Journal Title Here, 2022, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Stylometric Detection of AI-Generated Texts: Evidence from Human and Machine-Written Essays

Jingqi He,[1] Rongzhi Chen,[1] Shizhao Xiong[1] and Gordon J. Ross[1,*]

[1]School of Mathematics, The University of Edinburgh, The King's Buildings, EH9 3FD, Edinburgh, United Kingdom
[*]Corresponding author: gordon.ross@ed.ac.uk
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

The rise of large language models (LLMs) such as ChatGPT has intensified debates about authorship, originality, and integrity in academic and creative writing. Distinguishing between human- and AI-generated texts is therefore not only a technical task but also a pressing concern for the digital humanities, where style, creativity, and attribution remain central. In this study, we adapt stylometric techniques to this new setting, introducing a dataset of paired human- and AI-authored essays across 110 subject areas. We evaluate three established classifiers, analysing how essay length, training data size, and topical variation influence performance, and whether human writing is better represented as a single category or as distinct authorial voices. Our findings show that while AI-generated texts exhibit striking stylistic uniformity, human writing is marked by variability and individuality. This contrast demonstrates both the continuing effectiveness of stylometry for AI detection and its wider relevance for authorship, originality, and voice in the age of generative AI.

Key words: Authorship attribution, Function words, Plagiarism detection, Stylometry

Introduction

Stylometry is the quantitative analysis of linguistic style, long recognised in the digital humanities as a bridge between computational techniques and interpretive questions about authorship (Evert et al., 2017). Classical stylometry focuses on measurable textual features such as word frequency distributions, vocabulary richness, sentence length, and the use of function words. These features serve as a distinctive stylistic fingerprint of an author (Rudman, 2006; Eder et al., 2016).

Although stylometry originated in literary and philological research, its role has broadened in digital humanities to include plagiarism detection, historical authorship disputes, and the study of stylistic development across periods and genres. What unites these applications is the ability to operationalise 'style' in measurable ways, linking computational analysis to long-standing scholarly questions about authorship and individuality.

The emergence of large language models (LLMs) such as ChatGPT has introduced a new domain where these techniques are urgently needed. LLMs are increasingly used for automated content generation(Lu et al., 2023), code development (Biswas, 2023), and academic writing (Imran and Almussharaf, 2023). While these tools have clear benefits, they also create challenges for originality, attribution, and academic integrity. Detecting AI-generated writing has therefore become an important problem for educators, publishers, and researchers.

Existing work on detecting AI-generated text often uses black-box machine learning classifiers trained on large neural embeddings (Fabien et al., 2020; Barlas and Stamatatos, 2020). While these approaches can be effective, they offer limited interpretability and are not always accessible in a humanities research setting. Stylometry, by contrast, provides transparent, lightweight methods grounded in a long tradition of digital humanities research, making it a natural choice for our study.

Recent studies have shown that stylometric techniques can indeed separate human- and AI-generated texts Yeadon et al. (2023); Berriche and Larabi-Marie-Sainte (2024); Zaitsu and Jin (2023); Przystalski et al. (2026). These works highlight consistent stylistic patterns in AI writing, but also leave open key questions about how data characteristics and modelling choices influence results.

Accordingly, this study investigates not only how well stylometric methods perform in distinguishing human from AI writing, but also what these comparisons reveal about the notion of "authorship" in digital culture. If machine-generated texts display consistent stylistic patterns, while human writing is marked by diversity and idiosyncrasy, then authorship attribution becomes a way of examining the contrast between uniformity in AI writing and variability in human writing."

This study addresses four such questions: (i) whether human writing is best represented as a single category or as distinct authorial voices; (ii) how essay length influences classifier performance; (iii) how performance shifts with limited training data; and (iv) whether classifiers generalise across topics or rely on subject-specific cues. We evaluate three established stylometric methods—Burrows' Delta, Random Forest, and Support Vector Machines—on a novel dataset of 4,346 paired human- and AI-authored essays. By clarifying the conditions under which stylometry performs best, our work contributes both practical insights into AI text detection and evidence of the continuing value of interpretable, style-based methods within digital humanities research.

Background and Rationale

ChatGPT

ChatGPT is an AI model developed to generate human-like text, including essays, articles, and news reports. The model is built on generative pre-trained transformer (GPT) technology, which forms part of a broader class of large language models (LLMs) specifically designed for natural language processing (NLP)(OpenAI, 2022). Although our experiments employ ChatGPT outputs as the AI-generated texts, the methods are equally applicable to other LLMs, such as Claude (Anthropic), Gemini (Google DeepMind), and Grok (xAI) if the training data is adjusted accordingly.

Function Words

Function words are among the most established features in stylometric analysis, due to their stability across an author's work and their unconscious use in writing (Rudman, 2006; Eder et al., 2016; Hossain et al., 2017). Unlike content words, function words contribute little semantic meaning but are crucial for grammatical cohesion. Their extremely high frequency makes them especially useful for computational comparison: fewer than 0.04% of English vocabulary accounts for over half of all words used in everyday discourse (Chung and Pennebaker, 2007). In this study, we use a curated set of 70 function words, adapted from Mosteller and Wallace's landmark analysis of The Federalist Papers (Mosteller and Wallace, 1963), as listed in Table 1. In this study, each essay is transformed into a 71-dimensional vector, where the first 70 dimensions represent the proportion of each function word in the text, with the final dimension representing the proportion of non-function words in the text.

a, all, also, an, and, any, are, as, at, be, been, but, by, can, do, down, even, every, for, from, had, has, have, her, his, if, in, into, is, it, its, may, more, must, my, no, not, now, of, on, one, only, or, our, shall, should, so, some, such, than, that, the, their, then, there, things, this, to, up, upon, was, were, what, when, which, who, will, with, would, your

Table 1. The 70 most frequently used function words in this analysis.

We note that other feature representations – such as character n-grams or neural embeddings derived from pre-trained language models like Fabien et al. (2020) and Tyo et al. (2022) – have been shown to perform well in authorship attribution tasks, particularly in PAN-style benchmarks. However, they are often difficult to interpret, and their connection to traditional stylistic analysis is less direct. Our aim here is not to pursue state-of-the-art accuracy, but to test whether function words, as a transparent and well-established stylometric feature set, can reliably separate human and AI-generated texts. Future work could extend this comparison to embeddings or hybrid approaches, but our contribution is to establish a clear baseline within the stylometric tradition.

Data and Visualisations

Our dataset comprises 4,346 essays, evenly split between 2,173 human-written and 2,173 ChatGPT-generated entries. The human-written corpus was drawn from the Aeon Essays Dataset (Acharya, 2024), comprising long-form essays originally published by Aeon Media. They cover 110 subject areas, ranging

from the sciences and engineering to the arts and humanities, including topics such as “Architecture”, “Genetics” and “Stories and Literature”. A detailed list of the analysed topics is provided in Appendix A: List of Topics Analysed.

