Transformer-Based Authorship Attribution for Kant and Nietzsche

Xiaohan Wu

Jan 2nd 2026

Abstract

Authorship attribution in philosophical texts is challenging due to substantial semantic overlap across authors. This paper examines whether transformer-based models can distinguish authorial style independently of topic by studying Immanuel Kant and Friedrich Nietzsche. We fine-tune a DistilBERT classifier and evaluate cross-work generalization on entirely unseen books, achieving 97.2% accuracy. To control for semantic content, we combine topic diagnostics using BERTopic with embedding-based similarity matching between test chunks and opposite-author training text. Classification accuracy remains high under increasing semantic overlap (96.5% at cosine similarity ≥ 0.5 and 95.1% at ≥ 0.6), indicating that stylistic signals persist beyond topical cues. An interpretability analysis using LIME shows that predictions rely on linguistically meaningful features such as function words, logical connectors, and abstract philosophical terminology. These results demonstrate that fine-tuned transformer models can encode robust and interpretable authorial style in semantically dense domains.

1 Introduction

Authorship attribution has long been a central problem in linguistics, literary studies, and the digital humanities raising fundamental questions about whether and how individual writing style can be distinguished from topical content. Philosophical texts present an especially challenging case because they are highly abstract, conceptually dense, and often engage overlapping themes such as ethics, reason, and morality. As a result, semantic cues alone are often insufficient to separate authors, making stylistic discrimination challenging. Successfully identifying authors based on style provides a stringent test of whether language models capture deeper linguistic structure, such as syntax, rhetorical organization, and discourse patterns, rather than relying primarily on topic-specific vocabulary.

Recent advances in natural language processing, especially transformer-based models such as BERT [1] and DistilBERT [2], have achieved remarkable performance on text classification tasks. However, high accuracy alone does not guarantee that these models learn meaningful stylistic representations. In many cases, strong performance may reflect sensitivity to surface-level semantic differences or domain-specific keywords. This concern is particularly relevant in high-stakes applications such as forensic authorship verification, legal document analysis, and political speech attribution, where distinguishing *how* something is said rather than *what* is said is crucial.

From a theoretical perspective, this question connects to long-standing debates in computational stylometry about whether authorial style constitutes a stable and measurable signal. Classical approaches emphasized frequency-based features such as function words and sentence-length distributions [3], while later work incorporated syntactic and discourse-level features [4]. Transformer models offer a qualitatively different approach by encoding contextualized token interactions across entire sequences via self-attention. If such models can reliably distinguish authors even when semantic content overlaps, this would provide evidence that stylistic information is embedded in higher-order linguistic structure and is recoverable without explicit feature engineering.

Beyond theoretical interest, understanding stylistic learning in transformer models has practical implications. Fine-tuned language models are increasingly used in authorship verification, plagiarism detection, and personalized text analysis, and parameter-efficient methods such as LoRA [5] have lowered the barrier to deploying such systems. It is essential to establish whether these models genuinely encode stylistic signals and not merely topical correlations for their responsible use in applications where interpretability and trustworthiness matter.

2 Research Question and Methodology

We investigate whether transformer-based models can distinguish between the writing styles of two canonical philosophers, Immanuel Kant and Friedrich Nietzsche, independently of topical content. These authors provide a natural test case. Both engage extensively with overlapping philosophical domains such as ethics, morality, and reason, yet their writing styles are sharply contrasted in the literature: Kant's systematic, clause-heavy exposition differs markedly from Nietzsche's aphoristic and polemical prose [6, 7]. In addition, both authors wrote extensively, providing sufficient data for cross-work evaluation.

Our evaluation proceeds in two stages. First, we test whether a fine-tuned DistilBERT classifier can generalize across unseen works, assessing whether it learns transferable authorial style rather than book-specific patterns (H1). Second, we control for semantic content using topic modeling with BERTopic [8] and embedding-based similarity matching to assess whether classification performance persists when topical cues are minimized (H2). Finally, we use local explanation methods (LIME) [9] to identify which linguistic features drive model predictions.

