Exploring retrieval augmented generation in stylometric analysis

Mohamed Saied, Nada Mokhtar, Michael Adel, Eslam Atwan, Philopater Boles
School of information technology and computer science, Nile University

***Abstract*—This paper demonstrates the effectiveness of Llama models with a retrieval-augmented generation framework in identifying unique authorial styles. By combining advanced text representation and retrieval techniques, these models excel in stylistic analysis, with applications in forensic analysis, literature studies, and digital communications.**

***Index Terms*—NLP, Llama Models, Retrieval-Augmented Generation, Stylistic Analysis.**

I. Introduction

THE study of writing styles to determine the author of a work is known as stylistic analysis. It examines writing patterns such as sentence construction, word choice, and distinctive stylistic elements. This approach is helpful in a variety of domains, including forensic science, internet communication, and literature. With the rise of digital texts and sophisticated machine learning techniques in recent years, its significance has only increased.

For working with text, contemporary natural language processing (NLP) models such as LLAMA (Large Language Model Meta AI) are incredibly effective tools. These models have a thorough understanding of the text and are able to spot subtle variations in writing style. They are therefore fit for this task.

Enhancing Large language models with RAG enhances their performance and reduces hallucinations. We used RAG on a dataset of texts and their authors to perform stylometric analysis.

II. Related work

Stylometry is the study of writing style, the characteristic manner of writing unique to an individual. Recent developments broadened its use to improve language models by giving it stylistic properties, identifying AI-generated text,and settling questions of authorship. Stylistic differences in translations, as influenced by social or cultural factors, have also been studied by researchers. These studies testify to the increasing complexity of the ways that stylistic patterns can be captured,and how this process within and across languages and genres poses challenges for researchers.

It was the emergence of transformer-basedmodels that revolutionized stylometric analysis. Ksieżniak et al. (2024) introduced stylometric tags (such as sentence complexity, and punctuation usage) into the trainingdata of BERT-family models that allowed the model to become more sensitive to stylistic features. For smaller (having few variants, few stylistics) benchmark datasets their improvement was more pronounced and depends on datasetcomplexity. Verified across multiple datasets, demonstrated that while the approach yielded better generalization the hypothesis on stylistic sensitivity was notconclusive and required further exploration to evaluate possible augmentation options.

Zamir et al. (2024) proposed a merit-based late fusion framework for transformer architectures that uses techniques that include particle swarm optimization to combine the predictions. They improved results by 5% using their benchmarks, where they proved that special characters should not be removed during preprocessing. Moreover, the bidirectional long short-term memory neural net (BLSTM) based approach, although placing the switches between authors correctly, was expensive computationally, and it was not able to recognize multiple changes in writing styles for a single author (firstauthor). Aiming to overcome such challenges is an avenue for future work by enhancing the efficiency of these approachesas much as ensuring that they can be applicable to informal, heterogeneous texts.

The slightly different approach has recently been taken in AI-generated text detection stylometric methods. Zaitsu et al. (2024) Used a random forest classifier allowed for more than 99. 5% accuracy classifiers based on the Consumer -Burnout, Word- and Phrase-based features such as the function words frequency within the human and ChatGPT generated text; Language-specific aspects like use of limed feature performance (comma use)for studies focused on Japanese language, and results showed a need for more exploratory studies based on multilanguage and multitype of text-based data.

Kwok, Moratto, and Liu (2024) provided an analysis of stylistic differences in Hongloumeng between two English translations. Translators' education and techniques had stylization effects: one translation adapted to the conventions of English fiction, while the other favored the original. Similarly, Zaki and Mohamed (2024) conducted a comparison between two renderings of Awlad Haratina andfound distinct styles in the two translations and that Theroux's translation was lexically more diverse and complex. Both of the studies have highlighted the socio-cultural context of translation and suggested that further investigation should be carried out to keep the stylistic devices and the dynamics of the retranslation of the fact.

Agapitos and van Cranenburgh (2024) performed a computational analysis of authorship controversies applied to the case of Octavia and Hercules Oetaeus with methods including PCA and Bootstrap Consensus Trees. The plays were consistent with Seneca's corpus, but some stylistic oddities indicated

partial non-Senecan authorship. Join them as they detail their ideas about the challenge of attributing ancient texts - and the tools of computation needed to do so.