We treat each essay as originating from a unique author. This decision reflects the diversity of writing styles in the dataset and allows more precise modelling of variation in human authorship. Aeon essays are particularly suitable for our purposes as they cover a wide topical range, are authored by many distinct writers, and exhibit diverse stylistic characteristics. This diversity provides a strong test case for distinguishing human writing from AI outputs.

ChatGPT-Generated Essays

For each human-written essay, we produced a corresponding ChatGPT-generated version by providing the essay’s title to the GPT-4o mini language model. The following prompt was used:

“Hi ChatGPT, I am going to give you a title for an essay, and I would like you to write a 1000-word essay on this subject. Please do not include anything in your response except for the essay. Also, your essay should not have section headings or a title. The title is:”

Using this instruction, a total of 2,173 essays were generated. While real-world use of AI often involves more varied prompts and post-editing, we adopted a consistent procedure to ensure comparability across topics and to isolate stylistic differences between human and AI texts.

Comparative Example of Human and ChatGPT Essays

To illustrate the differences in writing style between human-authored and AI-generated texts, we present the first 100 words from a human-written and an AI-generated essay on the same topic: “Stories and Literature”.

Human-Written Essay (Excerpt)

In the surreal aftermath of my suicide attempt and amid the haze of my own processing, my best friend visited me in the hospital with a (soft-bound and thus mental-patient-safe) copy of David Foster Wallace’s Infinite Jest under his arm. It was the spring of 2021. A couple months earlier, I had slipped in a tub, suffered a concussion, and triggered my first episode of major depression, and those had been the most difficult months of my life. Though a lifelong ’striver’ and ’high achiever’, nothing I’ve ever done was harder than waging that war against myself while catatonic...

ChatGPT-Generated Essay (Excerpt)

Bereft and suicidal, I lay on my sofa, the weight of despair pressing down on me like an insurmountable force. Each day felt like an echo of the last, a numbing repetition that drained the vibrancy from the world around me. I found myself entrapped in a cycle of lethargy and hopelessness, wandering through the dark corridors of my mind. It was during these bleak hours that I stumbled upon something surprising: a novel that would become my lifeline. David Foster Wallace’s Infinite Jest emerged as a beacon in my darkness, its complexity and raw honesty tethering me to thoughts of life...

Data Visualisations

To visualise relationships between texts in this feature space, we employ Multi-Dimensional Scaling (MDS), a dimensionality-reduction technique commonly used in stylometry (Ross, 2019) that projects high-dimensional observations into two dimensions while preserving the original distance structure as closely as possible.

Figure 1 shows an MDS plot for our data where Each red point represents a ChatGPT-generated essay, while each blue point corresponds to a human-written essay. This plot shows a clear separation between the ChatGPT and human clusters. The ChatGPT cluster is compact and tightly grouped, reflecting the model’s ability to produce uniform text across all generated essays. In contrast, the human cluster is more dispersed, indicating greater variability in writing styles. This variability likely reflects differences in individual authors’ linguistic choices, levels of expertise, and personal interpretations of essay topics.

A small number of red points overlap with the blue cluster, suggesting that some ChatGPT-generated essays closely resemble human writing. This raises the question of whether to treat all human essays as a single representation of “human” writing or to treat each essay as written by a distinct author. Given the observed overlaps, the latter approach appears more appropriate.

Treating all human essays as a single “blob” assumes that human writing is relatively homogeneous. However, the evident dispersion within the blue cluster suggests this assumption is inaccurate. The diversity of human writing styles could compromise the accuracy of classification models. Treating each human essay as authored by a separate individual acknowledges this variability, potentially leading to more precise classifications by reflecting the differences among various human authors.

Methodology

Introduction to Stylometry Methods

We now turn our attention to conducting a formal authorship attribution study. In this phase of our research, we will employ three standard methods of authorship attribution that are widely recognised and commonly utilised in the stylometry literature. These methods—Burrows’ delta (Delta), Random Forest (RF), and Support Vector Machines (SVMs)—each present unique advantages and are frequently applied to analyse and classify authorship based on the linguistic and stylistic features of texts. By comparing these methodologies, we aim to gain valuable insights into their respective strengths and limitations

and identify the most suitable technique for our authorship attribution task.

Burrows' Delta

Burrows' Delta is a widely used method in classic stylometry (Hoover, 2004; Argamon, 2008; Evert et al., 2017), first introduced by John F. Burrows as a simple yet effective measure of stylistic difference based on word proportion patterns (Burrows, 2002).

In our study, each essay is represented as a vector of function-word proportions. We then measure the stylistic distance between texts using Euclidean distance, a common variant of Delta that emphasises larger deviations in word use. Classification is performed by assigning a text to the category with the closest average profile.

The appeal of Delta lies in its interpretability: unlike more complex machine-learning models, it provides a transparent measure of stylistic similarity grounded in a long-standing humanistic tradition. At the same time, its simplicity makes it sensitive to data conditions, as we explore in our experiments.

Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are a supervised machine learning method frequently used in stylometric authorship attribution (Jockers and Witten, 2010; Evert et al., 2017). The general principle is to construct a hyperplane that best separates two classes of data points. In practice, many datasets are not linearly separable. To address this, SVMs employ kernel functions such as the radial basis function (RBF) kernel, which allows the model to capture subtle stylistic differences that cannot be represented with a straight line.

Random Forest (RF)

Breiman (2001) introduced Random Forest as a powerful ensemble learning method, now widely used for both classification and regression tasks. It operates by constructing multiple decision trees during training and combining their predictions to enhance accuracy and reduce over-fitting. In stylometric analysis, Random Forest has been applied to authorship attribution by leveraging high-dimensional linguistic features such as function word proportions, part-of-speech tags, and character n-grams (Jockers and Witten, 2010; Stamatatos, 2009). Its ability to manage redundant or irrelevant features makes it well-suited to capturing the subtle stylistic cues that differentiate authors.

Introduction to Evaluation Metrics

In machine learning and data science, evaluating a model's performance is crucial for understanding its effectiveness. In our research, we use four standard metrics (accuracy, precision, recall, and the F1 score) to assess classifier performance. Each captures a different aspect of predictive ability, and together they provide a balanced view of how well models distinguish between human- and ChatGPT-generated texts.

Because our dataset contains an equal number of human and AI essays, accuracy serves as a useful overall measure. However, accuracy alone can be misleading, so we also report precision (the proportion of texts identified as AI that truly are AI), recall (the proportion of AI texts correctly identified), and the F1 score, which balances precision and recall. These metrics allow us to evaluate not just whether classifiers are correct on average, but how they manage the trade-offs between false positives (misclassifying human texts as AI) and false negatives (failing to detect AI texts).

Results and Discussion

We conducted a series of experiments to assess how well the three classifiers – Burrows' Delta (Delta), Support Vector Machines (SVMs), and Random Forest (RF) – could human-written from ChatGPT-generated essays. All classifiers were implemented with standard parameter settings, tuned via cross-validation where appropriate. Unless otherwise noted, classifiers were trained using a leave-one-out scheme so that each essay was tested on models built from all remaining data.

