Detecting Various Writing Styles in Documents Through Inherent Stylometric Analysis

Alisher Beisembekov
Computer Science and Information Technology
Rochester Institute of Technology Dubai
Dubai, UAE
ab4227@rit.edu

Tairlan Kairolla
Computer Science and Information Technology
Rochester Institute of Technology Dubai
Dubai, UAE
tk6878@rit.edu

Abstract— **In this project, we have developed an intelligent system that takes a single document and classifies different writing styles within the document using stylometric analysis. The classification is done using K-Means Clustering (an unsupervised machine learning method). First, the document is divided into chunks of text using a standard chunk size. Then, for each chunk of text, a vector of stylometric features is computed. After that, the chunks are clustered using their vectors of stylometric features. This is where unsupervised machine learning comes into play. The chunks with the same style are clustered together, and the number of clusters made corresponds to the number of different writing styles that the document has. In this approach, the value of K is determined using the Elbow Method. We also ran an experiment for a document with two writing styles, and our system successfully identified that the document had two different writing styles. Our approach separates out text with the same style from that of different styles and can be used to detect plagiarism as well.**

I. Background

The study of measurable features of literary style, such as sentence length, readability scores, vocabulary richness, and various frequencies (of words, word lengths, word forms, etc.), has been around since at least the middle of the 19th century and has found numerous practical applications and interesting problems in the modern era of Artificial Intelligence and Machine Learning. This study of the literary style of a document is called Stylometry. Stylometry has evolved from earlier techniques for analyzing texts for evidence of authenticity, author identity, and other questions. The development of computers and their capacities for analyzing large quantities of data has greatly enhanced this type of effort. In the current era of research, Stylometry is widely used in intrinsic plagiarism detection, genre separation, authorship verification, authorship attribution, author gender detection, and more. The main task is to classify different writing styles from the text, which can further be used to solve the aforementioned problems.

II. Literature Review

An in-depth study of stylometry, authorship attribution, authorship obfuscation, along with a bit of theory on linguistics (e.g., Zipf's law), was conducted. These fields utilize stylometric features to address various problems. Since the work is related to literary writing style, studying linguistics was necessary to understand the nuances of these features. The study began with survey [1], covering the basics of stylometry, its applications in computer science, and its history. It was found that many researchers are using stylometric features for text analytics. Several research papers with significant citations were thoroughly studied. One research from the University of Illinois [2] at Chicago used stylometry to distinguish between actual human-readable text and paraphrased machine-written text using stylometric analysis. Features such as n-grams and lexical features such as frequency count, punctuations, and special characters were learned from this research paper. Intrinsic plagiarism detection, another application of stylometric analysis, was explored in [3]. Given a document, suspicious sections for plagiarism were identified. Another survey [4] described the use of machine learning and statistics with stylometry.

III. Introduction

Our system aims to determine variations in writing styles in a text document. These variations can be due to different authors or different genres of writing, such as stories, research papers, dramas, etc. Unlike other approaches to intrinsic plagiarism detection (essentially different writing style detection) that require a large corpus of texts from different authors to train their models, our approach doesn't need such training. It extracts the essence of the text style of each chunk of text using stylometric features and then groups together the chunks that have the same writing styles. This process is repeated for every new document. In this report, we elaborate on our entire methodology, including feature selection and data preprocessing. It also includes the machine learning method used. We then run an experiment on a document with two different writing styles to demonstrate our approach. The results are explained, and some limitations of our work are presented. Finally, we conclude our report.

IV. Methodology

First, the document is divided into chunks of text using a standard chunk size window. Then, for each chunk of text, a vector of stylometric features is computed. After that, the chunks are clustered using their vectors of stylometric features. This is where unsupervised machine learning comes into play. The chunks with the same style are clustered together.

A. Data Set Selection

We selected our dataset from the internet. http://textfiles.com/stories/ is an online repository that encompasses a vast collection of stories from different authors and with varying difficulty levels. While we aim to cluster different literary styles, we used this dataset for demonstration purposes. Our system can be run on any document.

B. Features Selection

The heart of our system lies in feature extraction. We need features that capture the style of the text, so we carefully crafted features for our project based on those studied during the literature review. To distinguish a chunk of text based on its literary style, we focused on three major categories: Lexical Features, Vocabulary Richness Features, and Readability

Scores. These include features like Shannon Entropy and Simpson's Index. Simpson’s index, for example, measures the diversity of a text, which is useful for our project. We used Python to calculate these features. The list of features extracted includes:

1) Lexical Features

1. Average Word Length
2. Average Sentence Length by Word
3. Average Sentence Length by Character
4. Special Character Count
5. Average Syllables per Word
6. Functional Words Count
7. Punctuation Count

These basic features provide insights into the structure of the text, such as average word lengths, the presence of special characters, and punctuation usage. Functional words are used to express grammatical relationships within sentences. Additionally, average syllables per word measures complexity, which is used in calculating other features related to readability scores.

