R&D Project Report

Academic Year- 2022-23

Identifying the writing pattern from literary text.

Submitted by-

1. Aditya More	**BT20HCS348**	**CSE**
2. Kshitij	**BT20HCS224**	**CSE**

Prof. Ratna Sanyal and Prof. Anshima Srivastava

Introduction

Natural Language Processing (NLP) has revolutionized the way we analyze and understand human language. With the help of NLP techniques, we can now extract meaning, structure, and patterns from large volumes of text data. In this paper, we focus on the goal of identifying the author of an anonymous text by extracting features from literary text and training a machine learning model to recognize the unique writing style of each author.

The ability to identify the author of an anonymous text has important applications in fields such as forensics, literary analysis, and plagiarism detection. By leveraging NLP techniques, we can extract features from the text that are unique to each author, such as their choice of words, sentence structure, and use of literary devices. We can then train a machine learning model to recognize these features and use them to identify the author of an anonymous text.

Overall, the goal of this research is to demonstrate the power of NLP in extracting meaning and patterns from literary text. By leveraging machine learning techniques, we can identify the author of an anonymous text and gain deeper insights into the author's writing style and the structure of the text.

Problem Statement

In recent times, there has been an increase in cases of literary theft, infringement of literary rights, and attribution of authorship to the wrong person. It has become quite easy for someone to copy another person's work and present it as their own. Identifying the true author of a text is crucial in ensuring that they receive the proper recognition for their work. However, it is not always straightforward to detect when an article is falsely attributed to someone else. To address this issue, a system is needed that can analyze unstructured articles and accurately assign them to their original authors. This study will focus on the use of NLP analysis and machine learning algorithms to extract features from articles and predict the correct authorship.

Objective

The objective of this research paper is to develop a natural language processing (NLP) model trained on statistical features extracted from training data to accurately identify the author of an anonymous literary text. Additionally, we aim to detect other writing patterns from the text, such as sentence length, word frequency, and part-of-speech tags, to gain deeper insights into the author's writing style. The model also extracts other features like Type Token Ratio, LSA Vectors, Repetition Count, Rhyme Count, Alliteration Count, Keywords, Filtered Text, Collocations, Stemmed Words, Summary, Sentiment Analysis, Named Entities, etc.

Literature Review

Numerous studies have been conducted to develop NLP models for authorship attribution. Most of these studies have focused on statistical features extracted from the text, including word frequency, sentence length, and vocabulary richness. For example, Stamatatos et al. (2009) used a combination of lexical and syntactic features to develop an NLP model for authorship attribution. Their results showed that the model achieved high accuracy rates in identifying the author of anonymous texts. [1]

Similarly, Koppel et al. (2005) developed an NLP model based on a set of linguistic features, including word n-grams, function words, and part-of-speech tags. Their results showed that the model was able to accurately identify the author of anonymous texts with high accuracy rates.[2]

Another approach is to use machine learning algorithms to develop NLP models for authorship attribution. For example, Kestemont et al. (2013) used a support vector machine (SVM) algorithm to classify anonymous

texts based on their authorship. Their results showed that the SVM model achieved high accuracy rates in identifying the author of anonymous texts.[3]

In addition to statistical features, some studies have also focused on other types of features, such as psychological and social features. For example, Argamon et al. (2003) developed an NLP model that incorporates psychological features extracted from the text, including emotions and personality traits. Their results showed that the model achieved high accuracy rates in identifying the author of anonymous texts.
Overall, the literature suggests that developing NLP models trained on statistical features extracted from training data is an effective approach for authorship attribution. However, further research is needed to explore other types of features and to develop more advanced machine learning algorithms for this task. [4]

Proposed Methodolgy

- Data Collection: We will collect a dataset of literary texts written by known authors as well as a set of anonymous texts of unknown authorship.

- Data Preprocessing: We will preprocess the data by cleaning the text, removing punctuation and stop words, and tokenizing the text into words and sentences.

