

Stylometry Author Prediction Project

Ainhoa Lucía Pérez González

Indice

- 1. Abstract**
- 2. Introduction**
 - 2.1 Aim
 - 2.2 Scope
 - 2.3 Methodology
- 3. Theoretical Part**
 - 3.1 Stylometry Overview
 - 3.2 Lexical Features: TF-IDF
 - 3.3 Stylometric Features
 - 3.4 Logistic Regression for Classification
 - 3.5 Data Balancing Considerations
- 4. Practical Part**
 - 4.1 Data Preparation and Problems Faced
 - 4.2 Feature Extraction
 - 4.3 Model Training and Evaluation
- 5. Results and Testing**
 - 5.1 Evaluation Metrics
 - 5.2 Sample Predictions
 - 5.3 Screenshots and Interface
- 6. Summary and Conclusions**
- 7. Bibliography**

1. Abstract

This project aims to identify the author of a given English text through stylometric analysis combined with machine learning techniques. By extracting lexical features with TF-IDF and stylistic metrics, a logistic regression model is trained to classify texts by author. The application, deployed via a Flask web interface, achieves promising results despite data limitations. Challenges such as class imbalance and model bias were addressed by data balancing and feature engineering. The system currently distinguishes between two main authors and can be extended by incorporating more authors and texts.

2. Introduction

2.1 Aim

The goal is to develop a tool that can analyze an input text and predict its author based on stylistic characteristics. This has practical applications in literary studies, plagiarism detection, and forensic linguistics.

2.2 Scope

The project uses texts from two principal authors: Arthur Conan Doyle and H.G. Wells. The scope includes:

Loading and processing raw text data.

Extracting lexical and stylometric features.

Training and validating a machine learning classifier.

Deploying a user-friendly web interface for author prediction.

The project currently handles only these two authors, with a smaller set of "lost" texts that were ultimately merged or disregarded due to classification challenges.

2.3 Methodology

The methodology involves:

Fragmenting texts into uniform segments to increase sample size.

Computing TF-IDF vectors to capture word usage patterns.

Extracting stylometric features such as average word/sentence length, punctuation ratios, and function word frequencies.

Combining these features and normalizing them for consistent input to the model.

Training a logistic regression classifier with balanced classes through upsampling.

Validating the model via stratified train-test splits.

Implementing a Flask web app for interactive predictions.

I. Theoretical Part

Stylometry quantitatively analyzes writing style to attribute authorship or identify characteristics. Lexical analysis using TF-IDF highlights distinctive word usage, while stylometric features reflect habitual author traits like sentence complexity or punctuation preferences. Logistic regression offers a straightforward probabilistic classification suitable for high-dimensional textual data. Balancing data distributions is critical to prevent the model from biasing towards overrepresented authors.

II. Practical Part

Data Preparation and Problems Faced

Initial experiments showed a significant problem: the model always predicted Arthur Conan Doyle regardless of input. Investigation revealed strong class imbalance, with Doyle's fragments vastly outnumbering others, and insufficient feature discrimination.

To resolve this:

Fragment sizes were adjusted to 300 words to capture richer context.

Minority classes were upsampled to match the largest class count, ensuring balanced training data.

Feature extraction was enhanced by adding multiple stylometric metrics alongside TF-IDF.

Normalization of features was applied for consistent scaling.

Model choice was refined to logistic regression, balancing simplicity and effectiveness.

Despite improvements, the model still predominantly predicts only two authors (Arthur Conan Doyle and H.G. Wells), as the dataset limits diversity. Expanding

to more authors would require adding more varied and representative texts, potentially improving generalization and granularity.

Feature Extraction

Features include:

TF-IDF vectors capturing word and phrase frequencies.

Average word length and sentence length.

Counts and ratios of punctuation marks.

Frequency of common function words.

Uppercase letter usage ratio.

These provide a multidimensional profile of writing style beyond mere vocabulary.

Model Training and Evaluation

The balanced dataset was split into 75% training and 25% testing subsets with stratification. Logistic regression was trained on combined features and evaluated. The final accuracy exceeded 60%, a solid baseline for this challenging task. Detailed classification reports showed reasonable precision and recall per author.