2) Vocabulary Richness Features

Quantitative studies often rely on the concept of vocabulary richness. A text has low vocabulary richness if it repeats the same limited vocabulary, while it has high vocabulary richness if new words continually appear. These features indicate the diversity and richness of the vocabulary used in the text. Vocabulary richness features include:

1. Hapax Legomenon
2. Hapax DisLegemena
3. Honores R Measure
4. Sichel’s Measure
5. Brunet’s Measure W
6. Yules Characteristic K
7. Shannon Entropy
8. Simpson’s Index

a) Hapax Legomena and Hapax DisLegemena

Hapax Legomena, often referred to as "hapax," represents a word that appears just once within a specific context. This context could encompass the entire written language, an author's body of work, or even a single text. It's important to note that sometimes the term is mistakenly applied to describe a word found in only one of an author's works but occurring more than once within that specific work. The origin of the term lies in Greek, where "hapax legomenon" translates to "(something) being said (only) once." Similarly, Hapax DisLegemena pertains to words that occur precisely twice.

Now, to delve deeper into the remaining features, we utilize the following notation:

- Tokens (N): The length of the text, measured in words.
- Types (V): The count of distinct words within the text.
- Hapax legomena (V1): The number of words that appear just once in the text.
- Dislegomena (V2): The number of words that occur exactly twice in the text.
- Vi: The number of words that occur exactly 'i' times.

The type/token ratio is influenced by the text's length (generally lower for longer texts) but serves as a valuable metric for assessing vocabulary richness when comparing texts of equal length.

b) Honore’s measure R

Honore’s measure R [6] is based on the hapax legomena:

$$R = 100 * \frac{\log N}{1 - \frac{V1}{V}}$$

c) Sichel’s measure S

Sichel’s measure S [3] depends on the dislegemena and remains relatively consistent concerning 'N':

$$S = \frac{V2}{V}$$

d) Brunet’s measure W

$$W = N^{v-a}$$

Where 'a' represents a constant (typically 0.17). W is notably unaffected by text length and exhibits author-specific characteristics [4].

e) Yule’s characteristic K

Yule’s characteristic K [5] considers words of all frequencies:

$$K = 10000 * \frac{M - N}{N^2}$$

Where M is calculated as:

$$M = \sum_{i}^{n} i^2 * V_i$$

f) Shannon Entropy

In a broad sense, Entropy provides insights into the level of disorder within a given system. This concept has been applied within the context of our text-based project. Claude Shannon, the pioneer of information theory, introduced the Shannon entropy formula for quantifying the informational content conveyed by a word. The formula is represented as:

$$E = \sum_{i=0}^{N-1} P_i \log P_i$$

Here, 'P' represents the probability of a word occurring in the text passage, and 'N' signifies the total number of distinct words in the passage.

g) Simpson’s index

Simpson's Diversity Index serves as a metric for gauging diversity, and within the field of ecology, it is frequently employed to quantify the biodiversity within a specific habitat. This metric takes into consideration both the number of distinct species present and their respective abundances. Simpson's Index (referred to as 'D') quantifies the likelihood that two randomly selected individuals from a given sample will belong to the same species or another category beyond the species level. This concept can be extended to Natural Language Processing (NLP) applications to assess the diversity within a given segment of text. In our project, we have harnessed this biodiversity concept as a feature to evaluate the diversity of various text passages.

Mathematically, Simpson's Index (D) is calculated as follows:

$$S\ Index\ (D) = \sum (\frac{n}{N})^2$$

Here, 'N' represents the total number of words in the text, while 'n' denotes the total count of unique tokens within the text.

3) Readability Scores

Readability is the ease with which a reader can understand a written text. Features for readability are derived from linguistics and have been used to calculate readability scores in modern computer science. These readability features include:

1. Flesch Reading Ease
2. Flesch-Kincaid Grade Level
3. Gunning Fog Index
4. Dale Chall Readability Formula
5. Shannon Entropy
6. Simpson's Index

Each of these measures assesses the readability of the text from different perspectives, such as sentence length and word complexity. These features help gauge how easy or difficult it is for a reader to understand the text.

a) Flesch Reading Ease

The Flesch Reading Ease assessment, devised in 1948 as a readability evaluation tool, provides an approximate indication of the educational level required for comfortable comprehension of a given text [5]. This assessment generates a numerical score within the range of 1 to 100, although scores beyond this range can occasionally be computed. Subsequently, a conversion table is utilized to interpret the resulting score, wherein, for instance, a score falling between 70 and 80 corresponds to a seventh-grade reading level, signifying that the text should be readily understandable by the average adult reader. Originally developed to assist educators in selecting texts suited to their students' reading proficiency, the utility of the Flesch Reading Ease test has extended beyond the realm of education.