- Feature Extraction: Using NLP techniques, we will extract various statistical features from the preprocessed text, such as word frequency, sentence length, and part-of-speech tags. We will also use more advanced techniques such as sentiment analysis and named entity recognition to identify other writing patterns.

- Model Training: We will use machine learning algorithms to train a model to recognize these features in the anonymous literary text and

identify the author. We will use techniques such as k-NN, SVM, and neural networks to train and test the model.

- Model Evaluation: We will evaluate the performance of the model using various metrics such as accuracy, precision, recall, and F1 score. We will also perform cross-validation and use different evaluation techniques to ensure the robustness of the model.

- Writing Pattern Analysis: We will perform an analysis of the writing patterns in the text, such as sentence length, word frequency, and part-of-speech tags, to gain insights into the author's writing style.

- Result Interpretation: We will interpret the results of the model and writing pattern analysis to draw conclusions about the authorship of the anonymous text and gain deeper insights into the author's writing style.

- Feature Selection: We will perform feature selection to identify the most important statistical features for authorship identification. This will help to reduce the dimensionality of the feature space and improve the performance of the model.

- Hyperparameter Tuning: We will use techniques such as grid search and random search to tune the hyperparameters of the machine learning algorithms. This will help to optimize the performance of the model and improve its accuracy.

- Ensemble Methods: We will explore the use of ensemble methods such as bagging and boosting to improve the performance of the model. This will involve combining multiple models to produce a more accurate and robust prediction.

- Visualization Techniques: We will use visualization techniques such as word clouds and scatterplots to gain insights into the writing

patterns and identify any trends or patterns that may be present in the data.

- Error Analysis: We will perform an error analysis of the model to identify any cases where the model fails to correctly identify the author of the text. This will help to identify areas for improvement and inform future research.

- Comparison with Existing Methods: We will compare the performance of our proposed methodology with existing methods for authorship identification, such as stylometry and attribution studies. This will help to evaluate the effectiveness of our approach and provide insights into the strengths and limitations of different methods.

