Author Attribution Using Stylometry and Transformers

Final Report

Destiny Hillis
CSPB 4830 – Natural Language Processing
Instructor: Professor Dr. Curry Guinn
May 2025

Abstract

This project explores authorship attribution using two different methodologies: a traditional, feature-based approach grounded in stylometry and a modern transformer-based approach using deep contextual embeddings. I implemented both pipelines on literary texts from the NLTK Gutenberg corpus, segmented into 50-word chunks. The feature-based model relied on syntactic and lexical features processed by classical classifiers like SVM and Random Forest, while the transformer approach leveraged frozen BERT embeddings and a logistic regression classifier. The transformer-based model consistently outperformed the feature-based model across six author sets, providing stronger performance in accuracy, precision, recall, and F1-score. However, the feature-based approach offered advantages in interpretability and robustness to metadata leakage. These results illustrate the relative strengths of traditional and contextual techniques in authorship attribution and provide a nuanced basis for comparing their use in stylometry.

Introduction

Authorship attribution seeks to identify the author of a given text based on stylistic features rather than content. This has real-world applications in fields like forensic linguistics and plagiarism detection, where determining authorship can have significant consequences. However, writing style is often subtle, and distinguishing between similar authors poses challenges for both humans and machines. In this project, I compare two approaches to authorship attribution: a traditional, feature-based pipeline using stylometric features and classical classifiers, and a deep learning pipeline leveraging BERT embeddings. The goal is to evaluate how each performs on short text segments from the same authors under consistent experimental conditions.

Related Work

Authorship attribution has long relied on stylometry, the analysis of linguistic style through quantitative features. A foundational survey by Stamatatos (2009) outlines the evolution of these

techniques, highlighting the effectiveness of lexical, syntactic, and structural features. This work helped shape my initial feature selection strategy, especially the use of character and function word frequencies.

Kestemont (2014) investigates the predictive power of function words, arguing that their frequency and distribution provide more robust stylistic signals than content-based features. This directly supports my decision to include function word counts in the feature-based pipeline.

Bhargava et al. (2013) explore authorship attribution in the context of social media, using tweets as a test case for short-text attribution. The short length of their samples informed my decision to experiment with 50-word chunks to simulate similar challenges in style detection. Their results demonstrated that even brief texts can be attributed with high accuracy when well-chosen features are applied.

Data

The dataset consists of literary texts from the NLTK Gutenberg corpus. I selected a pool of six authors and extracted multiple 50-word chunks from each text, resulting in approximately 17,000 labeled samples. This chunk size was chosen to simulate the challenges of short-text attribution and to balance class representation. Shorter samples also avoid issues with stylometric features varying too much across a larger sample and help normalize features to avoid skewed results.

Preprocessing involved:

- Removing Project Gutenberg metadata such as headers and footers
- Lowercasing all text and stripping punctuation
- Feature extraction for the feature-based pipeline

Each chunk was labeled with its author, and the resulting dataset was divided into six balanced subsets, each containing three authors and about 2,800–3,000 samples. These subsets enabled multiple controlled experiments under similar conditions.

Methodology

The project compares two distinct pipelines:

Feature-Based Pipeline

Stylometric features were extracted at multiple levels:

- **Character-Level:** average word length, punctuation frequency, uppercase frequency, digit frequency, n-grams, top 5 bigrams and trigrams
- **Word-Level:** frequency of function words, type-token ratio

- **Syntax-Level:** POS tag frequencies, count of main POS categories (nouns, verbs, adjectives, and adverbs)

These features were passed to two classical classifiers: Support Vector Machines and Random Forests, both using Scikit-learn. These models were chosen for their established performance in stylometry and their ability to handle high-dimensional feature spaces.

