Statistical Authorship Attribution And Analysis Of Agatha Christie's Works

Adam Ryan McDaniel, Ahmad Amirivojdan , Vincent Gregory Broda , Mohamed Shatarah
EECS UTK, Knoxville, Tennessee
amcdan23@vols.utk.edu , aamirivo@vols.utk.edu , vbroda@vols.utk.edu , mshatara@vols.utk.edu

***Abstract*—This document is our report for Project #1 of COSC 524, Natural Language Processing. It outlines our statistics based approach to performing authorship attribution to given works, specifically targeting Agatha Christie.**

***Index Terms*—statistics, NLP, tokenization, supervised ML, unsupervised ML**

I. Introduction

Authorship attribution is a crucial task in the field of natural language processing, where the author of a given text is determined based on linguistic and stylistic patterns [3]. This project focuses on developing an authorship analysis model specifically for Agatha Christie, a famous crime and mystery novelist. The primary purpose of this project is to construct a statistical model capable of attributing authorship to Agatha Christie among a set of other arbitrary authors. The dataset used for training included 4 other authors: Maurice Leblanc, G. K. Chesterton, Lewis Carroll, and Herman Melville. The texts were sourced from Project Gutenberg, providing a diverse dataset for analysis. By analyzing stylistic and linguistic features extracted from the texts, the aim is to capture the unique aspects of Christie's writing style. Machine learning techniques such as logistic regression and clustering algorithms are employed to differentiate her works from others.

II. Methodology

A total of 64 novels from Project Gutenberg were collected to form the dataset. This included 12 novels by Agatha Christie and 52 novels by other authors: Maurice Leblanc, G.K. Chesterton, Lewis Carroll, and Herman Melville. Table II-C shows the novel counts for each author. These authors were chosen due to similarities in genre and style, providing a challenging dataset for authorship attribution. To better handle lengthy texts, the novels were split into chapters, and further into 100-sentence-long chunks. This segmentation improved the accuracy of the model by allowing finer feature analysis.

A. Data Preprocessing

To prepare the texts for feature extraction, standard text-cleaning steps were performed. This included:

1) **Tokenization:** Splitting text into sentences and words.
2) **Normalization:** Lowercasing and splitting punctuation into separate tokens to standardize text.
3) **Chunking:** Splitting novels into 100-sentence sections to handle lengthy texts effectively. To predict the author of a text, a majority vote of predictions for each chunk in the text is taken.

B. Feature Extraction

A variety of textual and stylometric features were extracted for each text or partial chunk of a whole text:

- Textual Features
  1) **N-grams:** Unigrams, bigrams, and trigrams captured frequently used word sequences, helping to identify patterns in word choice and syntax. [3]
  2) **TF-IDF**: Terms were weighted by their importance across all novels to prioritize unique vocabulary for each author. [2]

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\text{IDF}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

- Stylometric Features
  1) **Sentence Length Variation:** The variance of all the sentence lengths in the text, in terms of number of tokens. [4]
  2) **Word Length Variation:** The variance of all the word lengths in the text, in terms of number of characters. [3]
  3) **Average Sentence Length:** The average number of tokens in a sentence.
  4) **Average Word Length:** The average number of characters in a word.
  5) **Punctuation Frequency:** The frequency of punctuation marks in the text: commas, periods, exclamation marks, and question marks.
  6) **Lexical Diversity:** The ratio of unique words to total words in the text.
  7) **Adverb Frequency:** The frequency of adverbs in the text.
  8) **Contraction Density:** The ratio of contractions to total words in the text.
  9) **Pronoun Density:** The ratio of pronouns to total words in the text.
  10) **Quote Density:** The concentration of quotes in the text.
- Advanced Features

1) **Passive Voice Frequency:** The frequency of passive voice constructions in the text.
2) **Estimated Grade Level:** A number representing the estimated grade level required to understand the text. This is a commonly used readability metric called the Flesch-Kincaid Grade Level [1].

$$g = 0.39\left(\frac{\text{Words}}{\text{Sentences}}\right) + 11.8\left(\frac{\text{Syllables}}{\text{Words}}\right) - 15.59$$

These features were chosen to reflect both the content and the style of the authors, aiming to capture subtle differences in writing habits.