2.1 Hypotheses

We test two primary hypotheses:

- **H1: Cross-Work Generalization**

  A fine-tuned DistilBERT model achieves high classification accuracy ($> 80\%$) on completely unseen works, indicating that it learns generalizable stylistic patterns rather than memorizing work-specific features.

- **H2: Style Beyond Semantics**

  When semantic content is controlled either through topic modeling (BERTopic) or embedding-based similarity matching with cosine similarity ≥ 0.6, the model maintains high classification accuracy ($> 70\%$), demonstrating reliance on stylistic rather than purely topical cues.

These hypotheses clarify what transformer models learn during authorship classification and assess whether stylistic information can be reliably distinguished from semantic content in complex philosophical texts.

2.2 Data Collection and Cleaning

We constructed the corpus from Project Gutenberg[1], collecting twelve public-domain books in English translation: six works by Immanuel Kant and six by Friedrich Nietzsche. To prevent spurious authorship cues, we removed Project Gutenberg boilerplate (e.g., licensing information and metadata) by retaining only the main body text of each work. This preprocessing step ensures that classification performance reflects authorial language rather than platform-specific artifacts.

2.3 Chunking and Train/Validation/Test Split

Texts were tokenized using the DistilBERT tokenizer and segmented into contiguous chunks of 125 tokens, discarding segments shorter than 80 tokens. Token-based chunking produces uniformly sized inputs suitable for transformer fine-tuning while preserving local stylistic patterns such as function-word usage, clause structure, and punctuation.

To evaluate cross-work generalization and avoid leakage, data were split at the *work* level. One median-sized work per author was held out for testing (Kant: *Critique of Practical Reason*; Nietzsche: *Twilight of the Idols*), producing 1,377 test chunks. The remaining works formed the training and validation pool, with 85% of chunks used for training and 15% for validation, stratified by author. This design ensures that evaluation reflects generalization across books rather than memorization of shared content.

2.4 Fine-Tuning Setup

We fine-tuned `distilbert-base-uncased` for binary sequence classification (Kant vs. Nietzsche) using the Hugging Face `Trainer`. Training employed a learning rate of 2×10^{-5}, batch size 16 (32 for evaluation), weight decay 0.01, and macro F1 as the model-selection criterion. Models were trained for up to five epochs with early stopping (patience = 2) based on validation performance. The checkpoint with the highest validation F1 was retained for all downstream evaluations. All experiments are fully reproducible, with code and configurations available online.[2]

3 H1: Cross-Work Generalization

3.1 Results on Held-out Works

The fine-tuned DistilBERT model achieves strong cross-work generalization on unseen texts. On the held-out test set of 1,377 chunks, the model correctly classifies 1,338 samples, yielding an overall accuracy of 97.2%. Macro-averaged F1 is 0.971 and weighted F1 is 0.972, indicating balanced performance across authors despite moderate class imbalance.

[1] `https://www.gutenberg.org/`. Project Gutenberg is a free digital library providing public-domain literary works, including many historical philosophical texts.

[2] `https://github.com/juliawuxh/philosopher-style-classification`

Table 1 reports the confusion matrix. The model correctly identifies 589 of 610 Kant chunks (recall 96.6%) and 749 of 767 Nietzsche chunks (recall 97.7%). Precision is similarly high for both classes: 97.0% for Kant and 97.3% for Nietzsche. In total, only 39 misclassifications occur (2.83% of the test set), demonstrating that the learned representations generalize well beyond the training works.

Table 1: Confusion Matrix for Cross-Work Test Set (N=1,377)

		Predicted		
		Kant	Nietzsche	Total
True	Kant	589 (96.6%)	21 (3.4%)	610
	Nietzsche	18 (2.3%)	749 (97.7%)	767
	Total	607	770	1,377

Accuracy: 97.2% | Macro F1: 0.971 | Weighted F1: 0.972

Performance is stable across authors: Kant's recall is 96.6% while Nietzsche's recall is 97.7%.

This result is particularly notable given the sharp stylistic and structural contrast between the two works. Kant's *Critique of Practical Reason* exemplifies systematic, clause-dense ethical reasoning, whereas *Twilight of the Idols* is aphoristic and polemical, relying on rhetorical compression and cultural critique. The model achieving near-identical accuracy on these fundamentally different texts suggests it captures author-level stylistic patterns rather than work-specific structure or topic.

