facilitate identification of similarity at the meaning level. Watanabe [3] presented Latent Semantic Scaling (LSS), a semi-supervised approach that identifies fine-grained semantic cues in new domains and languages. Liu et al. [4] built upon this with hierarchical semantic networks and dual link prediction, allowing semantic similarity analysis over large corpora and technical fields.

These methods bypass the lexical constraints of classical tools by placing sentences within high-dimensional semantic space, enabling paraphrased or reorganized text to be identified as semantically equivalent. Significantly, these models are domain-independent, ensuring that they are applicable for cross-disciplinary academic writing wherein vocabulary and phrasing tend to differ significantly.

4. Hybrid Semantic–Stylometric Approaches

Whereas semantic embeddings are able to capture meaning and stylometry is able to capture the style of the author, current research calls for the need to bridge both together in detecting plagiarism. Distributed style representations, Ding et al. [2] demonstrated, can complement semantic cues, while Hu et al. [6] verified that topic-debiasing works significantly in minimizing false positives when combining stylometry with semantic features.

Boenninghoff et al. [10] introduced an attention-based similarity learning approach to authorship authentication for social media, which obtained explainable performance by pointing out semantic overlap and stylistic signals at the same time. Likewise, Nadeem et al. [8] created a multimodal system incorporating semantic similarity and stylometric signals for detecting fake news, highlighting the importance of hybrid pipelines for intricate text analysis tasks.

These studies cumulatively show that hybrid systems surpass single-method systems. Such pipelines, however, are underutilized in academic integrity applications where the majority of tools still depend on outdated string-matching methods.

5. Gaps Identified Through Research

Despite progress, major gaps exist: Explainability: Most existing systems hardly offer sentence-level, reviewer-friendly evidence [7], [10].

Cross-domain and cross-lingual detection: Semantic models are available but not yet widely applied in plagiarism tools [3], [4].

AI-based writing: Stylometric camouflage and ghostwriting create new challenges, demanding strong detection [5], [9].

Real-world integration: Scaleable, domain-adaptive hybrid systems have yet to be fully achieved for educational scenarios [6], [8].

6. Lessons for DocInsight

The literature reviewed leads toward a second-generation solution combining semantic embeddings and stylometric anomaly detection. DocInsight follows on directly from these lessons by: Applying SBERT + FAISS retrieval to model semantic paraphrases [3], [4].

Utilizing stylometric anomaly detection for authorship authentication [2], [5], [6], [9].

Merging semantic and stylistic proof with interpretable outputs [8], [10].

Overcoming the limitations of Turnitin through sentence-level transparency [1], [7].

By doing so, DocInsight conforms to cutting-edge research while meeting practical and ethical demands of academic honesty systems. Please do not revise any of the current designations.

TABLE I
RESEARCH GAPS MATRIX

Topic/Method	DL	TD	MR	MM	GB
Domain semantics	2	1	2	GAP	GAP
Authorship verify	3	2	1	1	1
Paraphrase detect	2	1	1	1	GAP
Cross-lingual	1	GAP	GAP	1	1
AI-text detect	2	GAP	GAP	1	GAP

Abbreviations: DL = Deep Learning Models, TD = Topic-Debiasing, MR = Manual Review, MM = Multimodal Approaches, GB = Graph-Based Models.

III. PROPOSED SYSTEM (METHODOLOGY)

The system under consideration, DocInsight, is a stylometric–semantic document intelligence system aimed at transcending the limitations of the existing plagiarism detection systems. Unlike the traditional string-matching methods, DocInsight combines semantic similarity detection with stylometric anomaly detection to offer sentence-level, interpretable proof of potential plagiarism or author inconsistencies.

1. System Architecture

The overall architecture of DocInsight is comprised of six interlinked modules: Document Parsing and Preprocessing

Supported formats: .docx and .pdf.

Libraries: PyMuPDF (PDF), python-docx (DOCX), and spaCy for sentence segmentation.

Preprocessing involves: removal of reference/bibliography section, normalisation of citations, and boilerplate filtering to reduce false positives.

Semantic Embedding Generator

Embedding model: Sentence-BERT (SBERT), pre-trained with paraphrase-mpnet-base-v2 or all-MiniLM-L6-v2. Fine-tuned on academic paraphrase datasets (PAWS, Quora QP, synthetic adversarial paraphrases) to understand academic writing semantics. Outputs high-dimensional vectors (384–768 dimensions) per sentence.