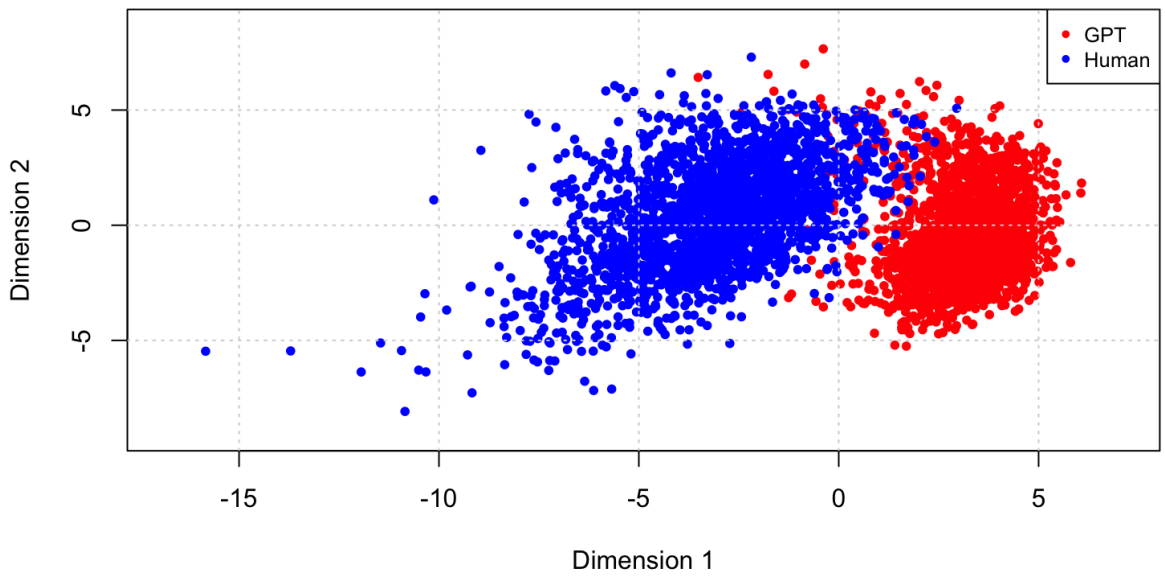


Fig. 1. Multi-dimensional scaling plot visualising stylistic similarity among essays authored by humans and generated by ChatGPT.

Study 1: Aggregated vs. Individual Treatment of Human-Written Essays in Authorship Attribution

As can be seen in Figure 1, human writing is inherently diverse, while ChatGPT-generated essays tend to be more uniform We therefore compare two representations of the human class: (i) an aggregated representation that treats all human essays as if produced by a single generic author, and (ii) an individual-author representation that treats each human essay as written by a distinct author. This comparison aims to determine whether accounting for individual variability in human writing improves classification performance.

Human Essays as a Single Aggregated Class

We first assume that all human-written essays share a common authorship style, effectively collapsing them into a single class. We hence have a binary classification task, with two classes: "Human" and "AI". Under this setting, the classifiers achieved moderate performance. As shown in Figure 2, Random Forest performed particularly well, while Delta lagged behind. Evaluation metrics such as recall, precision, and F1 score aligned closely with accuracy.

Next, we treat each human essay as if it were written by a unique individual. Although the dataset contains approximately 1,500 actual authors, this method assumes maximal stylistic variability by assigning a distinct author label to each human-written essay. It is important to note that, accuracy is still defined with respect to the binary task of distinguishing human versus ChatGPT essays. Consequently, if an essay written by one human author is misclassified as belonging to a different human author, this still counts as a correct prediction

Figure 2 shows this more granular approach yields substantial improvements across all classifiers, clearly demonstrating that treating human-written essays as authored by distinct individuals yields superior performance in distinguishing them from ChatGPT-generated texts. This approach better reflects the inherent stylistic variability of human writing, enabling classifiers to capture meaningful differences and reduce errors.

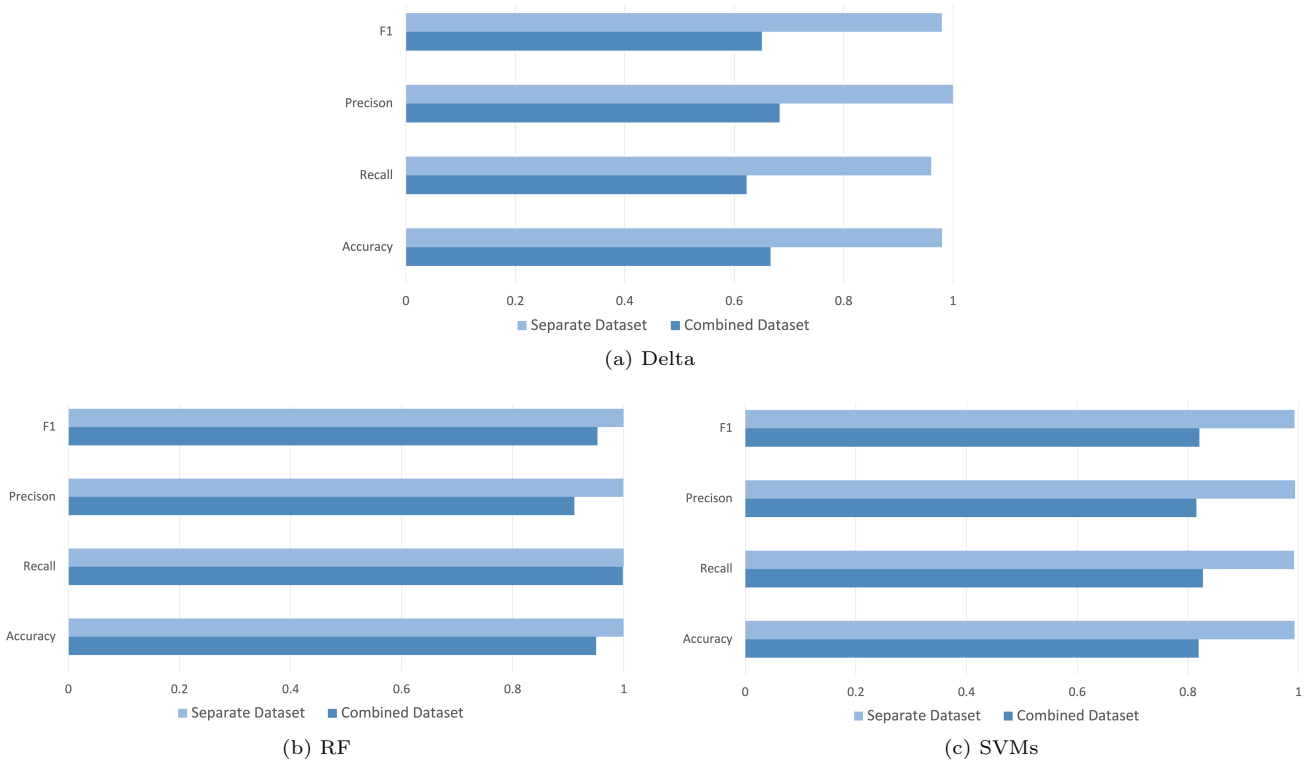
This contrast between the tight cluster of AI essays and the dispersed cluster of human essays highlights more than a technical classification problem. It reflects a deeper distinction: human authorship is bound up with individuality and variation, while AI writing projects a homogenised style. Stylometry thus makes visible, in quantitative form, a long-standing humanistic concern — the link between authorship and personal voice.