C. Experiments

The corpus is composed of the following works:

Author	Number of Complete Novels
Agatha Christie	12
Maurice Leblanc	17
G.K. Chesterton	26
Lewis Carroll	4
Herman Melville	5

TABLE I
NUMBER OF WORKS BY AUTHOR

The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing. The experiments were executed using Python in a Jupyter Notebook, with a Python 3.12 environment. The machine used has an M3-Max CPU with 16 single-threaded cores and 128 gigabytes of unified memory. The following experiments were conducted:

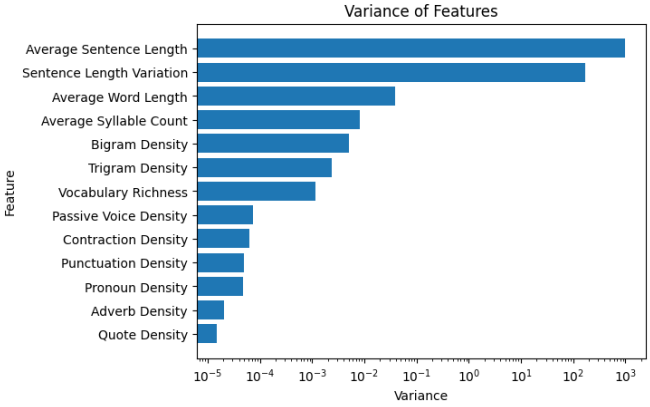
1) Variance Analysis Experiment: Each feature was analyzed for variance across the authors to determine its effectiveness in distinguishing between them. The distribution of different features were graphed in 3-dimensions to visualize the clusters and discriminatory power of each dimension.

2) K-Means Clustering Experiment: K-means clustering was performed on the extracted features to investigate whether the texts naturally group according to authorship without using labels, as clustering methods have proven effective in stylistic analysis. [4] Then, the greedy register allocation algorithm was repurposed to assign clusters to their appropriate authors. The model was trained on the entire novels and evaluated on the validation set.

3) Logistic Regression Binary Classification Experiment: The first supervised model was a logistic regression model, using TF-IDF to distinguish texts authored by Agatha Christie and those by other authors. The model was also trained on the feature vectors extracted, and then evaluated on the validation set as well.

4) Support Vector Machine Multiclass Classification Experiment: The second supervised model was an extended iteration of the logistic regression model, using a support vector machine (SVM) instead to classify texts into one of the five authors. The model was trained only on sentence chunks in order to identify authors, and was then tested against the validation set.

Fig. 1. Variance of Features Across Authors



5) Binary Decision Tree Classification Experiment: The final supervised model was a decision tree classifier, using the same features as the logistic regression model. The model was trained only on the chunks of sentences, and then additionally evaluated on the validation set.

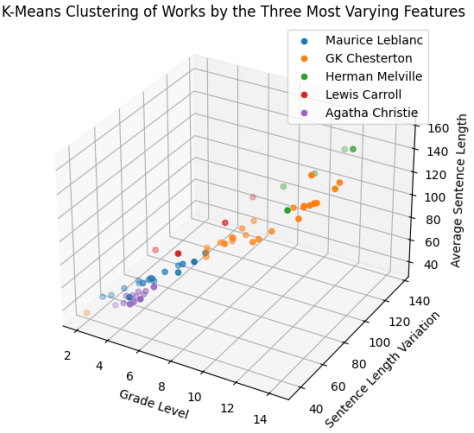
III. Results

A. Variance Analysis Experiment

The variance of each feature from author to author is plotted in a bar-chart in Figure 1. Few features have massive variance, while the majority of features vary only slightly between authors. The most important features were determined to be average sentence length, sentence length variation, and average word length.