III. Methodology

Our analysis focuses on developing a stylometric analysis tool to detect authorship for unique writing traits using state-of-the-art natural language processing (NLP) techniques. We use Llama models with a retrieval-augmented generation (RAG) framework to achieve an effective stylometric analysis tool.

Our analysis utilized three models of Llama: Llama 3.2 which had 3 billion parameters, Llama 3.1 which had 8 billion parameters, and Llama 2 which had 7 billion parameters. Each of these models was carved for one purpose - stylometric analysis, which offers varying levels of contextual understanding. All models have been improved through Retrieval-augmented generation (RAG) and prompt engineering. All of them are quantized.

Each text sample was transformed into high-dimensional embeddings that captured semantic and stylistic nuances using HuggingFace embeddings. For retrieving relevant samples during the RAG process, these were indexed using FAISS-Facebook AI Similarity Search and ChromaDB. This retrieval-augmented approach contextualized the analysis by including related samples from the dataset, which enhanced its ability to pick out subtle stylistic features.

A. Dataset

The dataset includes 19,579 text samples written by three writers, Edgar Allan Poe (EAP), H.P. Lovecraft (HPL), and Mary Shelley (MWS). Each instance is structured into three attributes: an Id, the body of text for analysis, and the corresponding author label. HuggingFace embeddings were used to represent the text. ChromaDB was used for saving and searching the embeddings during analysis.

B. RAG setup

We use Langchain to construct our RAG setup. Langchain is a popular framework for building AI applications.

IV. Results

Three models, Llama 2 7B, Llama 3 8B, and Llama 3.2 3B, along with their Weighted Voting mechanism, were used to perform the proposed analysis for the authorship identification problem. This section discusses the obtained results—accuracy scores, confusion matrices, and predicted author distributions. Each aspect provides insights into the performance of the individual and combined models in identifying authors through their stylometric features.

A. Comparisons of Accuracies

Table I summarizes the accuracy of the three models and the Weighted Voting mechanism:

- **Llama 2 7B:** Achieved an accuracy of 70%, showing fair success in capturing stylistic differences. The model worked well with some authors while struggling with others.
- **Llama 3 8B:** Delivered the lowest accuracy at 54%, suggesting its limited ability to differentiate between authorial styles.
- **Llama 3.2 3B:** Emerged as the best-performing model, achieving an impressive accuracy of 94%, demonstrating a robust understanding of subtle writing patterns.
- **Weighted Voting:** Attained an accuracy of 93%, closely matching Llama 3.2 3B's results. It combined predictions from all models with highly consistent performance.

TABLE I
Model Accuracies

Model	Accuracy (%)
Llama 2 7B	70
Llama 3 8B	54
Llama 3.2 3B	94
Weighted Voting	93

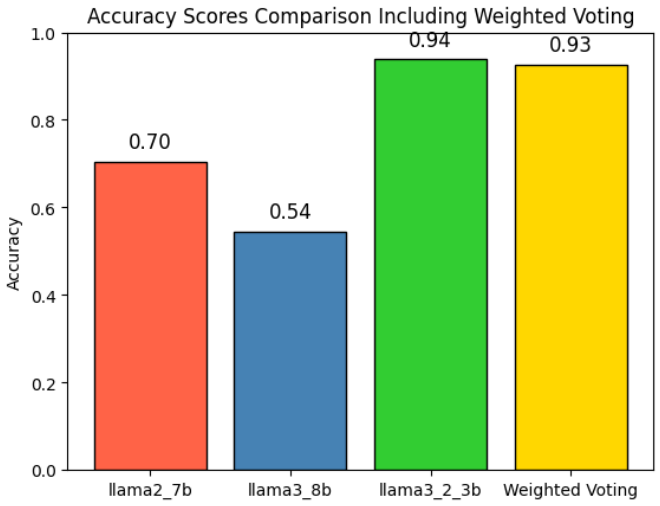

Fig. 1. Bar Chart Comparing Model Accuracies.

B. Predicted Author Distributions

The distribution of predicted authors provides insight into the decisions of each model:

- **Llama 2 7B:** Predicted Edgar Allan Poe (EAP) most frequently, followed by Mary Shelley (MWS). The number of misclassifications for H.P. Lovecraft (HPL) and "None" was moderate, showing some ability to capture stylistic features.
- **Llama 3 8B:** Demonstrated a consistent bias toward predicting Edgar Allan Poe (EAP) but did not predict any texts as "None." However, it struggled to discern patterns across other authors.
- **Llama 3.2 3B:** Made relatively balanced predictions across EAP, MWS, and HPL, with very few in the "None" category. It best captured subtle stylistic variations between authors.
- **Weighted Voting:** Combined strengths from all three models and obtained a well-distributed set of predictions, most of which were quite close to the actual values.