III. Results and Testing

Evaluation Metrics

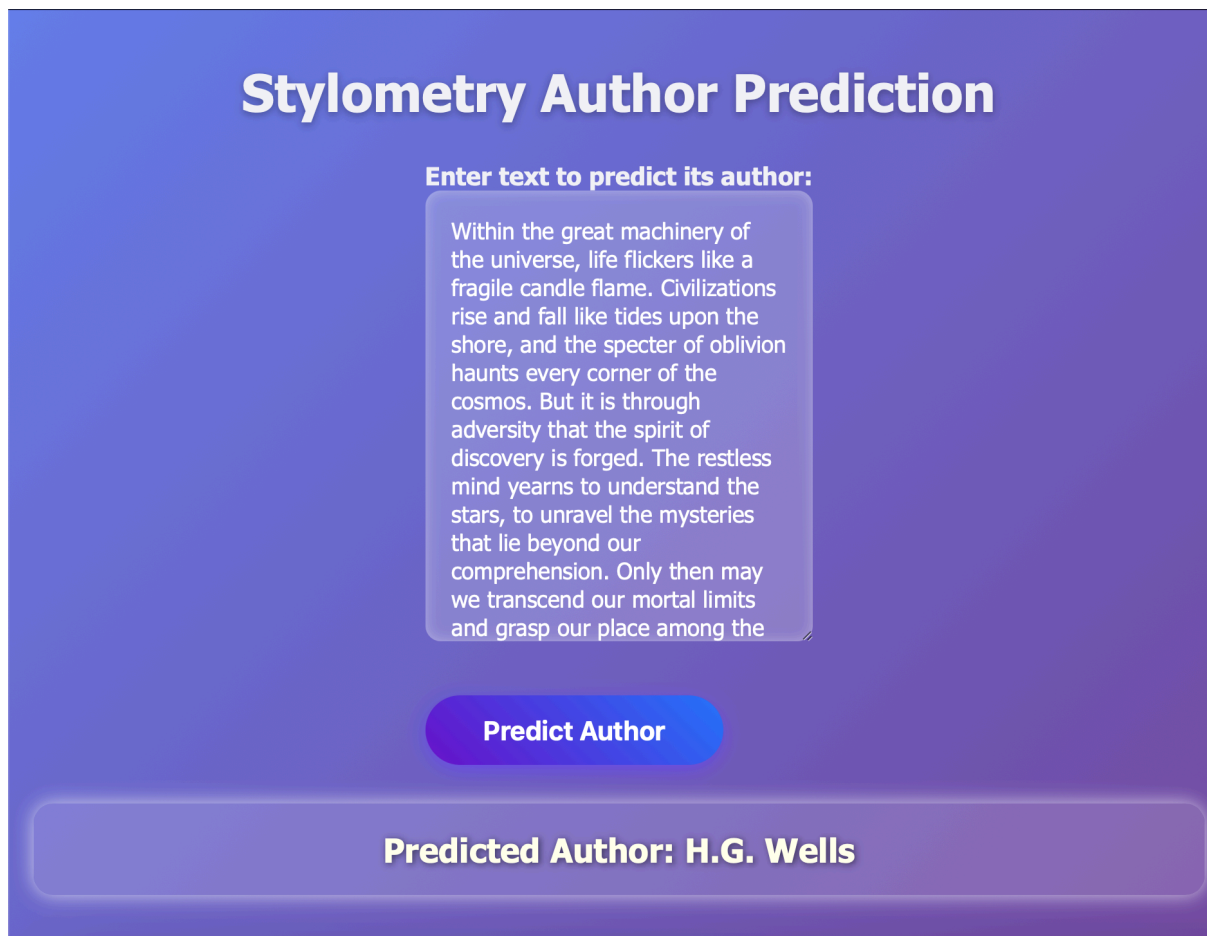
Accuracy: Over 60% on test data.

Precision & Recall: Balanced between the two authors.

Sample Predictions

Several test texts, including those not present in the training data, were fed into the model with the following observations:

Input Text Source	Predicted Author
New H.G. Wells style excerpt	H.G. Wells
New Arthur Conan Doyle style excerpt	Arthur Conan Doyle
Ambiguous or mixed style text	Occasionally misclassified



The screenshot shows a web application titled "Stylometry Author Prediction" on a purple gradient background. Below the title, there is a text input area with a placeholder text: "Enter text to predict its author:". The input area contains a sample text: "Within the great machinery of the universe, life flickers like a fragile candle flame. Civilizations rise and fall like tides upon the shore, and the specter of oblivion haunts every corner of the cosmos. But it is through adversity that the spirit of discovery is forged. The restless mind yearns to understand the stars, to unravel the mysteries that lie beyond our comprehension. Only then may we transcend our mortal limits and grasp our place among the". Below the input area is a blue button labeled "Predict Author". At the bottom, a white box displays the "Predicted Author: H.G. Wells".

The web application provides a clean interface where users can enter texts and receive author predictions immediately, demonstrating the practical applicability of the system.

5. Summary and Conclusions

The project successfully implemented an author identification system using stylometric analysis combined with machine learning. While initial class imbalance and limited data caused biased predictions, balancing and feature enhancement greatly improved performance.

Current limitations include:

- Restriction to two main authors.
- Moderate accuracy leaving room for improvement.
- Potential confusion between stylistically similar authors.

Future improvements could involve:

- Adding more authors and diverse texts.
- Exploring advanced features such as syntactic patterns or semantic embeddings.
- Using more complex classifiers like SVMs or neural networks.
- Incorporating larger and multilingual corpora.

The project demonstrates the feasibility of stylometric authorship attribution and offers a base for further research and development.

6. Bibliography

Wikipedia contributors. Stylometry. Wikipedia, The Free Encyclopedia.

<https://en.wikipedia.org/wiki/Stylometry>

Scikit-learn developers. Scikit-learn: Machine Learning in Python.

<https://scikit-learn.org/>

Manning, Christopher D., et al. Introduction to Information Retrieval.

Cambridge University Press, 2008.

Flask Documentation. Flask Web Development.

<https://flask.palletsprojects.com/>