Authorship Classifier for the Disputed Federalist Papers

João Silva
M2Ai – School of Technology, IPCA
Barcelos, Portugal
a16998@alunos.ipca.pt

***Abstract*—This paper presents an authorship classification approach for the Federalist Papers using feature extraction methods like Bag-of-Words and TF-IDF, along with stop word removal and Stemming for text processing. An SVC model achieved 76.5% accuracy in classifying the papers based on authors.**

***Index Terms*—Stylometry, Bag-of-Words, TF-IDF, SVC, Authorship Classifier**

I. Introduction

With the surge of new language models, that can generate human-like text, we are seeing a rise in concerns of plagiarism and misinformation [1]. Now, more than ever, the ability to determine the author of pieces of text that we find online has become much more important. Although the field of Authorship Attribution (AA) is not something new, given today's problem, the development of new techniques or solidification of existing ones is essential.

This work will serve as an introduction to the field of AA, staring with a bit of its history and motivation, but focusing on the exploration of its key concepts and techniques. I will be diving into more detail on the common text pre-processing and feature extraction techniques that are utilized in AA, and also clarify the methodology behind this work. After discussing these methods and ideas, the experimental part of this work begins.

Features extraction methods, like Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) will be experimented with and applied to a commonly used Corpus in AA studies: The Federalist Papers [2]. Afterward, these features will be fed into machine-learning models like Support Vector Classification (SVC) and predict the authors of specific essays of the federalist papers, the results and accuracy of the models will be analysed and finally conclusions and future work is going to be discussed in the last section.

All of the text processing, implementation and training of machine-learning models were done in Python, with the help of Natural Language Toolkit (NLTK) and Scikit-learn libraries.

II. Stylometry

Stylometry is the study of linguistic style, usually with a view to identifying and analyzing the characteristics of an authors writing. It involves quantitative analysis linguistic features such as sentence structure, vocabulary, punctuation and other textual patterns [3]. The goal of stylometry is to find patterns and distinctive traits in an authors written works, in order to identify the author and provide insight on their writing style.

Fig. 1: Federalist Papers Authors (Left to Right) : Hamilton, Madison and Jay

Stylometry has been utilized throughout history to identify the authors of anonymous or disputed documents and even detect forgeries or plagiarism. An example of this can be seen in the story of Lorenzo Valla from the 15th century, when he successfully demonstrated the forgery of the Donation of Constantine by comparing the Latin used during that period with its 4th-century equivalent [4].

Later in the 19th century, the basics of stylometry were established by Polish philosopher Wincenty Lutosławski, where used these new principles to develop a chronology of Plato's Dialogues [5].

However, it was only in the mid-20th century where the field of stylometry gained significant attention with the advent of computers and computational linguistics. One of the most notable figures in the history of stylometry is the statistician and authorship attribution pioneer, Frederick Mosteller. In the 1960s, Mosteller, along with David L. Wallace, conducted a groundbreaking study on the authorship of the disputed Federalist Papers, applying statistical analysis to linguistic features to determine the likely authors [6].

With the advancements in Natural Language Processing (NLP), Machine-Learning, and computational techniques, stylometry has witnessed significant progress. Today, with the emergence of chat models there is an increasing need to identify

AI-generated text to combat plagiarism, so it is important to keep studying and devoloping the field of Stylometry.

III. Stylometric Features

There are several different types of information that we can extract from text in order to identify distinct characteristics in someones writing style. They can derive from vocabulary, grammar, syntax, punctuation, word usage and sentence structure. These features contain unique information about writing styles and can be utilized in machine learning models. By examining these features, the model can detect patterns that are specific to a particular writing style and determine authorship. These features are typically categorized into three distinct groups: **Word-Based**, **Stylometric-Based**, and **Syntax-Based features** [7].

A. Word-Based Features

These features rely on the frequent usage of specific words or phrases by an author, which can be highly effective. Authors often write with their unique vocabulary and utilize specific adjectives, making this approach particularly valuable. The main issue when utilizing this type of features, is that when two authors are writing about the same topic, they will use the same vocabulary and this makes it harder to differentiate between the two.

Some examples of word-based feature extraction techniques are:

- **BoW**: This technique represents text as a bag of its words, disregarding grammar and word order. It is one of the simplest techniques for feature extraction in NLP [8].
- **TF-IDF**: This technique is used to reflect how important a word is to a document in a corpus. It assigns weights to each word based on its frequency in the document and the inverse frequency of the word in the corpus [9].
- **Word Embeddings**: This technique represents words as vectors in a high-dimensional space. It captures the semantic meaning of words and their relationships with other words [10].