Consequently, we adopt the individual-author approach in all subsequent studies to maximise classification effectiveness. Nevertheless, in computationally constrained scenarios, employing the aggregated "blob" representation with Random Forest still offers a practical and accurate alternative.

Study 2: Impact of Essay Length on Classification Performance

Intuitively, we would expect essay length to have a direct impact on classification accuracy. To assess this, we created shortened versions of each essay by taking the first 200 words as a continuous block from the original text. This method preserves the natural flow and context that occur in authentic writing. We then compared classifier performance between these shortened essays and the full-length versions.

Figure 3 show that essay length strongly impacts performance, with all classifiers experiencing declines across the four metrics when moving from full texts to 200-word excerpts. Recall consistently showed the smallest decline, indicating that even with shorter texts, classifiers could still identify ChatGPT



(a) Delta

(b) RF

(c) SVMs

Fig. 2. Results from Study 1: accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines using combined and separate datasets.

essays with relative effectiveness, though at the expense of precision and overall balance. Precision experienced the largest decrease, suggesting a greater likelihood of misclassifying human-written essays as ChatGPT-generated in the shorter-text condition. This indicates that while the classifiers can capture the characteristics of ChatGPT-generated essays with shorter texts, they may also mistakenly identify features in human essays due to the limited stylistic data available.

Delta and SVMs showed modest performance drops, suggesting robustness to moderate reductions in text length. In contrast, Random Forest exhibited the most pronounced performance drop: while their recall showed almost no change across lengths, accuracy and precision fell to about $\frac{1}{2}$ and the F1 score dropped to about $\frac{2}{3}$ in the 200-word condition. This pattern indicates a strong bias towards predicting essays as ChatGPT-generated when text length is reduced, leading to many false positives. As such, Random Forest is not reliable as a stand-alone classifier for short texts; if used for its high recall, it should be paired with a more accurate classifier to minimise errors.

SVMs consistently outperformed the other classifiers at both full and reduced lengths. This makes SVMs the most resilient option for datasets with varying or limited text lengths, including condensed excerpts.

In summary, shorter essays generally reduced classifier performance, particularly in terms of precision. SVMs delivered the most stable and balanced results, while Random Forest's high-recall but low-precision profile suggests its potential utility as a pre-classification filter. Burrows' Delta also proved relatively resilient, though slightly less robust than SVMs in short-text scenarios.

Study 3: Impact of Training Data Size on Classifier Performance

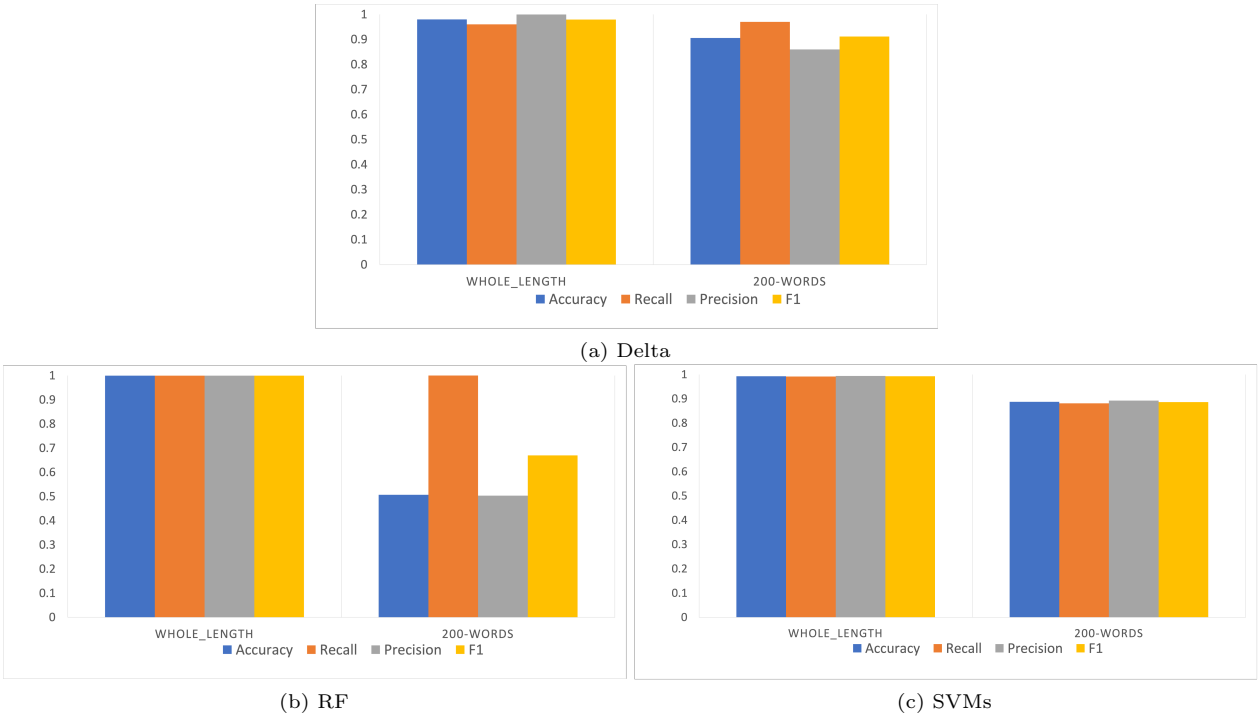
Training data availability is not a constraint, as ChatGPT texts can be generated in unlimited amounts and human essays are readily accessible. The purpose is to show the risks arising from small training sets, and to underline why such conditions are not recommended.

Therefore, in this study, we examined how reducing the number of training essays while keeping their original length impacts the performance of these three classifiers. For each topic, we randomly reduced the number of available training essays to 50%, 20%, and 10% of the original count. For example, if the original training set for a topic contained 100 essays, only 50, 20, or 10 essays were used in these reduced-data scenarios.

Figure 4 shows how classifier performance varies with training set size. When trained on the full dataset, all classifiers performed well, with Random Forest achieving near-perfect results. The detailed numerical outcomes for this condition are reported in Table 3, which complement the figure by showing the exact accuracy, precision, recall, and F1 scores.