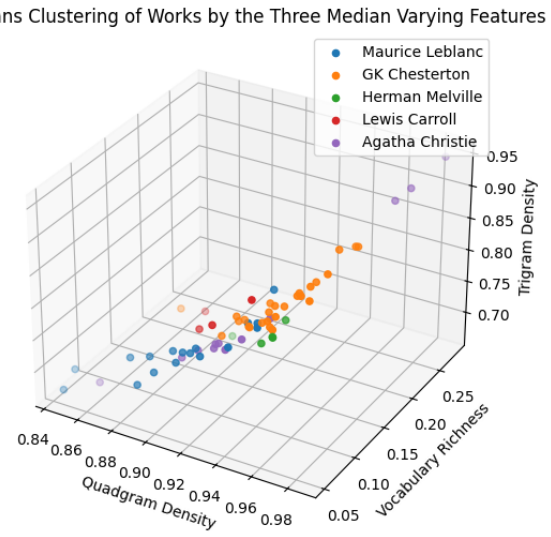
The three top features were plotted in 3D space to visualize the clusters formed by the authors. Figure 2 shows the clusters formed by the authors in the feature space. There is usually a clear separation between authors, indicating that these metrics are useful for distinguishing between them.

Fig. 2. 3D Plot of Top Features



Next, the three median features were plotted in 3D space to visualize the clusters formed by the authors. Figure 3

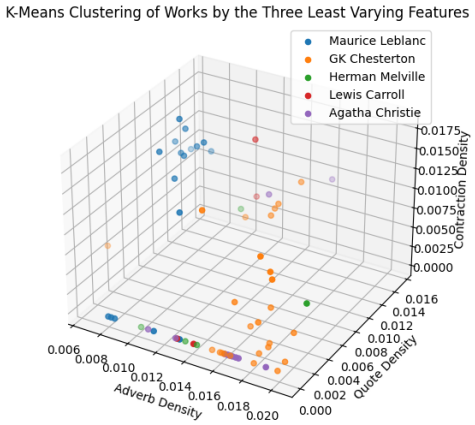
Fig. 3. 3D Plot of Features With Median Importance



shows the same visualization as Figure 2, but with the median features instead of the top features. There is much less clear separation between these groups, showing that these features are less discriminative.

Finally, the lowest three important features were plotted. Figure 4 shows the same visualization as Figure 2, but with the lowest features instead of the top features. There is significant overlap between the groups, showing that these features are not useful at all for distinguishing between authors.

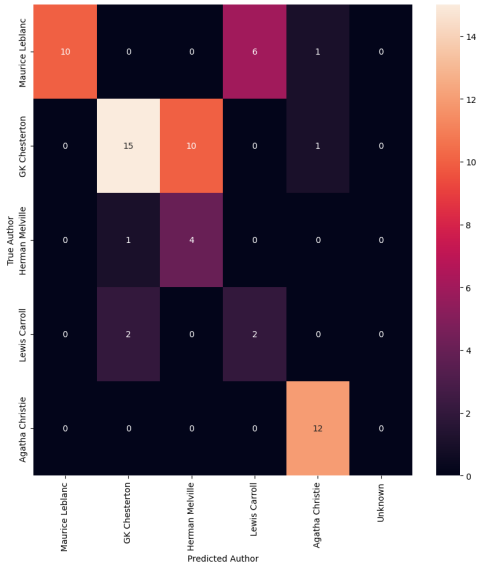
Fig. 4. 3D Plot of Features With Lowest Importance



B. K-Means Clustering Experiment

The K-means clustering model was able to cluster the feature vectors into five distinct clusters, corresponding to the five authors, however the classification for many works were very confused. Figure 5 shows the confusion matrix of the clustering. Agatha Christie's works were all correctly identified, but other authors were often misclassified. The overall accuracy across all works was 62.5%. This approach was inefficient and did not yield satisfactory results.

Fig. 5. K-Means Clustering Results



C. Logistic Regression Binary Classification Experiment

Classifying each text using TF-IDF and logistic regression yielded an accuracy of 100% for the custom validation dataset, distinguishing authors: the model was able to accurately predict whether Agatha Christie was the author of all the texts. Figure 6 shows the confusion matrix of the logistic regression model.

This method was successful at identifying texts, but the final, superior model was the decision tree based model for high explainability and reproducilbility while maintaining accuracy.