B. Stylometric-Based Features

Stylistic or Stylometric-based Features consider broader stylist aspects of text, they can happen on different levels of text:

- **Article-level**: Number of paragraphs, number of sentences or number of words.
- **Paragraph-level**: Number of sentences in a paragraph, average length of sentence, number of words in a paragraph [11].
- **Word-level**: Number of small words (less than four characters), average length of words.
- **Vocabulary Richness**: It can help differentiate between authors with distinct lexical preferences, styles, or levels of linguistic expertise. There are also different ways to quantify vocabulary richness, some include, Hapax Legomenon Ratio and Type-token ratio (TTR).

C. Syntax-Based Features

Syntax features can refer to the frequency and usage of function words, such as pronouns, articles, prepositions, conjunctions, and auxiliary verbs, as well as the way punctuation is used, frequency of commas, semicolons, colons, exclamation marks, or question marks. These features are less influenced by content and are more related to an author's unique writing style.

IV. Methodology

A. Corpus

The Federalist Papers are a collection of 85 essays that were a collaborative effort between Alexander Hamilton, James Madison, and John Jay (Fig. 1). They were initially published under the pseudonym "Plubius" to hide the identity of the authors. This makes it difficult to determine with absolute certainty who authored which papers, since Maddison and Hamilton claimed to have written the same papers. The common consensus is that some of the disputed papers were actually co-written by both of them [12] [6]. In order to train our machine-learning models, we obtained an e-book version of the papers from Project Gutenberg, a volunteer-driven initiative that aims to digitalize and archive cultural works, primarily books, in the public domain [13].

B. Pre-Processing

Since we want to use The Federalist Papers as a Dataset, it is important to pre-process the text data, so that it can be used in our models.

The first step is to parse through the entire text corpus and remove line breaks, convert text to lowercase and remove punctuation. Lowercasing text and removing punctuation in NLP tasks aid in normalization, reduce vocabulary size, eliminate noise, and promote text consistency for improved analysis and processing.

Aditionally, two more pre-processing techniques are applied to the text data with the help of the NLTK library, stop-word removal and stemming. This makes it so we have 3 different datasets to experiment with:

- Lower case and no punctuation.
- Lower case, no punctuation and stop-words removed.
- Lower case, no punctuation, stop-words removed and stemming.

Stemming is the process of reducing words to their root or base form. It involves removing common word suffixes to obtain the core meaning of a word. For example, the words "running","runs" and "ran" are converted to "run". The purpose of stemming is to normalize words so that variations of the same word are treated as a single term during analysis, helping to consolidate the vocabulary and reduce dimensionality of the data.

Another text processing technique, is stop-word removal. It basically consists in removing common words that do not carry significant meaning and occur frequently in text. Examples of these words include "and", "the" and "is". Removing these

words from text data before analysis helps in processing time since these can appear in large quantities within a given corpus and do not contribute much to the understanding of a document.

The second step is to label each chapter/paper of the Federalist Papers. The Project Gutenberg e-Book makes this easier since it names the authors behind each paper, the general consensus is that Hamilton wrote 51 of the papers, Madison wrote 26, Jay wrote 5 and the remaining 3 were a collaborative effort by Hamilton and Madison, (Fig. 1).

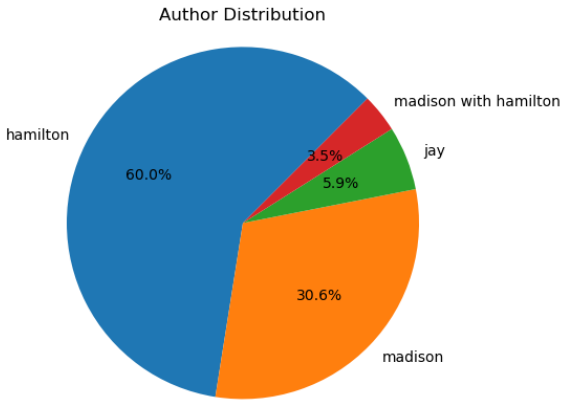

Fig. 2: Author Distribution in the Dataset: Pie Chart

Since the disputed papers were a collaborative effort, they were removed from the dataset.

C. Feature Extraction

Now that the data has been pre-processed, it can now be transformed into numerical features that machine-learning algorithms can work with.