When the training data was reduced to 50%, performance decreased slightly for most classifiers. Random Forest remained the most robust, achieving accuracy and F1 score of 99.82%, showcasing stable performance even with fewer training examples. SVMs exhibited a more noticeable decline, with accuracy dropping to 81.32% and F1 score to 81.48%, highlighting its sensitivity to smaller datasets. Interestingly, Burrows' Delta improved at the 50% training size (accuracy 99.54%). This improvement is not visible in Figure 4 but is evident when comparing the numerical results in Tables 3 and Table 9.

At 20% training data, the gap between classifiers widened. Random Forest continued to outperform the others, maintaining



(a) Delta

(b) RF

(c) SVMs

Fig. 3. Results from Study 2: accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines by essay length.

accuracy and F1 score around 99.93%. Visually, its lines appear almost flat, but the reported values confirm a small decline from 99.82% at 50% to 99.31% at 20%. Burrows' Delta experienced a moderate drop (F1 score = 97.83%). SVMs declined sharply to 79.08% accuracy and 78.87% F1 score, and while the downward slope looks modest in the plot, the numerical difference highlights a clear loss of stability as data availability shrinks, indicating growing vulnerability to reduce data availability.

With only 10% of the original training essays, performance declined for all classifiers. Random Forest still remained the most resilient, with accuracy and F1 around 98.4%. Burrows' Delta also showed reasonable robustness (accuracy 90.58%, F1 90.84%). SVMs suffered the largest drop, with both accuracy and F1 at around 75%.

Overall, these results show that Random Forest delivers the strongest performance across all training set sizes. Burrows' Delta also adapts relatively well, while SVMs are the most sensitive to small training data. The improvement of Burrows' Delta at 50% training size likely reflects a reduction in overfitting, as removing part of the training data may eliminate noisy or atypical samples and lead to cleaner frequency distributions. However, further reductions to 20% and 10% deprive the method of sufficient information to capture stable stylistic patterns, causing performance to decline. This suggests that Delta is sensitive not only to the amount of training data but also to its representativeness, performing best when the dataset strikes a balance between size and noise.

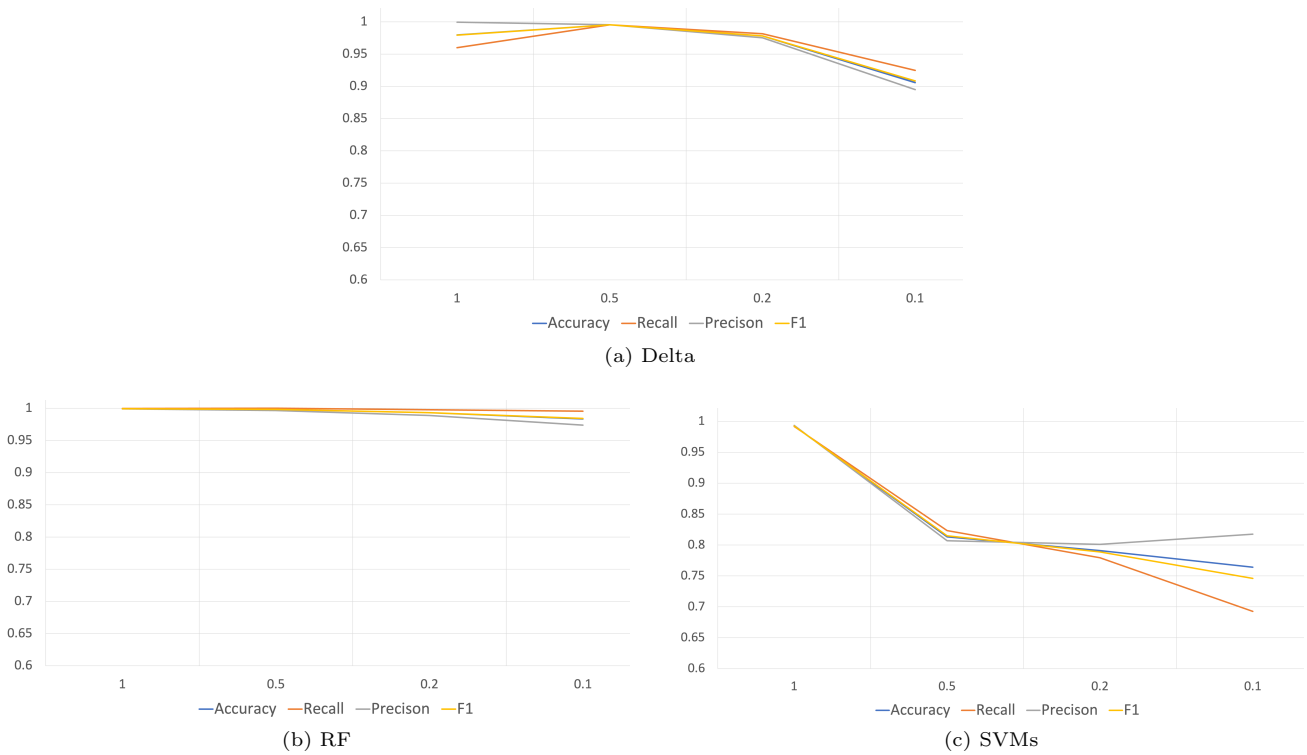
Study 4: Assessing the Influence of Topic Relevance on the Performance of Classifiers

This experiment investigates whether high classification accuracy depends on training with essays from the same topic as the test data, or whether models can generalise to unseen topics. In other words, we investigate whether topic relevance is a necessary condition for achieving strong performance. We therefore compare three scenarios: (i) training and testing on the same topic, which evaluates how well classifiers exploit topic-specific stylistic patterns; (ii) training on a single unrelated topic, which tests whether classifiers fail when the training data bears no topical relation to the test essays; and (iii) training on all topics except the test topic, which assesses whether classifiers can generalise from broad stylistic variation without direct exposure to the test topic. By contrasting these cases, we can evaluate the extent to which classifiers depend on topic-specific stylistic features versus topic-independent ones.

Scenario 1: Using Essays From the Same Topic

Scenario 1 restricted both training and testing to the same topic, creating a stricter, topic-specific setting.

As shown in Figure 5, Random Forest demonstrated outstanding performance, with all four metrics consistently around 99%, indicating high reliability when topic-matched training data were available. Burrows' Delta also performed well overall (93.5% accuracy), but its results varied sharply by topic; for "Making," where only two essays per class were available, accuracy, precision, recall, and F1 all fell to around 50%, revealing sensitivity to extremely small samples. By contrast, SVMs were the weakest on average (about 70% across metrics). A notable exception was "Teaching and Learning," where SVMs



(a) Delta

(b) RF

(c) SVMs

Fig. 4. Results from Study 3: accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines using 10%, 20%, 50%, and 100% of the training dataset.

achieved perfect scores, potentially due to this topic exhibiting distinctive and consistent stylistic patterns.

Overall, under same-topic training, Random Forest was robust, Burrows' Delta was competitive but data-sensitive, and SVMs generally struggled except on highly distinctive topics.