Fig. 6. Logistic Regression Binary Classification Results

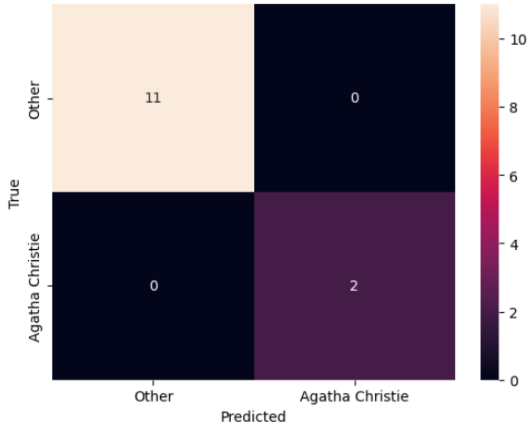
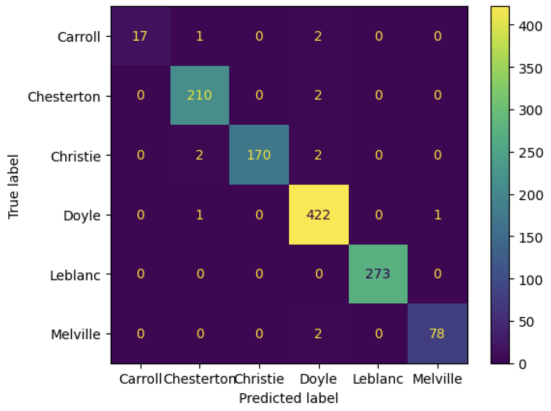


Fig. 7. SVM Multiclass Classification Results



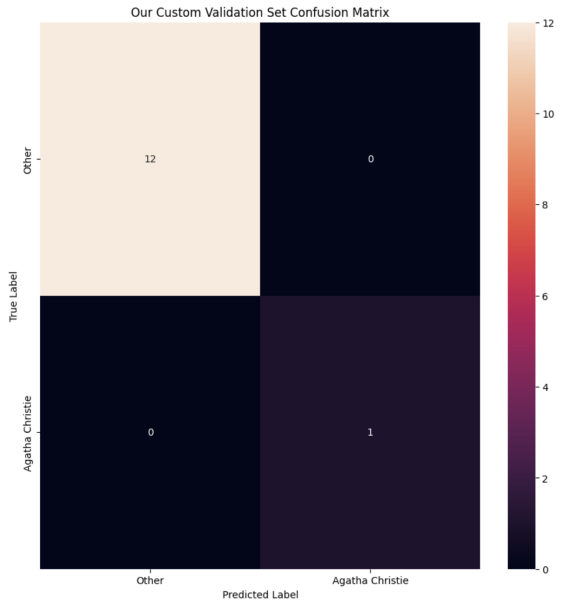
D. Support Vector Machine Multiclass Classification Experiment

The SVM model was trained only on sentence chunks and evaluated on the validation set. The model achieved an accuracy of 99% on the validation set, only misclassifiying 13 chunks out of over 1,000 total chunks. Figure 7 shows the confusion matrix of the SVM model.

E. Binary Decision Tree Classification Experiment

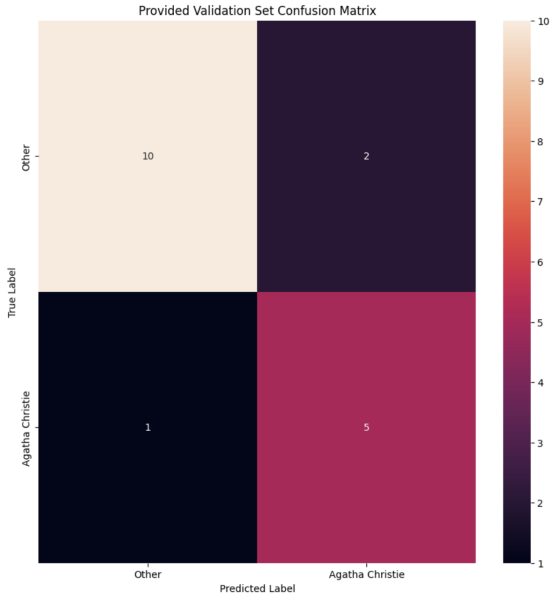
The chunk approach was adapted to perform prediction for a whole text. The work is chunked into blocks of N sentence, featurized, and a majority vote is taken based on the predicted classification of each chunk.

