Author Attribution Using Stylometry

Project Proposal

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.