Scenario 2: Using Essays From a Single Unrelated Topic

In this scenario, classifiers were trained using essays from a single randomly chosen topic that differed from the test topic.

As shown in Figure 6, Random Forest again outperformed the others, achieving near-perfect metrics with accuracy, precision, recall, and F1 scores all around 99%. SVMs demonstrated substantial improvement compared to Scenario 1, with all metrics exceeding 80%, indicating better generalisation across unrelated topics than when restricted to topic-specific training data. In contrast, Burrows' Delta achieved only moderate results (accuracy 65.74%, recall 56.72%, precision 56.27%, and F1 51.74%), underscoring its reliance on topic-relevant training data. Due to the randomness in topic selection, the performance of individual topics in this scenario was not analysed, as this would provide limited additional insights.

Scenario 3: Using Essays From All Topics Except the Test Topic

In this scenario, classifiers were trained on essays from all topics except the one being tested.

As shown in Figure 7, both Random Forest and SVMs achieved consistently high performance, with accuracy, precision, recall, and F1 scores all close to 99%, demonstrating strong generalisation across diverse training data. Burrows' Delta also performed well, though slightly lower. Random Forest exhibited near-perfect classification for most topics, with only "Biography and Memoir" and "History" recording slightly lower accuracies of 97% and 99%, respectively.

Cross-Scenario Performance Analysis

Analysing the performance of each classifier across the three scenarios reveals several patterns, shown in Figure 8 .

SVMs performed poorly in Scenario 1 but achieved strong results in Scenarios 2 and 3. This outcome may be explained by the way SVMs construct their decision boundaries. When the training essays all come from a single topic, topic-specific vocabulary and phrasing dominate the feature space, masking the stylistic signals that distinguish ChatGPT from human writing. As a result, the support vectors chosen by the model reflect mainly topic-related variation, which leads to overfitting and poor generalisation. In contrast, when trained on essays from diverse topics, the topic-specific noise is diluted and ChatGPT's consistent stylistic patterns become more prominent, enabling the model to classify more effectively.

In contrast, Random Forest maintained consistently high performance across all three scenarios, demonstrating its

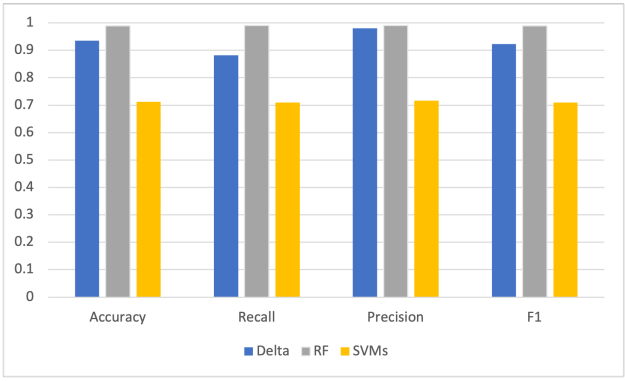


Fig. 5. Results from study 4: average accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines in Scenario 1.

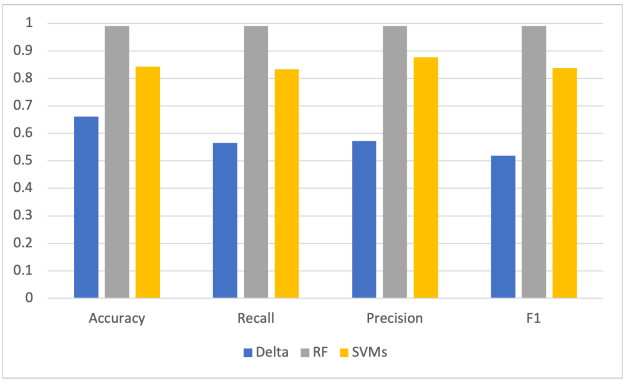


Fig. 6. Results from study 4: average accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines in Scenario 2.

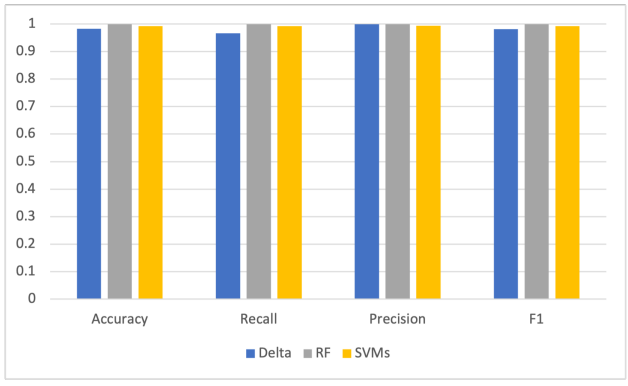


Fig. 7. Results from study 4: average accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines in Scenario 3.

robustness in identifying ChatGPT-generated essays, regardless of the topic or size of the training data.

Burrows' delta performed poorly in Scenario 2 but achieved good results in Scenarios 1 and 3. This indicates that its performance is heavily influenced by the training data size and relevance to the test topic. For these classifiers, having essays from the same topic is critical when data availability is limited. When same-topic training data is unavailable, including more essays from diverse topics can partially mitigate this limitation.

The influence of topic further underscores that authorship is not purely a matter of invariant linguistic features. Human writers often express individuality through topic-specific vocabularies, while AI writing displays a cross-topic sameness. In this sense, the very success of stylometry in detecting AI rests on a paradox: the more "general" the text appears, the less it resembles human authorship, which is always situated and variable.