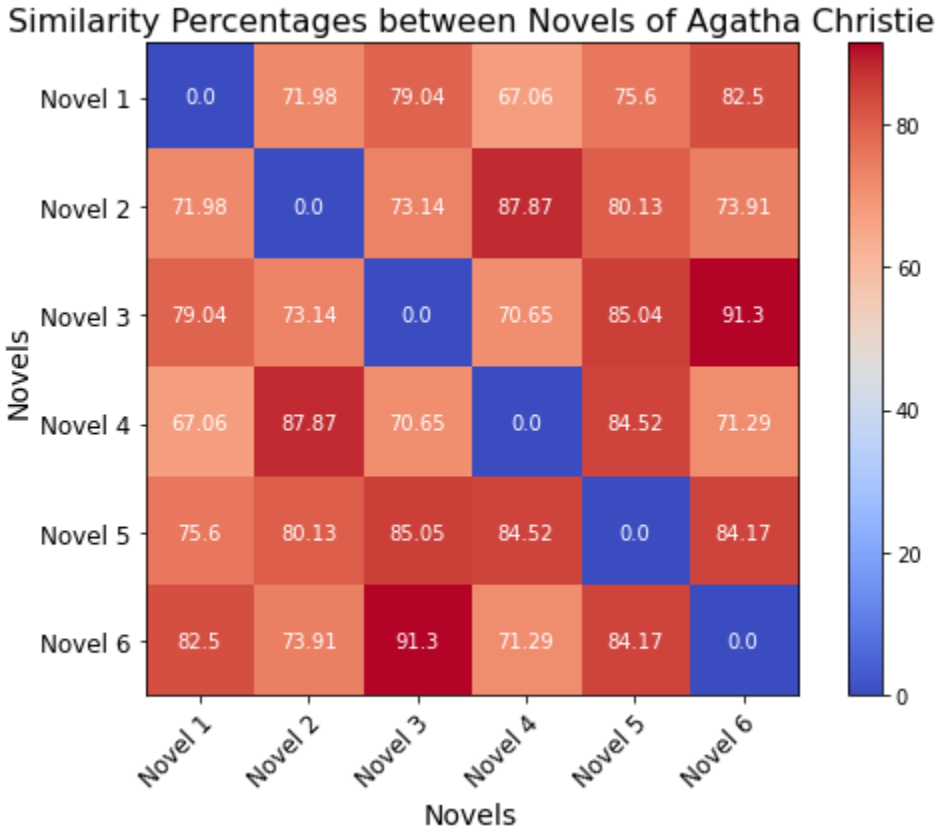
Results and Analysis

Cosine Similarity: When using a model for author identification, the similarity percentage can be used to determine the likelihood that two documents were written by the same author. If the similarity percentage is high, it suggests that the two documents are likely to have been written by the same author, while a low similarity percentage indicates that they were probably written by different authors.

However, it's important to note that a high similarity percentage does not necessarily guarantee that the documents were written by the same author.

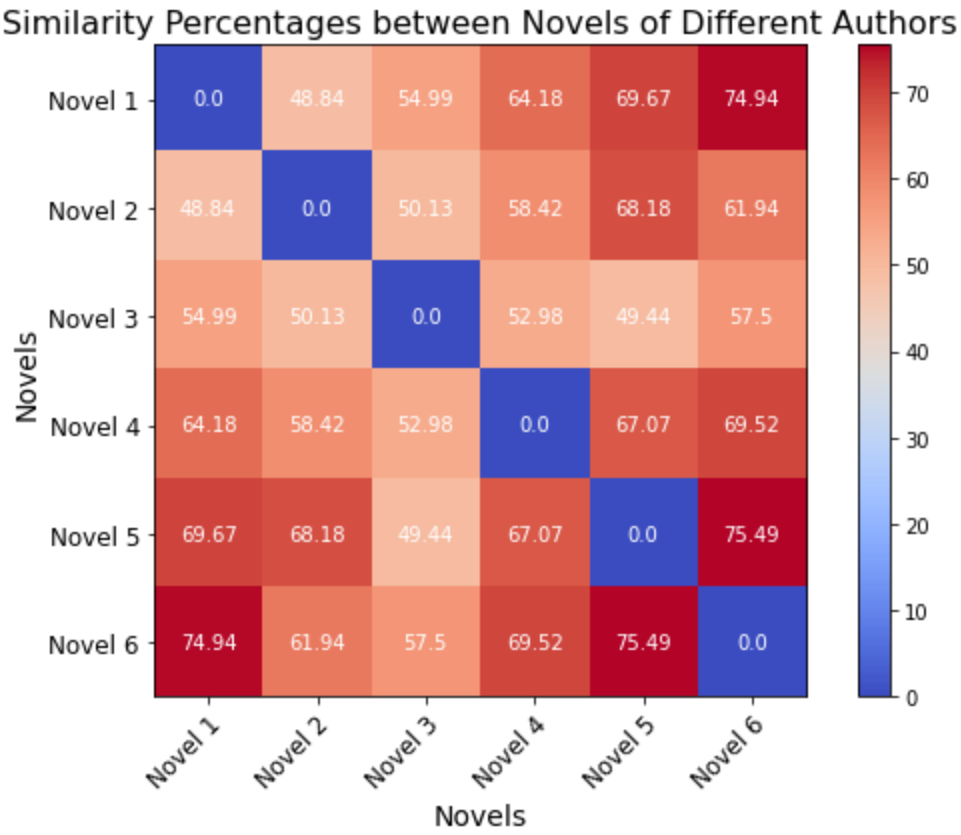
[5] uses Cosine Similarity over embedded vectors with accuracy of results of 66.6%.

Here is a result of similarity percentage check between Novels of Agatha Christie:



The average similarity between the novels of author Agatha Christine is 78.54%.

