Author Attribution Using Stylometry

Related Work and Bibliography

Project Description

Determining the author of a text based on writing style is a longstanding challenge in natural language processing. This project will implement two different methodologies for author attribution using stylometry. The first approach is feature-based, extracting key stylistic markers such as character-level features (e.g., average word length, punctuation frequency), word-level features (e.g., most common words, rare word usage), and syntax features (e.g., POS tag frequency, sentence structure). These features will then be used to train classifiers like Support Vector Machines (SVM) and Random Forest from Scikit-Learn.

The second approach will leverage deep learning, specifically transformer-based models like BERT (via Hugging Face), to learn authors' patterns from contextual embeddings. The evaluation will measure precision, recall, and F1-score on a dataset containing texts from multiple authors, such as Project Gutenberg (via NLTK's Gutenberg corpus) or the PAN Author Identification dataset. A comparative analysis of the results will highlight the strengths and weaknesses of traditional feature-based methods versus neural network approaches, providing insights into the effectiveness of modern deep learning techniques in capturing an author's unique writing style.

Related Work

Table of Contents

Authorship Attribution Using Stylometry and Machine Learning Techniques

Ramnial, Panchoo, and Pudaruth (2016) explore authorship attribution and plagiarism detection using stylometric analysis combined with machine learning algorithms. Their method leverages a set of 446 stylometric features from basics like sentence length and vocabulary richness to more specialized features like punctuation distribution, function word

frequency, and word endings. Using SMO and k-NN classifiers on segmented PhD theses, they achieved up to 98% attribution accuracy on 10,000 word samples. Their results also demonstrated that classification performance declines with shorter segments and more authors, underlining the importance of document length and the number of authors in the dataset when using stylometry.

This study offers a valuable comparison for the feature-based portion of my project. While my methodology uses a more compact set of stylometric markers combined with classifiers like Support Vector Machines (SVM) and Random Forest, their broader feature set and high accuracy can inform my feature selection process. Also, their segmentation experiments provide guidance on evaluating performance across different text lengths and are an important consideration for my comparative analysis of traditional and deep learning methods. Their approach reinforces the power of selecting specifically catered features and using classifiers as a strong baseline against which to evaluate modern transformer based models.

Stylometric Analysis for Authorship Attribution on Twitter

Bhargava, Mehndiratta, and Asawa (2013) address the challenge of authorship attribution on Twitter by applying stylometric analysis to the short-form, informal texts that are tweets. Their approach leverages features such as hashtag frequency, character-level n-grams, and lexical choices from curated tweet datasets. After preprocessing to remove noise (e.g., slang, URLs, and common stopwords), they use Support Vector Machines (SVM) to classify tweets based on stylistic patterns. Despite the constraints of tweet length and informal language, their experiments demonstrate that author style is discernible through consistent usage patterns, enabling effective attribution even in very short texts.

This work provides a relevant point of comparison for the feature-based portion of my project, which similarly relies on extracting character, word, and syntax level features to train classifiers like SVM and Random Forest. While my focus is on longer texts from structured sources like Project Gutenberg, their results suggest that stylometric signals are robust enough to function across smaller text lengths. Their use of lexical and character level traits as key features also aligns with my feature design.

Stylometry and Authorship Attribution: Introduction to the Special Issue

Calle-Martín and Miranda-García (2012) performed a survey of stylometry and authorship attribution, summarizing key historical methods and modern advancements. They outline a progression from early statistical metrics (e.g., Zipf's law, Yule's K) to advanced machine learning techniques, including unmasking and Support Vector Machines (SVM). They cover a diverse set of studies, from forensic cases and Alzheimer's affected writing to classic attribution problems like The Federalist Papers. Across these examples, a range of stylometric features (e.g, function word frequencies, POS tags, n-grams) combined with

techniques like Principal Components Analysis (PCA) Burrows's Delta demonstrate the versatility of computational stylometry in identifying authors from their works.

This work strengthens the methodology of my project, particularly in its emphasis on a feature-based approach. The authors' review supports the use of lexical, structural, and syntactic indicators, which align closely with my own feature extraction methods. Tools and techniques such as the Java Graphical Authorship Attribution Program (JGAAP), unmasking, and simplified Delta are presented as useful benchmarks for assessing the performance of traditional methods in contrast to deep learning models like BERT. By showing the applicability of stylometric analysis across various text genres and contexts, the article not only validates my strategy of comparing statistical and neural approaches but also encourages a more thoughtful selection of features to capture an author's distinctive qualities.

Neural Authorship Attribution: Stylometric Analysis on Large Language Models

In their study, Kumarage and Liu (2023) present an intriguing framework for neural authorship attribution by examining the writing styles of AI-generated text from large language models (LLMs). They extract 60 stylometric features (e.g., lexical indicators like type-token ratio and hapax legomena, syntactic features like POS tags, and structural features like punctuation and paragraph length) from a dataset of 6,000 news articles generated by LLMs. These features are used in a range of classifiers, including XGBoost and RoBERTa, as well as a hybrid model combining RoBERTa embeddings with stylometric inputs via an attention mechanism. The combined model achieved the highest attribution accuracy and demonstrates that even closely related models like GPT 3.5 and GPT 4 have discernible writing styles.