Conclusion

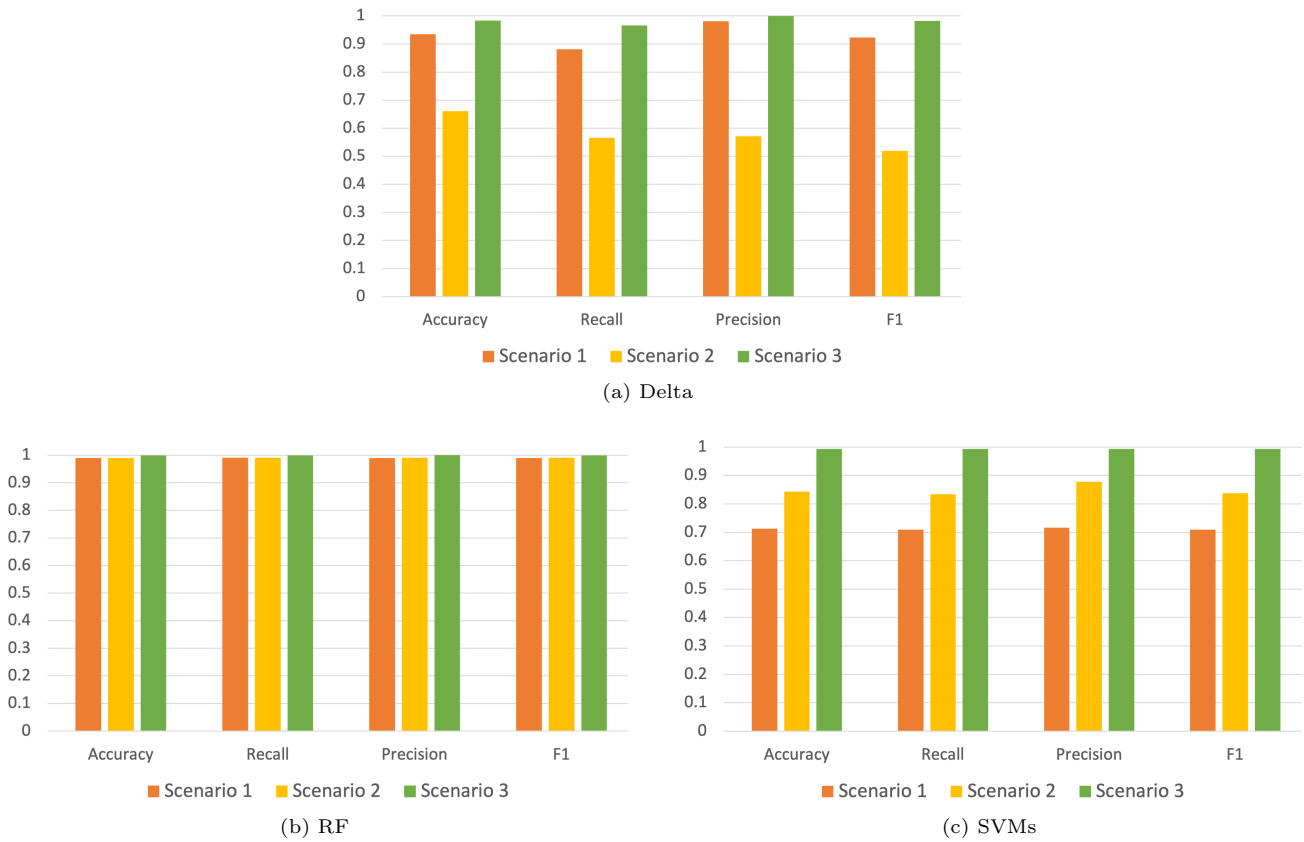
This experiment shows that topic relevance materially shapes classifier performance. Random Forest was consistently strongest across all scenarios, making it the safest default and a reliable option when no topic-matched training data are available. In contrast, SVMs performed poorly with same-topic training but improved markedly once training covered diverse topics, indicating reliance on topic-independent stylistic signals and vulnerability to topic-specific confounds. Meanwhile, Burrows' Delta was competitive with matched topics and broad coverage but fell to only moderate levels when trained on a single unrelated topic, suggesting sensitivity to both topic relevance and sample size.

From a practical perspective, when same-topic data are scarce, Random Forest should be regarded as the preferred method, with SVMs also suitable if the training corpus spans many topics. When limited same-topic data are available, Burrows' Delta can still be viable, provided that sufficient in-topic examples are included.

Limitations and Future Works

This study was designed as a controlled evaluation, and there are several directions in which the work can be extended. Our dataset, based on Aeon essays paired with ChatGPT outputs, ensured topical comparability and stylistic diversity, but future research could explore a wider range of genres such as student coursework, journalistic writing, or shorter informal texts. Hybrid cases, where human authors edit AI outputs, are especially relevant for practical applications and remain an important avenue for testing the robustness of stylometric classifiers.

We focused here on function words, a classic and interpretable feature set in stylometry. This choice allowed us to isolate clear stylistic signals and provide transparent explanations of classifier decisions. Future work could expand the feature space to include character n-grams, syntactic markers, or embeddings (Tyo et al., 2022), enabling comparison between interpretable baselines and more complex representations



(a) Delta

(b) RF

(c) SVMs

Fig. 8. Results from study 4: average accuracy, precision, recall, and F1 scores for Burrows' Delta, Random Forest, and Support Vector Machines across different scenarios.

Finally, we evaluated three established classifiers: Burrows' Delta, Random Forest, and SVMs. Together these provided a balance of classical stylometry and standard machine learning methods. Further research could extend the comparison to neural models (Fabien et al., 2020; Barlas and Stamatatos, 2020), investigate cross-lingual scenarios, and test outputs from other large language models beyond ChatGPT. Pursuing these directions would help to clarify the generalisability of our findings and map out the trade-offs between interpretability and raw predictive power.

Conclusion

This study evaluated the effectiveness of stylometric methods in distinguishing between human- and AI-generated essays across four dimensions: whether human texts are better treated as individual authors or a single class; the influence of essay length; the impact of training data size; and the ability of classifiers to generalise across topics. Using Burrows' Delta, Random Forest, and SVMs on a balanced dataset of 4,346 human and ChatGPT essays, we found that Random Forest performed most strongly overall, with accuracy improving as essay length and training size increased. Treating each human as a distinct author gave clearer separation than aggregating all human texts, and classifiers generalised reasonably well across topics.

Beyond these methodological results, our findings also highlight a more general pattern: AI-generated texts exhibited marked stylistic uniformity, while human writing was more variable and idiosyncratic. This contrast helps to explain why stylometric methods remain effective for AI detection, but it also has broader implications. For digital humanities, variability of style has long been central to how authorship is identified, contested, and valued. In this sense, stylometry remains relevant not only as a technical tool for classification, but also as a way of reflecting on the distinctiveness of human writing in the age of generative AI.

Future work should test the robustness of these findings on other genres and corpora, including mixed-authorship texts, student writing, and AI-generated outputs from models beyond ChatGPT. Expanding to multilingual contexts and incorporating additional feature sets would further clarify the scope and limits of stylometric approaches.

Appendix A: List of Topics Analysed

Addiction, Ageing and death, Animals and humans, Anthropology, Archaeology, Architecture, Art, Astronomy, Automation and robotics, Beauty and aesthetics, Bioethics, Biography and memoir, Biology, Biotechnology, Chemistry, Childhood and adolescence, Cities, Cognition and intelligence, Comparative Philosophy, Complexity, Computing and artificial intelligence, Consciousness and altered states, Cosmology, Cosmopolitanism, Dance and theatre, Death, Deep time, Demography and migration, Design and fashion, Earth science and climate, Ecology and environmental sciences, Economic history, Economics, Education, Engineering, Environmental history, Ethics, Evolution, Fairness and quality, Family life, Film and visual culture, Food and drink, Future of technology, Genetics, Global history, History, History of ideas, History of science, History of technology, Home, Human evolution, Human reproduction, Human rights and justice, Illness and disease, Information and communication, Knowledge, Language and linguistics, Life stages, Logic and probability, Love and friendship, Making, Mathematics, Meaning and the good life, Medicine, Mental health, Metaphysics, Mood and emotion, Music, Nations and empires, Nature and landscape, Neurodiversity, Neuroscience, Oceans and water, Palaeontology, Personality, Philosophy of language, Philosophy of mind, Philosophy of religion, Philosophy of science, Physics, Pleasure and pain, Political philosophy, Politics and government, Poverty and development, Progress and modernity, Public health, Quantum theory, Religion, Ritual and celebrations, Self-improvement, Sleep and dreams, Social psychology, Space exploration, Spirituality, Sports and games, Stories and literature, Subcultures, Teaching and learning, Technology and the self, The ancient world, The environment, The future, Thinkers and theories, Travel, Values and beliefs, Virtues and vices, War and peace, Wellbeing, Work, Psychiatry, Psychotherapy

Appendix B: Performance Results of Classifiers Across Studies

The tables present the numerical results for Studies 1 through 4 to further substantiate the respective findings.

