Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 274 (2025) 1238–1243

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Modelling and Applied Simulation (MAS 2025), held within the 22nd International Multidisciplinary Modeling & Simulation Multiconference (I3M 2025)

Experimental Modeling of Writing Styles for Authorship Verification via Punctuation Analysis

Roberto Dillon[a]*, Marco Gotelli [b], Agostino Bruzzone[b]

[a]*James Cook University, 149 Sims Drive, Singapore*
[b]*University of Genoa, Via Opera Pia 15, Genova, Italy*

Abstract

Authorship attribution is a critical task in forensic linguistics, literary studies, and digital forensics, where determining the origin of a text can have significant implications. This paper presents an experimental stylometric tool developed in Python, designed to model writing styles and assist in authorship determination. The tool extracts nine quantitative features from input texts, including metrics such as average words per sentence and the frequency of specific punctuation marks (e.g., commas, semicolons). By comparing these features across texts, the system computes a probability score indicating the likelihood that two samples share the same author.

To evaluate the tool's effectiveness, we conducted experiments using short stories authored by Charles Dickens, Ernest Hemingway, and Edgar Allan Poe. The results demonstrate that the tool can reliably distinguish between authors and identify stylistic consistencies within an author's body of work. The approach leverages statistical analysis to provide an interpretable and reproducible framework for authorship attribution, complementing more complex machine learning models. This work contributes to the growing field of computational stylometry by offering a transparent, feature-driven method suitable for both forensic and academic applications. Future research will focus on expanding the feature set, testing on larger and more diverse corpora, and integrating the tool with advanced classification algorithms to further enhance accuracy and applicability

* Corresponding author. Tel.:+65-6709-3711
E-mail address: roberto.dillon@jcu.edu.au

10.1016/j.procs.2025.12.122

Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

Stylometry, i.e. the task of determining the authorship of a given text, has long been a subject of interest in fields such as forensic linguistics, literary analysis, and digital forensics. The ability to reliably model writing styles and then identify or verify the author of a document has significant implications, ranging from resolving questions of literary provenance to supporting legal investigations and combating online misinformation. As digital communication proliferates and anonymous or pseudonymous writing becomes more common, the demand for robust, interpretable, and language-flexible authorship analysis tools continues to grow.

Traditional approaches to authorship attribution have relied heavily on qualitative analysis, expert intuition, or the use of language-specific dictionaries and lexical resources [1]. These methods often involve examining an author's vocabulary, preferred phrases, or thematic content, which can be effective but are inherently limited by their dependence on the language and subject matter of the texts under analysis. More recently, computational stylometry has emerged as a powerful alternative, leveraging statistical and machine learning techniques to quantify stylistic features and automate the attribution process. Common computational approaches include the use of n-gram models, function word frequency analysis, and machine learning classifiers trained on large, labeled corpora [2][3]. While these methods have achieved notable success, they often require extensive preprocessing, language-specific resources, or large datasets to perform optimally.

The approach presented in this work wants to simplify basic stylometry analysis and propose a light, cross-platform tool that addresses several of these limitations by focusing exclusively on quantitative, language-agnostic features of writing style.

Rather than relying on dictionaries or semantic content, the proposed model, in fact, analyzes structural and syntactic characteristics that are largely independent of the specific language or vocabulary used. This design choice enhances the tool's flexibility and applicability across a wide range of Western languages, provided they share similar punctuation and sentence structure conventions.

By focusing on these features, the present paper investigates whether a simple tool can still manage to capture subtle stylistic patterns that are often consistent within an author's body of work but distinct across different authors.

If successful, this approach, while very straightforward, can offer several advantages:

- **Language Independence:** The model does not require language-specific dictionaries or lexical resources, making it adaptable to any Western language with similar orthographic conventions.
- **Transparency and Interpretability:** The use of simple, well-defined quantitative features allows for clear interpretation of results, which is particularly valuable in forensic and academic contexts.
- **Efficiency:** The model can be applied to relatively short texts and does not require large training datasets, enabling rapid analysis and practical deployment.

In summary, this work contributes a light yet robust feature-driven methodology for authorship attribution that complements existing approaches by prioritizing language flexibility, interpretability, and ease of use, suitable for verifying authorships of essays, emails and other short text forms.

The following sections will the methodology and feature selection, and present experimental results demonstrating the effectiveness of the proposed tool on classic literary texts.

2. Methodology

Since the pioneering effort discussed in [4], computer based stylometry split into two closely related areas: Author Attribution (AA) and Author Verification (AV), where the former aims at assigning a likely author to an unknown text while the latter aims at verifying whether a specific author has written a given text instead.