Transformer-Based Pipeline

For the transformer approach, I used pretrained BERT (via Hugging Face) to generate embeddings for each chunk. BERT outputs were pooled using mean pooling, then passed to a logistic regression classifier. This setup allowed for comparison against the feature-based pipeline while holding the classification layer simple and consistent. The transformer model was used as a frozen feature extractor and not fine-tuned due to resource constraints, which also helped avoid overfitting to the relatively small dataset.

I evaluated both pipelines across six author sets, each containing text chunks from 3 authors. Within each set, I performed an 80/20 train-test split to assess model performance on unseen data. This structure enabled consistent side-by-side comparisons between the feature-based and transformer-based methods on the same data.

Results

To compare the pipelines, I evaluated accuracy, precision, recall, F1-score, and confusion matrices across the same author sets. These metrics were manually reviewed, with a few representative tables included below. Full results and code are available in the accompanying Jupyter notebook. Beyond standard evaluation, I also automated an analysis of both models' predictions against the ground truth, tracking accuracy, model agreement, and identifying cases of disagreement. This layered approach provided insight into where each method succeeded or failed. The consistent evaluation framework ensured a fair comparison under identical conditions.

Note: This analysis was conducted across all author sets, but truncated tables are shown here for brevity. Full comparison results and additional examples are available in the project's Jupyter notebook.

Overall Accuracy

Accuracy reflects the proportion of predictions that matched the true author label across all test samples. For example, in Set 1, the feature-based model achieved 78% accuracy, while the transformer-based model reached 98%; similarly, in Set 6, the feature-based model scored 76%, compared to 95% for the transformer-based approach.

Set	Feature-Based Accuracy	Transformer-Based Accuracy
Set 1	0.78	0.98

Set 3	0.78	0.98
Set 6	0.76	0.95

Precision

The transformer model consistently achieved over 97% precision across author classes in most sets, while the feature-based model showed greater variability. The table below shows per-class precision values for both models:

Set	Feature - Class 0	Transformer - Class 0	Feature - Class 1	Transformer - Class 1	Feature - Class 2	Transformer - Class 2
Set 1	0.79	0.98	0.74	0.96	0.84	0.99
Set 3	0.79	0.98	0.72	0.99	0.79	0.98
Set 6	0.69	0.91	0.73	0.94	0.87	0.99

Recall

BERT achieved near-perfect recall across most sets and authors, while the feature-based model often struggled to identify certain authors reliably. The following table illustrates recall by class:

Set	Feature - Class 0	Transformer - Class 0	Feature - Class 1	Transformer - Class 1	Feature - Class 2	Transformer - Class 2
Set 1	0.90	0.99	0.53	0.95	0.81	0.99
Set 3	0.47	0.96	0.63	0.95	0.91	1.00
Set 6	0.64	0.91	0.77	0.94	0.86	1.00

F1-Score

The transformer-based model achieved consistently high F1-scores, often above 0.95, while the feature-based model showed more fluctuation depending on the author. The table below presents per-class F1-scores:

Set	Feature - Class 0	Transformer - Class 0	Feature - Class 1	Transformer - Class 1	Feature - Class 2	Transformer - Class 2
Set 1	0.84	0.98	0.62	0.96	0.83	0.99
Set 3	0.59	0.97	0.67	0.97	0.84	0.99

Set 6	0.66	0.91	0.75	0.94	0.86	0.99

Confusion Matrices

These showed that the transformer model made fewer misclassifications overall, particularly in sets where authors shared similar genres or stylistic traits (e.g., Austen and Chesterton). In contrast, the feature-based model struggled more in distinguishing these cases.

Set 1 Confusion Matrix

	0 (Feature)	**0 (Transformer)**	**1 (Feature)**	**1 (Transformer)**	**2 (Feature)**	**2 (Transformer)**
0	849	946	79	13	18	1
1	196	23	233	404	10	0
2	33	0	2	2	149	180

Prediction Comparison

To better understand model behavior, I compared predictions from each pipeline to the ground truth for all sets. The analysis revealed meaningful differences in how the models handle stylistic overlap between authors.