Then we take 6 novels by 6 different authors, and perform Cosine Similarity on them. The result is shown in the form of a heatmap below.



The average similarity between these 6 novels of different authors is 61.55%.

From the two averages, we can conclude that the similarity percentage between the works of Agatha Christie is much more than that of works of different authors. This shows that, similarity percentage is a good feature in order to determine the author of an anonymous work.

Length Features: Next the model extracts features based on the length of the works, we extract the Word Count, Sentence Count and the Average Sentence Length. These features give an idea about how long does the author write, how long is each sentence in the text.

The Word Count, Sentence Count and Average Sentence Length of 6 works of Agatha Christie is given below:

	Word Count	Sentence Count	Average Sentence Length
Novel 1	72,946	5,418	13.46
Novel 2	97,883	8,080	12.11
Novel 3	80,828	4,565	17.70
Novel 4	93,905	7,513	12.49
Novel 5	89,970	6,715	13.39
Novel 6	80,320	5,915	13.57

From the above tabular data, we can conclude that in Agatha Christie's works the Average Word Count is 85,975 words per novel, the Average Sentence Count is 6,367 sentences per novel whereas the Average Length of each sentence is 13.78.

After that we compute the Length features for the six different works of six different works, the results for the same are shown below:

	Word Count	Sentence Count	Average Sentence Length
Novel 1	86,021	6,450	13.33
Novel 2	2,05,557	6,955	29.55
Novel 3	3,047	184	16.55
Novel 4	8,472	322	26.31
Novel 5	2,58,237	12,241	21.09
Novel 6	95,251	3,482	27.35

From the table above, it is quite evident that for different authors the Word Count, Sentence Count and Average Sentence Length vary by a large extent.

In [6] the researcher makes use of these Length based features for the analysis.

Sentiment Analysis: Sentiment analysis is a natural language processing technique that involves determining the emotional tone or polarity of a piece of text. In the context of author identification, sentiment analysis can be used to extract information about the emotions or attitudes expressed by an author in their writing.

One potential use of sentiment analysis in author identification is to help distinguish between different authors who may have similar writing styles or vocabularies. By analyzing the emotional tone of the text, it may be possible to identify patterns or tendencies that are unique to a particular author. For example, one author may tend to use more positive or negative language, or may express particular emotions more frequently than another author.

Again with the help of a table we will see the Sentiment Analyis of six works by Agatha Christie:

	Negative	Neutral	Positive	Compund
Novel 1	0.103	0.786	0.111	1.0
Novel 2	0.087	0.804	0.109	1.0
Novel 3	0.108	0.782	0.111	0.9998
Novel 4	0.081	0.801	0.118	1.0
Novel 5	0.081	0.795	0.124	1.0
Novel 6	0.093	0.788	0.119	1.0

From the table we can conclude that Agatha Christie's sentiment while writing is Neutral, the average value of Neutral Sentiment for the six works is 0.792.

[7] proposes a method for authorship attribution in online reviews that uses sentiment analysis to extract emotion-related features from the text.

Stylistic Features: Alliteration count is a feature that can be used in author identification models in natural language processing (NLP). Alliteration refers to the repetition of the same sound or letter at the beginning of adjacent or closely connected words. Rhyme count refers to the number of times that an author uses rhyme in their writing.Repetition count, on the other hand, refers to the number of times that an author repeats words or phrases in their writing. Repetition is another literary device that can be used to emphasize certain ideas or themes in a text. By counting the number of times that an author repeats specific words or phrases, we can capture another aspect of their writing style.

Let us see the results when we compute these features on the six works of Agatha Christie:

	Alliteration Count	Rhyme Count	Repetition Count
Novel 1	2,916	455	86
Novel 2	3,742	579	164
Novel 3	3,242	376	12
Novel 4	5,150	655	259
Novel 5	3,549	490	60
Novel 6	3,483	516	76

From these we can determine that in Agatha Christie's Novels the Average Alliteration Count is 3,680. The Average Rhyme Count is 511 rhymes per novel and the Average Repetition Count is 109 repetitions per novel.

