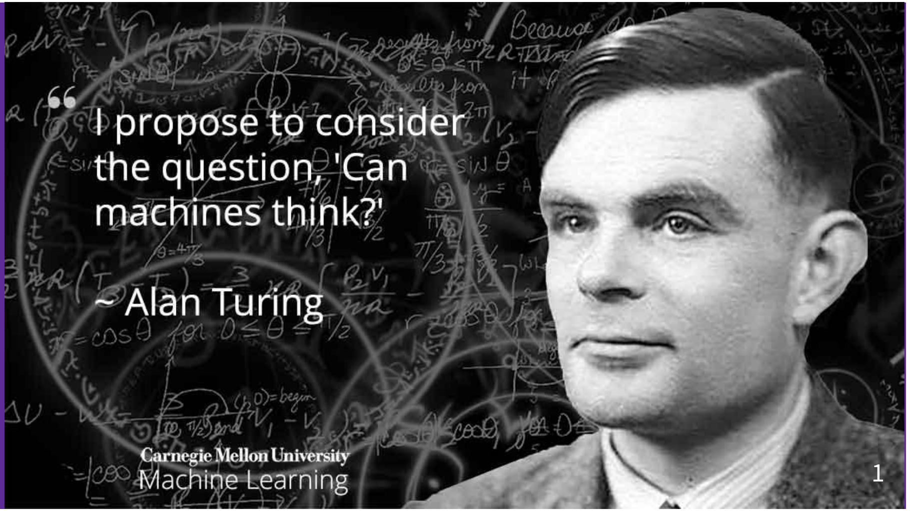
Natural Language Processing

Project 1: Authorship Attribution Using Statistical Models

Team 10

- Adam Ryan McDaniel
- Ahmad Amirivojdan
- Vincent Gregory Broda
- Mohamed Shatarah



Introduction

- Authorship attribution is the process of identifying the author of a text based on linguistic features.
- This project focuses on attributing crime novels to the target author, Agatha Christie, against other authors.
- This was achieved by extracting meaningful textual and stylometric features.
- Finally, we interpret the results of our various model approaches, their performances, and feature importances.

Dataset Preparation

- Data Source: Project Gutenberg
- Authors Included:
  - Agatha Christie (12 novels)
  - Maurice Leblanc (17 novels)
  - GK Chesterton (26 novels)
  - Lewis Carroll (4 novels)
  - Herman Melville (5 novels)

Data Preprocessing

- Data Splitting: Split novels into chapters for effective training.
- Cleaning: Removed punctuation, converted text to lowercase.
- Tokenization: Split text into words and sentences using custom functions.
- Feature Extraction: Calculated unique words, sentence lengths, word lengths, etc.
- Sentence Chunking: Chunks of sentences of varying lengths were split and processed as single cohesive units this way.

Feature Engineering

- Stylometric Features:
  - Sentence Length Variation: Measure the variability in sentence lengths
  - Word Length Variation: Analyze diversity in word lengths
  - Punctuation Frequency: Count usage of punctuation marks
- Advanced Features:
  - Passive/Active Voice: Measuring the frequency of passive voice usage in the text
  - Grade Level: Composite feature calculated by average syllables per word and word uniqueness
  - Adverb Density: Frequency of adverbs used
  - Pronoun Density: Usage rate of pronouns
  - Contraction Density: Prevalence of contractions in the text
- Textual Features:
  - N-grams: Capture common word sequences (bigrams, trigrams, etc.)
  - TF-IDF: Highlight important words by their frequency and uniqueness