In Set 1, for example, the transformer-based model outperformed the feature-based model across all authors. Notably, both models agreed on the correct author for most samples, but 1,429 instances of disagreement were found—many involving authors with similar stylistic fingerprints, such as Austen and Chesterton. These areas of confusion help expose the limitations of traditional stylometry and highlight where deeper contextual representations offer advantages.

The table below shows per-author accuracy and agreement rates in Set 1:

Author	**Feature-Based Accuracy**	**Transformer-Based Accuracy**	**Model Agreement**
Austen	0.8183	0.9950	0.9218
Chesterton	0.5684	0.9892	0.9971
Shakespeare	0.8593	0.9978	0.9100 (est.)

Note: Agreement reflects the percentage of samples where both models made the same prediction, regardless of correctness.

Below are a few representative examples where the models disagreed. In each case, the feature-based model misattributed the author, while the transformer-based model predicted correctly:

Text Snippet (Excerpt)	True Author	Feature Prediction	Transformer Prediction
“Emma Woodhouse, handsome, clever, and rich...”	Austen	Chesterton	Austen
“...nursed her through the various illnesses of childhood. A large debt of gratitude was owing...”	Austen	Chesterton	Austen
“...such an affection for her as could never find fault. How was she to bear the change?”	Austen	Chesterton	Austen

Discussion

The transformer-based pipeline outperformed the feature-based approach across all author sets and evaluation metrics. BERT's contextual embeddings provided a robust representation of writing style, allowing the model to distinguish between authors even when stylistic overlap was present.

The feature-based model showed more variability in its predictions. It struggled in cases where chunks were too short for syntactic features to capture meaningful signals. This may be due to the limited expressiveness of syntactic features in 50-word chunks, which do not always capture an author's deeper stylistic patterns. Additionally, these features are more sensitive to surface-level variation and less effective at modeling high-level language structure.

Despite this, the feature-based approach was more interpretable, easier to run, and faster to train. It also showed more resistance to metadata leakage compared to the transformer pipeline, illustrating a trade-off between interpretability and performance that is common in NLP tasks.

Disagreement analysis revealed that the transformer model better handled stylistic nuance, while the feature-based model tended to misattribute when stylistic overlap or genre similarities were present. Further investigation into these differences could inform future modeling strategies.

Conclusion & Future Work

This project compared a feature-based stylometric approach with a transformer-based pipeline for authorship attribution, using a consistent evaluation framework across six author sets. The transformer-based model consistently outperformed the feature-based method in all metrics, demonstrating the strength of contextual embeddings in capturing subtle stylistic patterns. That said, the feature-based approach still offered key advantages: it was more interpretable, easier to run, faster to train, and showed greater resistance to metadata leakage compared to the transformer pipeline.

Looking ahead, several promising directions could extend this work. One approach would be to develop a hybrid model that combines stylometric features with transformer embeddings,

leveraging both interpretability and deep contextual information. Fine-tuning the transformer model, rather than using frozen embeddings, may also improve performance by allowing the model to better adapt to the authorship task. Further experiments could explore larger or more diverse datasets, including multilingual or genre-specific corpora.

Finally, deeper analysis of prediction disagreements could help identify specific linguistic features that distinguish successful classifications from errors—insights that could inform future model design and feature selection strategies.

Bibliography

Bhargava, Akash, et al. “Who Wrote the Tweet? Using Stylometry for Twitter Author Identification.” *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013, pp. 370–377. https://doi.org/10.1145/2492517.2492630.

Kestemont, Mike. “Function Words in Authorship Attribution: From Black Magic to Theory?” *Proceedings of the 3rd Workshop on Computational Linguistics for Literature,* 2014, pp. 59–66. https://doi.org/10.3115/v1/W14-0908.

Stamatatos, Efstathios. “A Survey of Modern Authorship Attribution Methods.” *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, 2009, pp. 538–556. https://doi.org/10.1002/asi.21001.