Author Attribution Using Stylometry

Interim Project Report

1. Project Summary

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.

---

2. Data

The primary datasets under consideration are:

- **Project Gutenberg Corpus** (via NLTK): A collection of literary texts from various authors, well-suited for experiments involving longer, structured documents.
- **PAN Author Identification Dataset**: A benchmark dataset widely used in authorship attribution tasks and competitions, providing labeled samples for evaluation.
- **Works by F. Scott and Zelda Fitzgerald**: As an interesting stretch goal, the project may apply attribution methods to investigate historical claims of whether F. Scott Fitzgerald published any of his wife's writing as his own.

Preprocessing steps

(in progress or planned)

- Tokenization and sentence segmentation
- Removal of metadata and non-authorial content (e.g., prefaces, footnotes)
- Extracting character, word, and syntactical features
- Normalization of stylistic features to control for variation in text length and formatting

---

3. Approach

Two distinct methodologies are being applied:

- **Feature-Based Stylometry**: Extract character, word, and syntax level features (e.g., average word length, punctuation frequency, POS tag distributions) and use them as input to supervised machine learning classifiers like **SVM** and **Random Forest** using **Scikit-learn**. This aligns with established methods in authorship attribution literature.

- **Transformer-Based Deep Learning**: Use **BERT** (via **Hugging Face Transformers**) to extract contextual embeddings, then fine-tune for authorship classification. Evaluation metrics include precision, recall, F1-score, accuracy, and confusion matrices to gain further insights into the model's performance.

The results from each methodology will be compared to assess which approach better captures the distinctiveness of individual authorial styles.

---

4. Progress So Far

- Conducted extensive literature review of related work.
- Identified a range of stylometric features from existing studies (e.g., average word, sentence, or paragraph length, vocab richness, hapax legomena—words that only appear once, POS tag distribution, punctuation frequency, etc.).
- Collected performance benchmark results from past studies for comparison.
- Designed the experimental pipeline for both the feature-based and deep learning approaches.

- Selected relevant libraries: Scikit-learn, NLTK, and Hugging Face.
- Started dataset preprocessing.
- Defined an evaluation strategy aligned with methodologies from related work.

5. Next Steps

- Complete preprocessing and perform feature extraction from selected corpora (Gutenberg and PAN to start).
- Implement and evaluate the feature-based models using traditional classifiers (SVM and Random Forest).
- Fine-tune BERT on the same dataset for comparative analysis.
- Potentially use visualization libraries to assess/interpret model decisions.
- Document and compare evaluation metrics (precision, recall, F1-score, accuracy, and confusion matrices) for both methodologies.

Anticipated challenges:

- Ensuring a fair and meaningful comparison between models trained on different feature sets.
- Normalizing and validating features across varied text lengths and styles.
- Aligning experimental findings with results reported in previous research, especially where methodology or scope may differ

6. Questions or Concerns

- How can I ensure that the BERT model is learning authorial style rather than just picking up on topic specific content?
- Would it be beneficial to explore feature fusion by combining handcrafted stylometric features with contextual embeddings or would it be more appropriate to keep the models distinct to maintain clarity in comparative analysis?