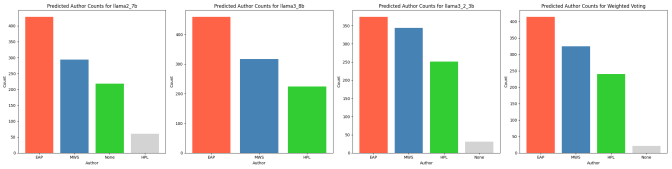

Fig. 2. Predicted Author Distributions for Each Model.

C. Confusion Matrices

Confusion matrices provide an overview of classification accuracy by highlighting correct and incorrect predictions:

- **Llama 2 7B:** Performance was moderate; most EAP and MWS texts were correctly classified, with only a few HPL texts wrongly categorized as EAP or MWS.
- **Llama 3 8B:** Struggled across all categories, frequently misclassifying texts. It lacked a distinct pattern of classification among MWS and HPL but avoided "None" misclassifications entirely.
- **Llama 3.2 3B:** Delivered the most accurate predictions, with minimal misclassification and clearer distinctions between authors.
- **Weighted Voting:** Outperformed individual models by leveraging their strengths and reducing their errors, increasing the reliability of the classification.

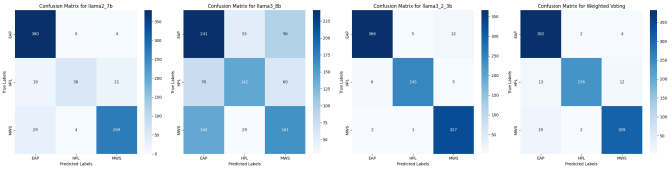

Fig. 3. Confusion Matrices for Each Model and Weighted Voting.

D. Summary Observations

- **Individual Model Performance:** Llama 3.2 3B was the best, with its 94% accuracy rating demonstrating great success in spotting subtle authorial styles.
- **Weighted Voting:** The Weighted Voting mechanism closely matched Llama3.2 3B with 93% accuracy by smoothing the models' predictions, mitigating any weaknesses, and providing strong results.
- **Practical Implications:** These results confirm the utility of using a combination of several Llama models with a Weighted Voting mechanism in performing stylometric analyses. This approach balances the limitations of individual models and provides reliable predictions on diverse datasets.

V. Conclusion

In conclusion, this paper highlights the powerful capabilities of advanced natural language processing techniques, particularly Llama models with a retrieval-augmented generation framework, in uncovering the unique stylistic fingerprints of authors. By combining sophisticated tools for text representation and efficient retrieval methods, these models excelled at capturing the nuances of individual writing styles. This work demonstrates how modern NLP technologies can bring new precision and depth to fields like forensic analysis, literature studies, and digital communications, paving the way for more refined and insightful applications of stylistic analysis.

References

[1] Agapitos, P., Van Cranenburgh, A. (2024). A stylometric analysis of Seneca's disputed plays. Authorship verification of Octavia and Hercules oetaeus. *Journal of Computational Literary Studies*. https://doi.org/10.48694/jcls.3919

[2] Księżniak, E., Wecel, K., Sawiński, M. (2024). Team OpenFact at PAN 2024: Fine-Tuning BERT Models with Stylometric Enhancements. https://www.semanticscholar.org/paper/Team-OpenFact-at-PAN-2024

[3] Kwok, H. L., Moratto, R., Liu, K. (2024). Activity versus Descriptivity: A Stylometric Analysis of Two English Translations of Hongloumeng. *Glottometrics*, 56, 1–21. https://doi.org/10.53482/2024$_5$6$_4$14

[4] Zaitsu, W., Jin, M., Ishihara, S., Tsuge, S., Inaba, M. (2024). Can we spot fake public comments generated by ChatGPT(-3.5, -4)?: Japanese stylometric analysis expose emulation created by one-shot learning. *PLoS ONE*, 19(3), e0299031. https://doi.org/10.1371/journal.pone.0299031

[5] Zamir, M. T., Ayub, M. A., Gul, A., Ahmad, N., Ahmad, K. (2024). Stylometry Analysis of multi-authored documents for authorship and author style change detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.06752