3.2 Prediction Confidence and Error Patterns

Correct predictions are typically made with high confidence, indicating strong separation in the model's representation space. Misclassified samples, however, do not primarily reflect low-confidence boundary cases. Among the 39 errors (2.83% of the test set), the mean prediction confidence is 0.90 and the median confidence is 0.98, with 33 errors occurring at confidence levels above 0.7.

This suggests that most errors reflect systematic ambiguity rather than uncertainty at the decision boundary. Error rates differ slightly by author: 3.44% for Kant (21/610) and 2.35% for Nietzsche (18/767). Many high-confidence errors involve passages where Nietzsche adopts a more formal, analytical tone or explicitly engages Kantian concepts, blurring stylistic distinctions despite strong overall separation.

3.3 Statistical Robustness

Using the Wilson score method, the 95% confidence interval for test accuracy is approximately [96.1%, 97.9%]. A binomial test decisively rejects both random guessing (50%) and the H1 threshold of 80% accuracy ($p < 0.001$ in both cases). These results confirm that the observed performance is statistically robust and not attributable to chance or favorable test composition.

3.4 Interpretation

The results provide strong support for H1. The model achieves 97.2% accuracy on unseen works, with balanced precision and recall across authors and a low overall error rate. The consistency of performance across distinct test texts, combined with high-confidence correct predictions, indicates that fine-tuned DistilBERT learns stable, transferable stylistic representations rather than memorizing content or exploiting work-specific features.

At the same time, the presence of high-confidence errors, often involving stylistic convergence or explicit philosophical overlap, highlights that semantic content can still exert influence. While cross-work generalization strongly suggests stylistic learning, it does not fully disentangle style from meaning. This motivates the semantic control analyses in H2, which explicitly test whether high accuracy persists when topical similarity between authors is enforced.

However, H1 results alone do not definitively establish that the model relies on stylistic rather than semantic features. Although cross-work generalization suggests the model learns transferable patterns, it remains possible that topical or thematic consistency within each author's corpus drives classification. Kant and Nietzsche engage fundamentally different philosophical projects—systematic epistemology and ethics versus existential critique and cultural polemic—which may produce distinct semantic signatures that correlate with but are separable from style. H2 addresses this concern by explicitly controlling for semantic content and evaluating whether high performance persists when topical overlap is enforced.

4 H2: Distinguishing Stylistic Signals from Semantic Content

To evaluate whether the classifier relies on stylistic cues rather than topical content, we assess performance under semantic control using two complementary approaches: topic modeling to diagnose topical separation, and embedding-based similarity to test robustness across varying degrees of semantic overlap.

4.1 Topic overlap diagnostic (BERTopic)

We apply BERTopic, an unsupervised framework combining sentence-transformer embeddings with density-based clustering, to characterize topical overlap between authors. BERTopic identifies 51 topics across the corpus with limited author overlap: mean balance ratio (minority/majority author share) is 0.072 (median: 0.018), and 47 of 51 topics are highly imbalanced ($<$20% minority representation).

Applying a viability criterion (balance ≥ 0.30, $n \geq 20$) yields three balanced topics, all appearing in the test set with 32 samples (12 Kant, 20 Nietzsche). On this topic-controlled subset, the model achieves 87.5% accuracy (28/32 correct; 95% CI: [0.72, 0.95]). While this exceeds the H2 threshold of 70%, the 9.7 percentage point reduction from overall accuracy (97.2%) suggests topical cues contribute to classification. The small sample size and class imbalance limit statistical precision, motivating a more scalable semantic control approach.

4.2 Semantic Similarity Control

For a more robust test, we evaluate performance on test chunks with high cosine similarity to opposite-author training content. For each test chunk, we identify its most similar opposite-author training chunk using sentence-transformer embeddings (`all-MiniLM-L6-v2`).

Figure 1 shows accuracy as a function of the similarity threshold. At threshold 0.60 ($n = 286$), accuracy is 95.1%—only 2.1 percentage points below overall performance. Accuracy remains above 94% through threshold 0.62, declining only at very high thresholds (≥ 0.68) where sample sizes drop below 20, reflecting statistical variance rather than systematic failure.