LSA (Latent Semantic Analysis) vector feature is a technique that can be used in author identification models in natural language processing (NLP). LSA is a statistical method that aims to capture the latent semantic structure of a text by identifying patterns of word co-occurrence.

In the context of author identification, LSA can be used to represent the semantic content of a text as a vector of numerical features. The LSA vector feature can be used to capture the underlying meaning or topic of a text, and can be compared across different texts to identify patterns or tendencies that are unique to a particular author.

The LSA Vector Count of six works of Agatha Christie is as follows:

	Novel 1	Novel 2	Novel 3	Novel 4	Novel 5	Novel 6
LSA Vector Count	1071.13	1625.70	1138.97	1585.03	1532.22	1053.16

The average LSA Vector Count of Agatha Christies's works is 1334.37.

Then we compute the LSA Vector Count of different works by different authors, so that we can compare the difference between them, the data is shown in the table below:

	Novel 1	Novel 2	Novel 3	Novel 4	Novel 5	Novel 6
LSA Vector Count	1695.96	1828.14	58.78	122.15	3329.66	1309.11

The lowest LSA Vector Value is 58.78.

The difference between the lowest LSA Vector Value by another author and the Average LSA Vector Value by Agatha Christie is 1275.59, The LSA vector values of two documents can be compared to identify similarities or differences in their underlying semantic content. A large difference, such as the difference of 1275.59 in this case, suggests that the two documents have very different semantic content and may be related to different topics or themes.

Type-token ratio (TTR) is a feature that can be used in author identification models in natural language processing (NLP). TTR is a measure of lexical diversity that indicates the ratio of unique words (types) to total words (tokens) in a text.

In the context of author identification, TTR can be used to capture aspects of an author's writing style, such as their vocabulary and their use of complex or rare words. An author with a high TTR may have a more diverse and varied vocabulary, whereas an author with a low TTR may tend to use the same words or phrases repeatedly.

Let us see the TTR Values of works by Agatha Christie:

	Novel 1	Novel 2	Novel 3	Novel 4	Novel 5	Novel 6
TTR Value	0.081	0.081	0.091	0.080	0.071	0.092

The Average TTR Value of Agatha Christie's works is 0.082.

	Novel 1	Novel 2	Novel 3	Novel 4	Novel 5	Novel 6
TTR Value	0.072	0.070	0.308	0.215	0.049	0.096

If we analyze the TTR values, we can conclude that the Average TTR Value of Agatha Christie's works is much less than that of a different author who has a TTR Value of 0.308.

There are several other features that can be extracted and analysed to train a model to identify the author of a text document, the next part of the project will have better results and advanced techniques and algorithms on which the model will be trained.

References:

[1]Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2009). Computer-based authorship attribution without lexical measures. Computers & the Humanities, 43(1), 143-157.

[2]Koppel, M., Schler, J., & Argamon, S. (2005). Authorship attribution in the wild. Language Resources and Evaluation, 39(1), 81-102.

[3]Kestemont, M., Daelemans, W., & Van Vaerenbergh, L. (2013). Authorship attribution using support vector machines and multiple linguistic features. Digital Scholarship in the Humanities, 28(3), 461-477.

[4]Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2003). Automatically profiling the author of an anonymous text. Communications of the ACM, 46(2), 121-127.

[5] "Author Identification: Using Text Mining, Feature Engineering \& Network Embedding," by A. a. S. S. Adhikari

[6] "Research on Author Identification Based on Deep Syntactic Features," by C. a. S. W. a. L. L. a. D. C. a. Z. X. Zhao, 2017 10th International Symposium on Computational Intelligence and Design (ISCID), pp. 276-279, 2017.

[7] "Authorship attribution using sentiment information from online reviews" by M. A. Al-Ghassani and S. S. Al-Riyami (2013)