One of the feature extraction techniques used was Bag-of-words (BoW), it treats the dataset as an unordered collection of words. It involves the following steps :

1) **Tokenization**: where the text is split into individual words or tokens.
2) **Vocabulary creation**: A unique vocabulary is created by collecting all the unique words from the text data of each chapter of the dataset.
3) **Vectorization**: Each chapter will be represented by a vector of the same dimensions as there are words in the entire vocabulary of the entire corpus. The vector shows the number of times a word occurs in each chapter of the corpus (Fig. 3).

The other feature extraction technique used was Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is an extension of the BoW model that also takes into account the importance of words withing the entire corpus. It aims to find the most relevant words by assigning them weights. It calculates this using two components:

1) **Term Frequency**: It measures the frequency of a word in the corpus, by calculating the ratio of the number of occurrences of a word to the total number of words in the corpus.
2) **Inverse Document Frequency**:It measures the rarity or importance of a word across the entire corpus, by calculating the logarithm of the ratio between the total number of chapters and the number of chapters that contain that word.

	Movie reviews
Review 1	This movie is good.
Review 2	The movie is not good.
Review 3	I love this movie. Watch, you will love it too.

	This	Movie	Is	The	Good	Of	Times	Not	I	Love	Watch	You	Will	It	too
Review 1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
Review 2	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0
Review 3	1	1	0	0	0	0	0	0	1	2	1	1	1	1	1

Fig. 3: Word-Count Vector

The TF-IDF score for a word in a document is obtained by multiplying its TF and IDF values. A higher score indicates that the word is more important within the document.

TF-IDF helps in reducing the impact of common words such as ”and”, ”the” and ”is” that appear frequently across documents, while emphasizing words that are specific to certain documents.

V. Experimental Results and Analyses

Based on the results from both feature extraction methods, it is evident that the combination of text processing techniques, such as stop-word removal and stemming, helped the model obtain better results.

TABLE I: SVC with TF-IDF

Processing	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LC+P[a]	58.8	34.6	58.8	43.5
LC+P+SW[b]	58.8	34.6	58.8	43.5
LC+P+SW+STE[c]	**64.7**	**72.1**	**64.7**	**55.3**

[a] Lower case and no punctuation
[b] Lower case, no punctuation and stop-words removed
[c] Lower case, no punctuation, stop-words removed and stemming

The BoW approach also yielded better results compared to the TF-IDF method, which was unexpected. This might be due to the distribution of words in the dataset being better captured by BoW, which focuses on word presence rather than their importance. Also, BoW lower dimensionality can help the classifier find patterns more easily, especially since we have a limited training sample.

TABLE II: SVC with Bag-of-Words

Processing	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LC+P	58.8	34.6	58.8	43.5
LC+P+SW	76.5	72.4	76.5	73.8
LC+P+SW+STE	**76.5**	**72.4**	**76.5**	**73.8**

VI. Conclusion

In addition to its primary goal of authorship classification in the Federalist Papers, this work also served as an introductory exploration into the field of stylometry. This work focused on delving into key concepts and techniques of stylometry, with a particular emphasis on feature extraction methods such as BoW and TF-IDF, as well as text processing techniques including stop word removal and Stemming.

While this work represents a preliminary step in the field of stylometry, it lays the groundwork for further investigations into more advanced concepts and techniques.

For future work, incorporating additional features such as syntax analysis and stylometric attributes could provide better results. Another way to improve these results would be to explore dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) to enhance the model's performance.

References

[1] Chris M Anson. Ai-based text generation and the social construction of" fraudulent authorship": A revisitation. *Composition Studies*, 50(1):37–179, 2022.

[2] Electronic Classics Series. The federalist papers. 2001.

[3] Helena Gómez-Adorno, Juan-Pablo Posadas-Duran, Germán Ríos-Toledo, Grigori Sidorov, and Gerardo Sierra. Stylometry-based approach for detecting writing style changes in literary texts. *Computación y Sistemas*, 22(1):47–53, 2018.

[4] Thomas Renna. Lorenzo valla and the donation of constantine in historical context, 1439–40. *Expositions*, 8(1):1–28, 2014.

[5] Wincenty Lutoslawski. Principes de stylométrie appliqués à la chronologie des œuvres de platon. *Revue des études grecques*, 11(41):61–81, 1898.

[6] Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[7] Ojaswi Binnani. Author identification using traditional machine learning models. In *CS & IT Conference Proceedings*, volume 12. CS & IT Conference Proceedings, 2022.

[8] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[9] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[10] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

[11] G Udny Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390, 1939.

[12] MG Kendall. Inference and disputed authorship: The federalist, 1966.

[13] Bryan Stroube. Literary freedom: Project gutenberg. *XRDS: Crossroads, The ACM Magazine for Students*, 10(1):3–3, 2003.