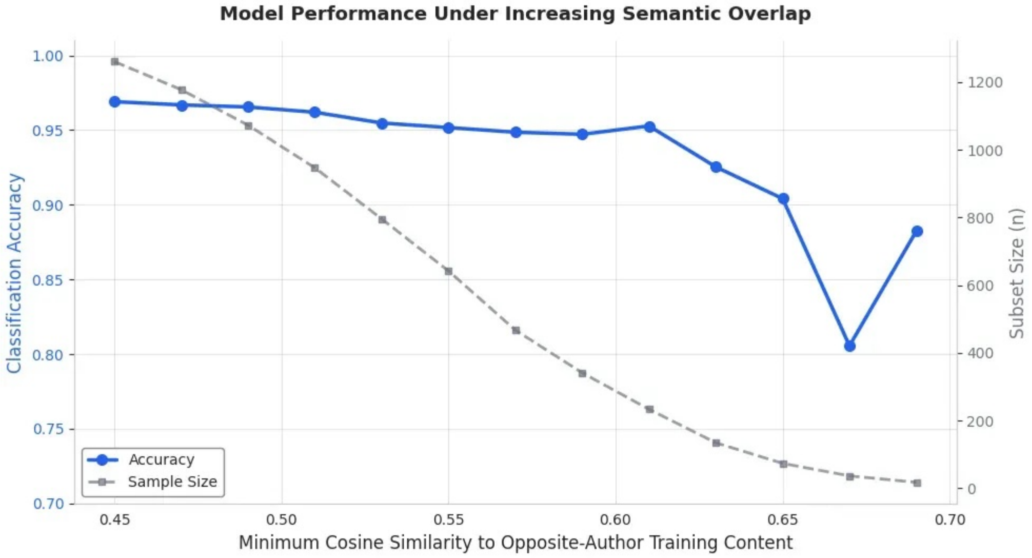

Figure 1: Classification accuracy under semantic similarity control. The blue line (left axis) shows accuracy on subsets defined by a minimum cosine similarity to opposite-author training content; the gray dashed line (right axis) shows the subset size. Accuracy remains stable (> 94%) through threshold 0.60 ($n = 286$), indicating robust stylistic discrimination despite strong semantic overlap.

4.2.1 H2 Interpretation

Both approaches support H2, though with complementary strengths. Topic-based control demonstrates discrimination on tightly constrained semantic domains (87.5%, n=32) but is sample-limited due to minimal topical overlap (mean balance: 0.072). Semantic similarity analysis provides stronger evidence with a larger, more robust sample (95.1%, n=286), showing only minimal performance degradation under semantic control.

These results indicate the model learns genuine stylistic features—syntactic structure, sentence complexity, rhetorical patterns— that generalize beyond topical content. The minimal accuracy reduction under similarity control (2.1 pp) suggests classification relies primarily on style rather than topic-specific vocabulary.

5 Interpretability Analysis: LIME Feature Attribution

To assess whether the classifier relies on linguistically meaningful stylistic cues rather than spurious correlations, we conduct an interpretability analysis using LIME. LIME provides local explanations by perturbing individual inputs and fitting a sparse linear surrogate model around each prediction, allowing us to identify which tokens contribute most strongly to classification decisions.

We generate LIME explanations for eight representative test chunks: six correctly classified cases (three per author) and two misclassifications. Across these explanations, we extract 80 feature attributions spanning 50 unique tokens. Table 2 summarizes the scope of this analysis and the overall strength of the extracted signals.

Table 2: Summary statistics for LIME feature attributions.

Statistic	**Value**
Explained test chunks	8
Total feature attributions	80
Unique tokens	50
Mean absolute feature weight	0.022
Median absolute feature weight	0.021
Mean |weight| (correct predictions)	0.024
Mean |weight| (incorrect predictions)	0.017

Correct predictions exhibit larger mean absolute feature weights than misclassifications (0.024 vs. 0.017), indicating that accurate decisions are driven by fewer, more decisive stylistic cues rather than diffuse or weak signals. This pattern suggests that clearer stylistic structure corresponds to higher model confidence and accuracy.