Study 1: Aggregated vs. Individual Treatment of Human-Written Essays in Authorship Attribution

Model	Accuracy	Recall	Precision	F1
Delta	66.65%	62.26%	68.30%	65.05%
RF	95.03%	99.82%	91.11%	95.26%
SVMs	81.94%	82.70%	81.52%	82.08%

Table 2. Classifier results (Delta, RF, SVMs) on the combined dataset.

Model	Accuracy	Recall	Precision	F1
Delta	97.98%	96.00%	99.95%	97.93%
RF	99.93%	99.95%	99.91%	99.93%
SVMs	99.26%	99.17%	99.36%	99.26%

Table 3. Classifier results (Delta, RF, SVMs) on the separate dataset.

Study 2: Impact of Essay Length on Classification Performance

Model	Accuracy	Recall	Precision	F1
Whole Length	97.98%	96.00%	99.95%	97.93%
200-word	90.59%	97.00%	86.00%	91.16%

Table 4. Performance metrics for Delta on full-length essays and the first 200 words of each essay.

Model	Accuracy	Recall	Precision	F1
Whole Length	99.93%	99.95%	99.91%	99.93%
200-word	50.69%	100.00%	50.35%	66.98%

Table 5. Performance metrics for RF on full-length essays and the first 200 words of each essay.

Model	Accuracy	Recall	Precision	F1
Whole Length	99.26%	99.17%	99.35%	99.26%
200-word	88.72%	88.12%	89.22%	88.64%

Table 6. Performance metrics for SVMs on full-length essays and the first 200 words of each essay.

Study 3: Impact of Training Data Size on Classifier Performance

Model	Accuracy	Recall	Precision	F1
Delta	90.58%	92.49%	89.49%	90.84%
RF	98.36%	99.55%	97.39%	98.42%
SVMs	76.42%	69.26%	81.75%	74.60%

Table 7. Classifier Results (Delta, RF, SVMs) with 10% of each topic used as training data.

Model	Accuracy	Recall	Precision	F1
Delta	97.82%	98.16%	97.54%	97.83%
RF	99.31%	99.77%	98.87%	99.31%
SVMs	79.08%	77.93%	80.09%	78.87%

Table 8. Classifier Results (Delta, RF, SVMs) with 20% of each topic used as training data.

Model	Accuracy	Recall	Precision	F1
Delta	99.54%	99.54%	99.55%	99.54%
RF	99.82%	100.00%	99.64%	99.82%
SVMs	81.32%	82.34%	80.70%	81.48%

Table 9. Classifier Results (Delta, RF, SVMs) with 50% of each topic used as training data.

Study 4: Assessing the Influence of Topic Relevance on the Performance of Classifiers

Access the output at:
`https://github.com/xiongshizhao/stylometric-analysis-human-and-chatgpt-texts/tree/main/results/study4/tables`

Appendix C: Code and Data Repository

In this appendix, we provide a link to the GitHub repository that contains all the source code and datasets used in the development of this report. The repository includes the following:

- R scripts for data analysis and visualisation.
- Raw and processed data files.

Access the repository at:
`https://github.com/xiongshizhao/stylometric-analysis-human-and-chatgpt-texts`

References

M. Acharya. Aeon essays dataset, 2024. URL `https://www.kaggle.com/dsv/7371248`. Dataset.

S. Argamon. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, June 2008.

G. Barlas and E. Stamatatos. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, volume 583 of *IFIP Advances in Information and Communication Technology*, pages 255–266. Springer International Publishing, Cham, 2020.

L. Berriche and S. Larabi-Marie-Sainte. Unveiling chatgpt text using writing style. *Heliyon*, 10(12):e32976, 2024.

S. Biswas. Role of chatgpt in computer programming. *Mesopotamian Journal of Computer Science*, 2023:002, 2023.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

J. F. Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, Sept. 2002.

C. K. Chung and J. W. Pennebaker. The psychological functions of function words. In K. Fiedler, editor, *Social Communication*, pages 343–359. Psychology Press, 2007.

M. Eder, J. Rybicki, and M. Kestemont. Stylometry with r: A package for computational text analysis. *The R Journal*, 8 (1):107–121, 2016.

S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, and T. Vitt. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16, Dec. 2017.

M. Fabien, E. Villatoro-Tello, P. Motlicek, and S. Parida. BertAA: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137. NLPAI, 2020.

D. L. Hoover. Testing burrows's delta. *Literary and Linguistic Computing*, 19(4):453–475, Nov. 2004.

M. T. Hossain, M. M. Rahman, S. Ismail, and M. S. Islam. A stylometric analysis on bengali literature for authorship attribution. In *Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–5, 2017.

M. Imran and N. M. Almusharaf. Analyzing the role of chatgpt as a writing assistant at higher education level: a systematic review of the literature. *Contemporary Educational Technology*, 15(4):1–14, 2023.

M. L. Jockers and D. M. Witten. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2):215–223, June 2010.

G. Lu, S. B. Larcher, and T. Tran. Hybrid long document summarization using c2f-far and chatgpt: A practical study, 2023.

F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963. ISSN 0162-1459, 1537-274X.

OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022. Accessed: 10 November 2024.

K. Przystalski, J. K. Argasiński, I. Grabska-Gradzińska, and J. K. Ochab. Stylometry recognizes human and llm-generated texts in short samples. *Expert Systems with Applications*, 296:129001, Jan. 2026.

G. Ross. Tracking the evolution of literary style via dirichlet–multinomial change point regression. *Journal of the Royal Statistical Society: Series A*, 183(1):149–167, 2019. ISSN 2738-2184.

J. Rudman. Authorship attribution: Statistical and computational methods. In *Encyclopedia of Language and Linguistics*, pages 108–115. Elsevier, 2 edition, 2006.

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

J. Tyo, B. Dhingra, and Z. C. Lipton. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*, 2022.

W. Yeadon, O.-O. Inyang, A. Mizouri, A. Peach, and C. Testrow. The death of the short-form physics essay in the coming ai revolution. *Physics Education*, 58(3):035027, 2023.

W. Zaitsu and M. Jin. Distinguishing chatgpt (3.5, 4)-generated and human-written papers through japanese stylometric analysis. *PLOS One*, 18(8):e0288453, Aug. 2023.