This approach aligns closely with my project, which also combines feature-based stylometry and transformer based classification (using BERT) for human authorship attribution. The study's use of interpretability tools like SHAP to understand which features most influence classification outcomes offers a potentially useful technique for analyzing and selecting the best features for my project. Also, their findings on model comparison, particularly the similarity of LLaMA 2 to proprietary models like GPT 3.5 and GPT 4, underscore the importance of fine grained feature analysis when distinguishing between closely related authors or models, in their case. Their methodology and results provide a strong technical and conceptual comparison and several points of potential integration into my own work.

Kafka's Literary Style: A Mixed-Method Approach

In the most recent study I found, Strathausen, Shang, and Kazakov (2025) explore Franz Kafka's literary style through a mixed-method approach that integrates stylometry with

interpretive literary analysis. They analyzed three corpora, Kafka's self-published works, his posthumous texts edited by Max Brod, and Brod's own writings, to determine whether AI can detect stylistic differences introduced by editorial changes. Using pre-trained, fine-tuned German BERT models and Bi-LSTM Q-Learning, they achieve high attribution accuracy and show conclusively that even subtle changes in editing can be stylistically measured. Complementary linguistic analyses of features like sentence length and modal verb frequency further support the idea that Kafka's style remains distinctive even when altered by his editor.

This work aligns closely with my project because not only does it focus on a classical author, it also evaluates authorship through a hybrid methodology combining stylometric feature extraction with deep learning models like BERT. Both my work and this study emphasize the value of contextual embeddings for detecting stylistic patterns, while also promoting the benefits offered by classical features. Where my project compares multiple authors, theirs demonstrates the sensitivity of authorship classifiers even within a single author's body of work, highlighting the nuanced interplay between author's voice and editor's influence.

A Weighted TF-IDF-Based Approach for Authorship Attribution

Abedzadeh, Ramezani, and Fatemi (2021) introduce a language independent author attribution method using a refined version of TF-IDF called TF-IDF AARR Weighted. Their method differs from traditional classification based approaches because it treats author attribution as an information retrieval problem. By calculating the similarity between an unknown document and a set of known documents from a particular author using normalized term frequencies, their technique minimizes the influence of document length on word frequency metrics. The system was evaluated across six datasets including IMDB, PAN2011, and Project Gutenberg with results showing up to 43% improvement in accuracy over baseline models, especially in longer text scenarios.

This approach complements the first part of my project, which also leverages stylometric features for author classification. Although my implementation involves explicit feature extraction and training classifiers like SVM and Random Forest, the TF-IDF AARR Weighted model offers a comparable method of comparing documents based on patterns of word usage. Their results on datasets like Gutenberg and PAN2011, which I also plan to use, provide a benchmark for evaluating my work. Also, the normalization technique used to account for document length offers an interesting strategy I might consider applying to reduce bias in my feature set.

Bibliography

Ramnial, Hoshiladevi, Shireen Panchoo, and Sameerchand Pudaruth. “Authorship Attribution Using Stylometry and Machine Learning Techniques.” *Intelligent Systems Technologies and Applications*, edited by S. Berretti et al., Springer, 2016, pp. 113–125. *Advances in Intelligent Systems and Computing*, vol. 384, https://doi.org/10.1007/978-3-319-23036-8_10.

Bhargava, Mudit, Pulkit Mehndiratta, and Krishna Asawa. “Stylometric Analysis for Authorship Attribution on Twitter.” *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, Proceedings*, edited by Vasudha Bhatnagar and Srinath Srinivasa, Springer, 2013, pp. 37–47. *Lecture Notes in Computer Science*, vol. 8302, https://doi.org/10.1007/978-3-319-03689-2_4.

Calle-Martín, Javier, and Antonio Miranda-García. “Stylometry and Authorship Attribution: Introduction to the Special Issue.” *English Studies*, vol. 93, no. 3, 2012, pp. 251–258. Taylor & Francis, https://doi.org/10.1080/0013838X.2012.668788.

Kumarage, Tharindu, and Huan Liu. “Neural Authorship Attribution: Stylometric Analysis on Large Language Models.” *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, IEEE, 2023, pp. 51–54. https://doi.org/10.1109/CyberC58899.2023.00019.

Strathausen, Carsten, Wenyi Shang, and Andrei Kazakov. “Kafka’s Literary Style: A Mixed-Method Approach.” *Humanities*, vol. 14, no. 3, 2025, p. 61. MDPI, https://doi.org/10.3390/h14030061.

Abedzadeh, Ali, Reza Ramezani, and Afsaneh Fatemi. “A Weighted TF-IDF-Based Approach for Authorship Attribution.” *2021 11th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, 2021, pp. 188–193. https://doi.org/10.1109/ICCKE54056.2021.9721474.