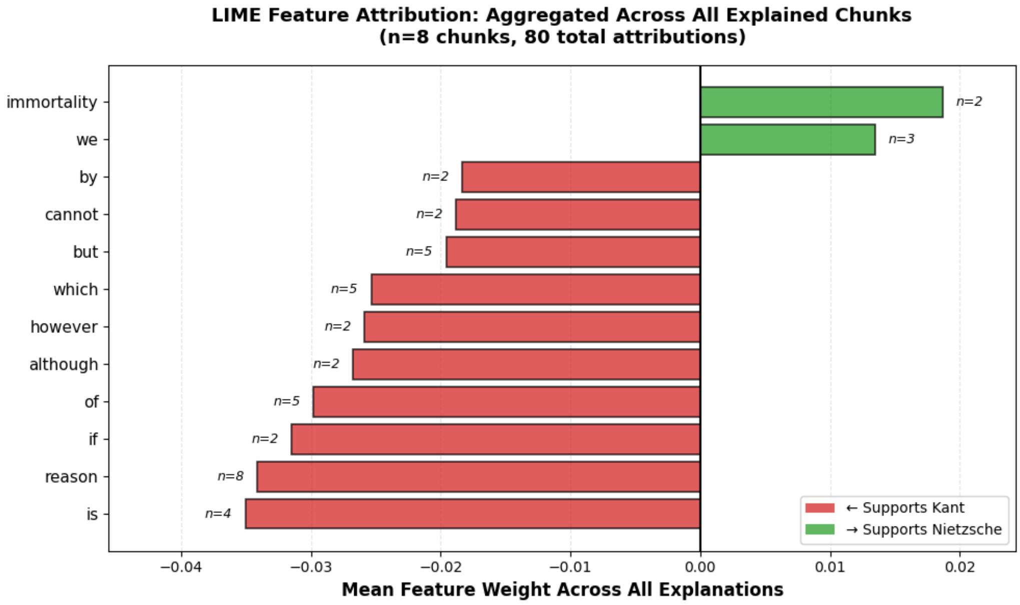

Figure 2: Aggregated LIME feature attributions across eight representative test chunks. Negative weights indicate features supporting Kant, while positive weights indicate features supporting Nietzsche. Bar lengths correspond to mean feature weight across explanations, and annotations show the number of occurrences of each token.

5.1 Feature patterns

Figure 2 summarizes the average contribution of the most influential tokens across all explanations. Features supporting Kant are predominantly function words and abstract philosophical terms. Tokens such as *reason* (mean weight -0.034, $n = 8$), *is* (-0.035, $n = 4$), *of* (-0.030, $n = 5$), *if* (-0.031, $n = 2$), and *although* (-0.027, $n = 2$) consistently push predictions toward Kant. These markers reflect Kant's clause-dense, highly structured argumentative style, characterized by frequent logical connectors and abstract terminology.

In contrast, features supporting Nietzsche are fewer and more content-specific. The strongest positive indicators are *immortality* (+0.019, $n = 2$) and the inclusive pronoun *we* (+0.013, $n = 3$), consistent with Nietzsche's rhetorical and often exhortative mode of address. The asymmetry in feature counts reflects Nietzsche's broader stylistic variability rather than weaker signal strength.

5.2 Linguistic categories

Aggregating feature attributions by linguistic category provides a complementary view of the same pattern. Table 3 shows that philosophical terms and function words contribute most strongly to Kant classifications, while Nietzsche-supporting features are fewer and more evenly distributed across categories.

Table 3: LIME feature attributions by linguistic category. Mean weights are signed; negative values indicate support for Kant and positive values indicate support for Nietzsche.

Category	**Kant-supporting**	**Nietzsche-supporting**	**Mean weight**	**Count**
Philosophical terms	8	0	−0.034	8
Function words	10	3	−0.020	13
Content words	27	7	−0.016	34
Short words	11	7	−0.010	18
Long words	4	3	−0.007	7

The prominence of function words is especially informative, as such tokens are largely topic-independent and have long been recognized as stable markers of authorial style in computational stylometry [3, 10].