Different approaches have been proposed, with [5] being particularly successful for AA related problems. AV is a more general problem and has received relatively less attention, with fewer datasets made publicly available [6].

While deep learning approaches and experiments are getting increasingly common [7][8], most stylometric approaches are feature-based where author specific features are extracted from n-grams [9], i.e. particular sequences of characters or words that can be taken as signatures of specific writing styles and authors. These approaches, while effective, are language specific and usually require built-in dictionaries, making the corresponding tools larger and their analysis more resource intensive. In this exploratory work, instead, we aim at defining a very lightweight approach based on language independent features mostly relying on punctuation analysis.

Specifically, given a text of arbitrary length in input, the model extracts and analyzes nine quantitative features on a per-sentence basis:

- Word count
- Number of capitalized words
- Frequency of commas, colons, semicolons, parentheses, and quotation marks
- Number of words longer than four characters
- Number of words equal to or shorter than four characters

This feature set is then used to verify authorship via Random Forest included in a small dataset of short stories of various lengths from famous authors such as Charles Dickens, Edgar Allan Poe and Ernest Hemingway. Two short stories from each author were selected and compared with each other to determine the likelihood of the stories been written by the same author.

In particular the following short stories were selected (Table 1):

Table 1. English language short stories used in the dataset

Story	Number of Words	Author	ID
Berenice	3197	Edgar Allan Poe (1809-1849)	AP1
Eleonora	2407	Edgar Allan Poe (1809-1849)	AP2
Nobody’s Story	2211	Charles Dickens (1812-1870)	CD1
A Child’s Story	1699	Charles Dickens (1812-1870)	CD2
The Snows of Kilimanjaro	9165	Ernest Hemingway (1899-1961)	EH1
Old Man at the Bridge	754	Ernest Hemingway (1899-1961)	EH2

Two stories are considered as written by the same author if the model reports a similarity score (SS) higher than a threshold of 0.60. This final SS output is an “ensemble similarity score” between 0 1representing the probability that two texts share the same author. Unlike a standard Random Forest classifier, which outputs discrete classification predictions, this score combines statistical hypothesis testing with machine learning uncertainty to provide a more nuanced measure of stylistic similarity. In other words, while a traditional Random Forest model assigns each text sample a probability of belonging to Author A or Author B, authorship verification requires a different approach. For this reason, instead of relying solely on classification confidence, we compute the SS by taking into consideration the following components:

- A Machine Learning Uncertainty (MLU) score, computed from the original RF output probabilities, where the model evaluates how uncertain it is when distinguishing between the two texts by taking into account an entropy measure, defined in the python script as follows by using the numpy (np) library:

$$\text{entropy} = -\ \text{np.mean}\ (\text{probabilities} * \text{np.log}\ (\text{probabilities} + 1\text{e-}10)) \quad (1)$$

High entropy in predictions suggests the texts are stylistically similar, making classification difficult while low entropy indicates clear differences, implying distinct authors. This uncertainty is normalized into a “ML Similarity” score (MLS) ranging between 0, i.e. clearly different, to 1, i.e. indistinguishable.

- A Statistical Feature (SF) score, which is obtained by comparing independent t-tests for each of the considered nine stylistic features between the two texts. The proportion of features (PF) without statistically significant differences ($p \geq 0.005$) as well as the Cohen’s Distance (CD), a measure of the average effect size of differences, contributes to the score as follows:

$$SF = PF\,(1 - CD) \tag{2}$$

The final SS is then computed as a weighted ensemble sum of MLS and SF and can be interpreted as an overall probability of having the same authorship among the two texts:

$$SS = 0.4\,MLS + 0.6\,SF \tag{3}$$

3. Results

3.1. Random Forest

The Random Forest model was implemented in Python by using the sklearn library and related functions, and was defined by the following parameters (Table 2):

Table 2. Random Forest implementation via the sklearn python library. Specific values were chosen to avoid the risk of overfitting.

Random Forest Parameter	Value
Number of Estimators	50
Max Depth	5
Minimum Samples Split	10
Minimum Samples Leaf	5

The model works having both files in input, extracting feature vector from each sentence and then taking 70% of available data for training purposes, using cross validation, while leaving the rest for testing. As output, the model provides an overall Similarity Score (SS) ranging from 0.0 to 1.0 as defined in the previous section.

Results are shown in Table 3. We can set a similarity threshold of 0.6 to determine whether consider the two texts as belonging to the same author.

Table 3. Similarity Scores (SS) from Random Forest Analysis . In bold results indicating same authorship (i.e. SS > 0.60).