Fast Candidate Retrieval (FAISS)

Embeddings are indexed with Facebook AI Similarity Search (FAISS). Each question sentence fetches the top-k candidate matches (default k=5) for effective semantic similarity search.

Cross-Encoder Reranker

A transformer cross-encoder (BERT/SRoBERTa) reranks the retrieved candidates. Outputs higher precision by scoring semantic equivalence among sentence pairs to minimize false positives.

Stylometric Feature Extractor and Anomaly Detector

Features extracted are: average length of sentence, variance of sentence length, frequency of function words, POS distribution ratios, density of punctuation, Flesch Reading Ease, and type-token ratio (TTR).

Classifier: RandomForest or XGBoost for style classification; Isolation Forest for outlier detection. Identifies inconsistencies in writing style likely to signal ghostwriting or AI-aided insertions.

Evidence Fusion and Explainable Reporting

Fuses semantic similarity scores, cross-encoder probabilities, and stylometric anomaly scores through a calibrated logistic regression model. Produces side-by-side explainable reports with: Suspicious sentences highlighted. Corresponding source snippets matched. Semantic similarity and stylometric anomaly comments. Outputs in PDF, JSON, and an interactive Streamlit interface.

2. Workflow Summary

Input documents are preprocessed and uploaded. Each sentence is embedded using SBERT and indexed in FAISS. Candidate matches are retrieved and reranked by the cross-encoder. Stylometric features are analyzed for detecting anomalies. Evidence is combined and presented in an explainable report.

3. Key Innovations

Hybrid Detection: Integrates semantic embeddings with stylometric analysis, performing better than single-method techniques [2], [6], [8]. Sentence-Level Explainability: Offers evidence at the sentence level, as opposed to Turnitin's coarse document-level scores [7]. AI-Resilience: Identifies ghostwriting and AI-supported writing through stylometric irregularities [5]. Custom Corpus Support: Supports comparison with proprietary datasets for institutional and academic use. Privacy-Preserving: Complete offline functionality, such that sensitive documents are never left in the cloud.

IV. IMPLEMENTATION

V. EVALUATION METRICS

VI. CONCLUSION AND FUTURE WORK

REFERENCES

[1] O. Emmanuel, A. Ogbonnaya, I. Christiana, C. Comfort, and D. Nsongurua, "Intricacies of utilizing the Turnitin tool in agricultural extension content writing in Nigeria," *African Journal of Agricultural Research*, vol. 18, no. 8, pp. 393–401, 2023.

[2] S. Ding, B. Fung, F. Iqbal, and W. Cheung, "Learning stylometric representations for authorship analysis," *IEEE Transactions on Cybernetics*, vol. 49, no. 1, pp. 107–121, Jan. 2016.

[3] K. Watanabe, "Latent semantic scaling: A semisupervised text analysis technique for new domains and languages," *Communication Methods and Measures*, vol. 15, no. 2, pp. 81–102, 2020.

[4] Z. Liu, J. Feng, and L. Uden, "Technology opportunity analysis using hierarchical semantic networks and dual link prediction," *Technovation*, vol. 125, pp. 102872, 2023.

[5] W. Zaitsu and M. Jin, "Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis," *PLOS ONE*, vol. 18, no. 8, pp. 1–18, Aug. 2023.

[6] X. Hu, W. Ou, S. Acharya, S. Ding, R. D'Gama, and H. Yu, "TDRLM: Stylometric learning for authorship verification by topic-debiasing," *Expert Systems with Applications*, vol. 233, pp. 120745, 2023.

[7] A. Mphahlele and S. McKenna, "The use of Turnitin in the higher education sector: Decoding the myth," *Assessment & Evaluation in Higher Education*, vol. 44, no. 7, pp. 1079–1089, 2019.

[8] M. Nadeem, K. Ahmed, Z. Zheng, D. Li, M. Assam, Y. Ghadi, F. Alghamedy, and E. Tag-Eldin, "SSM: Stylometric and semantic similarity-oriented multimodal fake news detection," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 6, pp. 101559, 2023.

[9] J. Grieve, "Register variation explains stylometric authorship analysis," *Corpus Linguistics and Linguistic Theory*, vol. 19, no. 1, pp. 47–77, 2023.

[10] B. Boenninghoff, S. Hessler, D. Kolossa, and R. Nickel, "Explainable authorship verification in social media via attention-based similarity learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 36–45.