The formula for calculating the Flesch Reading Ease (FR) score is expressed as:

$$FR\ Score = 206.835 - 1.015\left(\frac{Total\ Words}{Total\ Sentences}\right) - 84.6(\frac{Total\ Syllables}{Total\ Words})$$

During the mid-1970s, the United States Navy sought a method for assessing the complexity of technical manuals employed in their training programs. Consequently, the Flesch Reading Ease test underwent revisions, along with other readability assessments, to better suit the navy's needs. This modified calculation was subsequently designated as the Flesch-Kincaid Grade Level [6], and the determination of grade levels was informed by the scores achieved by participants in a trial group.

The formula for computing the Flesch-Kincaid Grade Level (FKG) is given by:

$$FKG\ Level = 0.39\left(\frac{Total\ Words}{Total\ Sentences}\right) + 84.6\left(\frac{Total\ Syllables}{Total\ Words}\right) - 15.59$$

b) Gunning Fog index

In the field of linguistics, the Gunning fog index represents a readability assessment tool specifically designed for English text. This index serves as an estimate of the number of years of formal education required for an individual to comprehend the text upon initial reading. To illustrate, a fog index of 12 implies that the reader should possess the reading proficiency equivalent to that of a United States high school senior, typically around 18 years of age. The development of this test can be attributed to Robert Gunning, an American entrepreneur with a background in newspaper and textbook publishing, who introduced it in 1952. The formula for computing the Gunning fog index is presented as follows:

$$G = 0.4 * [\left(\frac{Words}{Sentences}\right) + 100(\frac{Complex\ Words}{Words})]$$

Here, "complex" words are defined as those comprising three or more syllables.

C. Data Pre-processing

After downloading the dataset from textfiles.com, which consists of various text files from different authors and genres, we selected a children's story and a research paper for proof of concept. The document is then divided into small chunks, with the chunk size set to an average of 10 sentences (adjustable as needed). Lexical features are computed for each chunk. For all other features, we removed punctuation and special characters and performed tokenization, as lexical features use punctuation and special characters.

D. Machine Learning Algorithm

Our approach uses unsupervised learning, and for clustering, we employed the widely-used K-Means algorithm.

E. PCA and Data Visualization

To visualize the clusters, Principal Component Analysis (PCA) is used to convert the high-dimensional feature vectors into a 2D space. In this space, chunks with the same writing style are plotted with the same color, enhancing the visual identification of different writing styles.

V. Experimental Settings

For the proof of concept, we chose two documents: one is a story named "Jim (Story)" and the other is a research paper named "AuthAttr (Paper)." We merged these two documents into one for the experiment. This combined document contains two different writing styles (a story and a research paper), and our system should identify these two writing styles.

A. K-Means

We used the K-Means algorithm to identify the number of different centroids in a text, each corresponding to a different writing style. The number of centroids (K) corresponds to the number of different writing styles in the document.

B. Value of K

To determine the optimal value of K, we used the Elbow Method. This method involves computing the sum of squared error (SSE) for different values of K (e.g., 2, 4, 6, 8) and selecting the K value where the SSE decreases abruptly. In our case, the Elbow Method successfully returned K = 2.

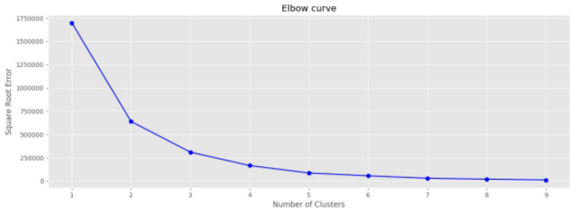
The Elbow method was employed in order to identify the most suitable value for 'K.'

C. Elbow Method

The elbow method can be delineated as follows:
Initially, the computation of the Sum of Squared Error (SSE) is carried out for various values of 'k,' such as 2, 4, 6, 8, etc. The SSE is defined as the summation of the squared distances between each cluster member and its respective centroid. Mathematically, this is expressed as:

$$SSE = \sum_{i=1}^{k} \sum x \in c_i \; dist(x, c_i)^2$$

Plotting 'k' against the SSE reveals a pattern where the error diminishes as 'k' increases. This phenomenon occurs because an increase in the number of clusters leads to a reduction in distortion. The fundamental principle behind the elbow method is to select the value of 'k' at which the SSE experiences a sudden and pronounced decline. This selection results in an "elbow effect" in the graphical representation, as depicted in the accompanying illustration.



In this particular instance, the optimal value for 'k' is determined to be 'k = 2'.

It is important to note that the elbow method operates as a heuristic, and consequently, its effectiveness may vary depending on the specific case at hand. There are instances where more than one elbow may be observed in the plot, or conversely, no discernible elbow may exist. In such scenarios, the determination of the most suitable 'k' value often entails evaluating the performance of the k-means clustering algorithm within the context of the clustering problem being addressed.