5.3 Interpretation

The LIME analysis indicates that the classifier relies on structurally meaningful stylistic cues, logical connectors, syntactic markers, and abstract vocabulary, rather than isolated topic words. The features identified by LIME align closely with established descriptions of Kant's systematic prose and Nietzsche's rhetorical style, complementing the semantic-control evidence in H2 and supporting the conclusion that the model encodes genuine authorial style.

6 Limitations

Several limitations qualify the interpretation of our findings. First, although generalization is evaluated using held-out works, inference is conducted at the chunk level. Adjacent chunks drawn from the same book are not independent, so confidence intervals that treat chunks as i.i.d. observations are likely optimistic. More fundamentally, the data have a hierarchical structure: chunks are nested within books, which are nested within authors. The current modeling approach treats all chunks as exchangeable, implicitly assuming that the model can infer this hierarchy on its own. While transformer models may capture some higher-level regularities implicitly, they are not explicitly constrained to respect this nesting. As a result, the classifier may partially rely on work-level or translation-specific regularities rather than purely author-level style. Addressing this issue would require hierarchical or multilevel approaches—such as document-level aggregation, mixed-effects evaluation, or hierarchical neural architectures—which remain an active area of research in machine learning and computational linguistics.

Second, the corpus is drawn from English translations hosted on Project Gutenberg. As a result, some stylistic signals may reflect translator or editor conventions, residual paratext (e.g., prefaces), or formatting artifacts rather than the authors' original German prose. While we removed standard Gutenberg boilerplate, translation effects cannot be fully eliminated.

Third, token-based chunking (125 tokens) can cut across sentence or paragraph boundaries, potentially fragmenting rhetorical units and introducing noise. This may affect both topic modeling and interpretability analyses by mixing stylistically heterogeneous material within a single input.

Fourth, semantic controls are necessarily approximate. BERTopic reveals strong topical separation between authors but yields few balanced topics, resulting in a small topic-controlled test set. Embedding-based similarity provides a broader control but depends on the choice of sentence-transformer model and similarity threshold; at very high thresholds, evaluation becomes statistically unstable due to small sample sizes.

Finally, interpretability analysis via LIME is local and based on a limited number of representative examples. While aggregated token attributions reveal coherent stylistic patterns, they should be interpreted as suggestive evidence rather than a complete characterization of the model's global decision rules. Future work could extend interpretability to larger samples or complementary methods such as attention analysis or probing classifiers.

7 Conclusions

This study demonstrates that transformer-based models can reliably encode authorial style in philosophically dense texts, even when semantic content overlaps substantially. Using Kant and Nietzsche as a stringent test case, we show that a fine-tuned DistilBERT model generalizes across unseen works and maintains high accuracy under explicit semantic controls.

Beyond performance, the combined use of topic diagnostics, embedding-based similarity, and local interpretability provides converging evidence that classification decisions rely on structural and stylistic signals rather than topical vocabulary alone. In particular, the persistence of high accuracy under semantic overlap and the prominence of function words and logical connectors in LIME explanations align closely with established stylometric theory.

More broadly, these results contribute to ongoing debates about what pretrained language models learn when fine-tuned for style-sensitive tasks. They suggest that transformers capture higher-order linguistic regularities that support principled authorship attribution, while also highlighting the importance of careful evaluation design and interpretability when deploying such models in humanities and forensic contexts.

References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 4171–4186.

[2] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.

[3] Burrows, J. (2002). *"Delta": A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing*, 17(3), 267–287.

[4] Stamatatos, E. (2009). *A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv preprint arXiv:2106.09685.

[6] Kaufmann, W. (1974). *Nietzsche: Philosopher, psychologist, antichrist* (4th ed.). Princeton University Press.

[7] Guyer, P. (2006). *Kant's conception of fine art*. In P. Guyer (Ed.), *The Cambridge companion to Kant and modern philosophy* (pp. 214–246). Cambridge University Press.

[8] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv preprint arXiv:2203.05794.

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

[10] Koppel, M., Schler, J., & Argamon, S. (2009). *Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Acknowledgments

AI-assisted tools were used to support LaTeX formatting and Python code debugging during manuscript preparation. All modeling decisions, analyses, and interpretations were developed and verified by the author.