Story	AP1	AP2	CD1	CD2	EH1	EH2
AP1	**0.70**	**0.63**	0.58	0.56	0.24	0.29
AP2	**0.63**	**0.70**	0.47	0.47	0.21	0.22
CD1	0.58	0.47	**0/70**	**0.62**	0.26	0.35
CD2	0.57	0.47	**0.62**	**0.70**	0.26	0.30
EH1	0.24	0.21	0.26	0.26	**0.70**	**0.63**
EH2	0.30	0.23	0.35	0.31	**0.63**	**0.70**

It should be noted that, in the previous empirical testing, even for texts confirmed to be from the same author, the SS has a tendency to cap around 0.70. This conservative score is due because the following reasons:

- Natural variations in writing: an author's style is never 100% consistent and fluctuates across different passages based on narrative needs, preventing perfect feature alignment.
- Model Calibration: the Random Forest model was implemented so to minimize the risk of overfitting (e.g. shallow trees, limited features) to avoid extreme certainty.
- Entropy-Based Similarity: Since the ML component measures uncertainty rather than confidence, scores remain moderate even for highly similar texts.

This should not be seen as a flaw, though, but as a safeguard against false positives. In forensic applications, it is preferable to err on the side of caution rather than overestimate stylistic matches [10].

Adjustments (e.g., weighting the statistical or ML components differently) can refine sensitivity, but the current balance ensures robustness across diverse texts.

4. Discussion

The similarity scores align well with the known stylistic differences among the authors considered in the experiment and the model was able to identify all matching texts as well as correctly point out differences in 100% of the cases under analysis.

For instance, the high SS values for intra-author comparisons reflect the tool's sensitivity to consistent stylistic features outlined by high level punctuation usage and sentence related features. Conversely, the low scores for inter-author comparisons (e.g., 0.24 for Poe's "Berenice" and Hemingway's "The Snows of Kilimanjaro") highlight the tool's ability to differentiate distinct writing styles clearly.

The conservative nature of the scores is a deliberate design choice to minimize false positives, which remains a critical consideration in forensic and academic applications where overconfidence in attribution could have significant consequences.

5. Conclusions

The experimental results presented in this study demonstrate the efficacy of the proposed stylometric tool in distinguishing authorship across short stories and writings by different authors. The tool's ability to generate consistent similarity scores (SS) above the 0.60 threshold for texts by the same author, while maintaining lower scores for inter-author comparisons, underscores its potential as a reliable method for authorship verification across a variety of forensic applications.

Most importantly, the tool is designed without references to an explicit corpus of dictionary words, making it language-agnostic. By avoiding reliance on lexical or semantic content, the proposed approach remains adaptable across Western languages with similar orthographic conventions. This approach contrasts with more complex models based on n-gram analysis or function word frequency that depend on language-specific dictionaries or large training corpora, making the proposed tool both lightweight and versatile and suitable for real-world applications where resources may be limited.

Additionally, the transparency of the feature set, comprising easily interpretable metrics like punctuation frequency and average word length per sentence, facilitates clearer interpretation of results, a significant advantage in forensic and academic settings where explainability is paramount.

This study is also directly relevant to industrial and business contexts where it is critical to know who is writing. Corporate and Industrial communication channels are frequent targets of impersonation, including emails that appear to originate from executives, coworkers and different stakeholders

Such tactics can coerce employees into harmful actions, and as AI systems become more capable, detecting messages authored by an impostor rather than the purported human sender will likely become more difficult when the actual individual is not behind the laptop.

A lightweight and interpretable authorship signal can serve as an additional control within email gateways, ticketing systems, and approval workflows by flagging anomalous messages, validating sensitive instructions, and providing auditors with traceable evidence. As a future development, we will undertake a systematic evaluation that includes deploying on enterprise systems to score drafts in real time, together with a comprehensive assessment on a dedicated email database.

References

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

[2] Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233-334.

[3] 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.

[4] Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed *Federalist* Papers. *Journal of the American Statistical Association*, *58*(302), 275–309. https://doi.org/10.1080/01621459.1963.10500849

[5] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In Proceedings of the 17th International Conference on Natural Language Processing (ICON), pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

[6] Tyo, Jacob & Dhingra, Bhuwan & Lipton, Zachary. (2022). On the State of the Art in Authorship Attribution and Authorship Verification. 10.48550/arXiv.2209.06869.

[7] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multichannel convolutional neural networks for large-scale authorship attribution. arXiv preprint arXiv:1609.06686.

[8] Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes-y Gómez, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In EACL (2), pages 669–674.

[9] Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character ngram features. Journal of Law and Policy, 21(2):421–439.

[10] Robson Samuel, Searston Rachel, Thompson Matthew, Tangen Jason. A guide to measuring expert performance in forensic pattern matching. Behav Res Methods. 2024 Sep;56(6):6223-6247. doi: 10.3758/s13428-024-02354-y