D. Parameter Tuning of K-Means

We applied the K-Means algorithm from the Python 'sklearn' library in our analysis. Initially, we determined the value of 'K' using the Elbow method. However, there are additional parameters that require careful consideration due to their significance. Following multiple experimental runs, we identified the following parameter values as optimal for our specific scenario:

a) n_init

Given that K-Means is a heuristic-based algorithm and its performance can be influenced by the initial seeds of centroids, we set the value of 'n_init' to 10. This parameter essentially involves random reinitialization of centroids, causing the K-Means algorithm to run 'n_init' times with different centroid seeds. The final output is selected as the best result from these 'n_init' consecutive runs, based on the inertia metric.

b) max_iter

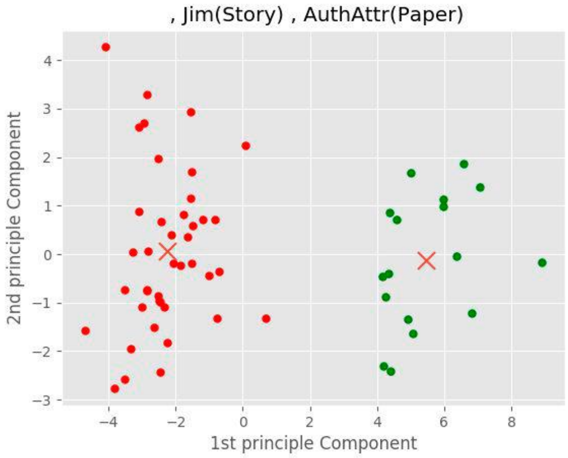
We configured 'max_iter' to be equal to 500, along with a minimum tolerance for convergence. 'max_iter' represents the maximum number of iterations allowed for the K-Means algorithm during a single run.

c) n_jobs

The ‘n_jobs’ parameter determines the number of CPU cores to be employed for computation. We opted for ‘n_jobs = -1,’ which effectively utilizes all available CPU cores on the host machine, enhancing computational efficiency.

VI. Results

We ran the experiment on a document containing two different writing styles: a story and a research paper. Our system correctly identified and clustered these two different writing styles, and the Elbow Method successfully returned K = 2, as expected. Visualization of the clusters using PCA further reinforced our results.



A document containing texts with 2 different writing styles (a story and a research paper) are clearly distinguished indicating the correctness of our approach

VII. Limitations

There are some limitations to our approach:

1. PCA Conversion: When converting high-dimensional vectors to 2D using PCA, there is a possibility of some loss of information and the potential for clusters not appearing as distinguishable as expected due to this loss. However, this is not a limitation of our system's ability to correctly identify the number of writing styles, which is the primary goal.

2. Single Author Scenario: In documents written by a single author, there may still be multiple clusters formed by our method, indicating different writing styles within the document. This is because even within a single author's work, there can be variations in writing style, such as differences in lexical features or the presence of poetry or technical content.

VIII. Conclusion

In conclusion, our system successfully identifies different writing styles within a text document using stylometric analysis and K-Means clustering. It does not require a large corpus of training data and can be applied to various documents. Our approach provides insights into both the number of writing styles and the dissimilarity between those styles within a document. This methodology has potential applications in plagiarism detection and genre classification, among others.

References

[1] E. Stamatatos, “A survey of modern authorship attribution methods,” *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, Mar. 2009, doi: https://doi.org/10.1002/asi.21001.

[2] U. Shahid, Z. Shafiq, S. Farooqi, P. Srinivasan, R. Ahmad, and F. Zaffar, “Accurate Detection of Automatically Spun Content via Stylometric Analysis,” 2017. Accessed: Sep. 13, 2023. [Online]. Available: https://web.cs.ucdavis.edu/~zubair/files/shehroze-text-spinner-icdm2017.pdf

[3] B. Stein, N. Lipka, and P. Prettenhofer, “Intrinsic plagiarism analysis,” *Language Resources and Evaluation*, vol. 45, no. 1, pp. 63–82, Jan. 2010, doi: https://doi.org/10.1007/s10579-010-9115-y.

[4] “APA PsycNet,” *psycnet.apa.org*. https://psycnet.apa.org/record/1949-01274-001

[5] J. Kincaid, R. Fishburne, L. Richard, B. Rogers, and Chissom, “Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel 1-1-1975.” Accessed: Aug. 08, 2019. [Online]. Available: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary

[6] A. Honore, “Some Simple Measures of Richness of Vocabulary, Association for Literary and Linguistic,” *Computing Bulletin*, vol. 7, no. 2, pp. 172–177, 1979.

[7] R. Flesch, “A new readability yardstick,” *Journal of Applied Psychology*, vol. 32, pp. 221–233, 1948.