Fig. 8. Decision Tree Binary Classification Results On Custom Validation Set



The chunk size of 100 was found to be an ideal candidate for the chunk size: with larger chunk sizes, there are fewer sentences per chunk and their feature vectors aren't representative of the work's author. With chunks that are too small, the feature vectors are too noisy and don't represent the work's author, either. The decision tree model achieved an accuracy of 100% on the validation set, with no misclassifications. Figure 8 shows the confusion matrix of the decision tree model on the custom validation dataset. With the provided dataset from the class, the following confusion matrix was obtained for the decision tree model. The model achieved an accuracy of 100% on the validation set, with no misclassifications. Figure 9 shows the confusion matrix of the decision tree model on the provided dataset. This final model achieved 100% accuracy on the custom validation dataset, and 83.33% accuracy on the provided dataset. This final model was chosen for its explainability and equally high accuracy.

Fig. 9. Decision Tree Binary Classification Results On Provided Validation Set



IV. Conclusion

Several key features that distinguish Agatha Christie's writing style from other authors. Average sentence length, sentence length variation, and average word length are the most important features for prediction.

Our experiments demonstrate that supervised machine learning models, particularly decision tree and logistic regression models with TF-IDF features, are highly effective for authorship attribution in this context. The perfect accuracy on the custom validation set, and the high accuracy on the provided set, indicates that Agatha Christie's writing style is distinctly different from the other authors in the dataset.

It was revealed that unsupervised clustering may not be ideal for authorship attribution, as it yields lower precision and more confusion between authors. While still lower in accuracy, though, unsupervised clustering results still showed that stylometric features carry authorial signals.

Text	Predicted Author	Actual Author
Text #1	Not Agatha Christie	Agatha Christie
Text #2	Agatha Christie	Agatha Christie
Text #3	Agatha Christie	Agatha Christie
Text #4	Agatha Christie	Agatha Christie
Text #5	Agatha Christie	Agatha Christie
Text #6	Agatha Christie	Agatha Christie
Text #7	Not Agatha Christie	Arthur Conan Doyle
Text #8	Not Agatha Christie	Arthur Conan Doyle
Text #9	Not Agatha Christie	Arthur Conan Doyle
Text #10	Not Agatha Christie	Arthur Conan Doyle
Text #11	Agatha Christie	Arthur Conan Doyle
Text #12	Agatha Christie	Arthur Conan Doyle
Text #13	Not Agatha Christie	G.K. Chesterton
Text #14	Not Agatha Christie	G.K. Chesterton
Text #15	Not Agatha Christie	G.K. Chesterton
Text #16	Not Agatha Christie	G.K. Chesterton
Text #17	Not Agatha Christie	G.K. Chesterton
Text #18	Not Agatha Christie	G.K. Chesterton

TABLE II
PREDICTED AUTHORSHIP ATTRIBUTION OF AGATHA CHRISTIE ON PROVIDED DATASET

The decision tree model was demonstrated to be the most useful while maintaining effectiveness and interpretability. The method of featurizing N-sentence chunks of text with a majority vote proved to be the most effective method for authorship attribution.

The final model achieved 100% accuracy on the custom validation dataset, and 83.33% accuracy on the provided dataset. The model was able to accurately predict the author of Agatha Christie's works, and was able to distinguish her works from those of other authors.

In conclusion, this work successfully presents an explainable, reasonably accurate, and reproducible decision tree model for authorship attribution of Agatha Christie's novels.

REFERENCES

[1] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel. Technical Report ADA006655, Chief of Naval Technical Training, U. S. Naval Air Station, 1975.

[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[3] A. Misini, A. Kadriu, and E. Canhasi. A survey on authorship analysis tasks and techniques. *SEEU Review*, 17(2):153–164, 2022.

[4] A. Roelleke. Stylogenetics: Clustering-based stylistic analysis. Unpublished.