Model Development

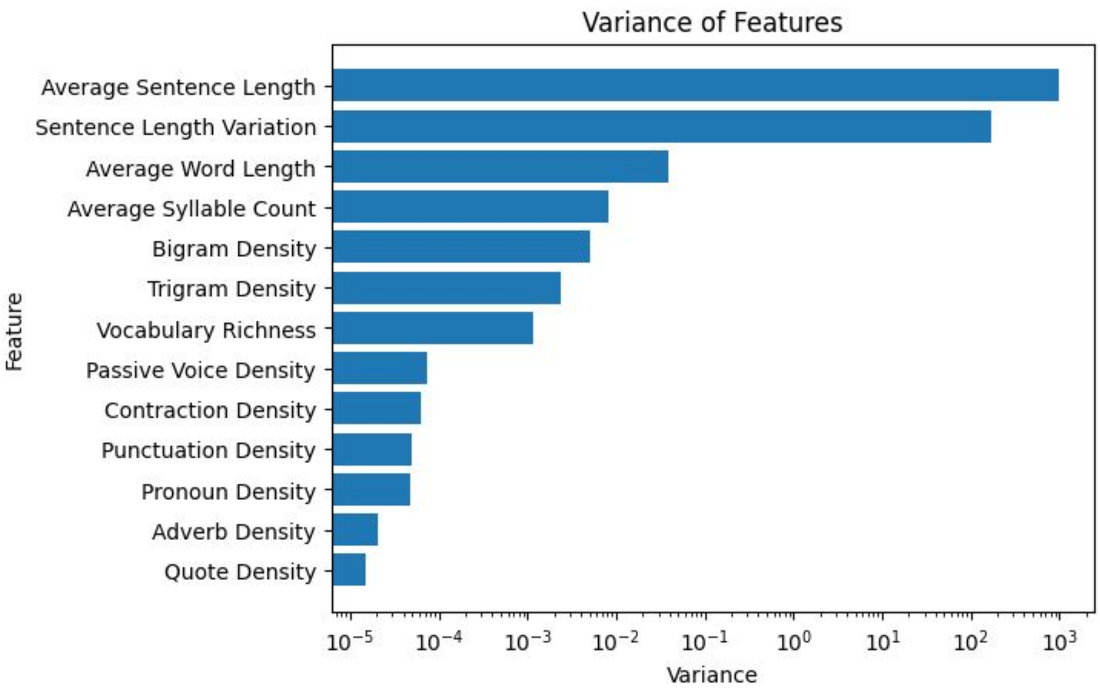
- Unsupervised Clustering
    - K-Means clustering on the features extracted from the text, assigning works to clusters. Authors are then attributed to each cluster based on the number of their works in each cluster.
    - Classification can be performed by checking if new works are bucketed in a given author's cluster.
- Supervised Classification
    - Attempted classification using Logistic Regression with TF-IDF features.
    - Attempted classification using SVM on 100-sentence chunks.
    - Finally, we modified the chunking approach to use a binary decision tree classifier as our final chosen method to classify the provided texts from the class. This method produced the best results.

Results

Feature Variance Analysis

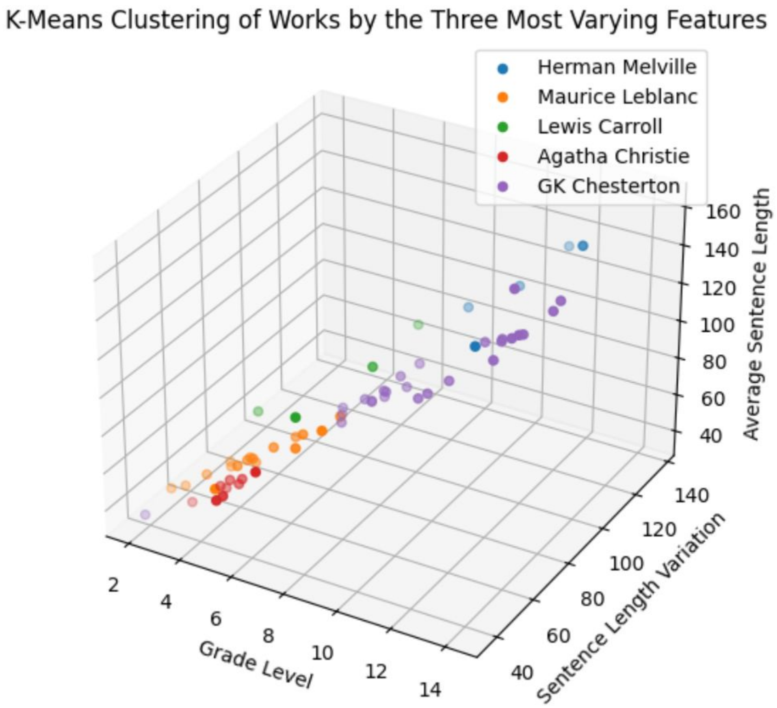
- First, we looked at how each feature varies between authors to identify which features contribute most to author prediction.

- Few features have massive variance, while the majority of features vary only slightly between authors.



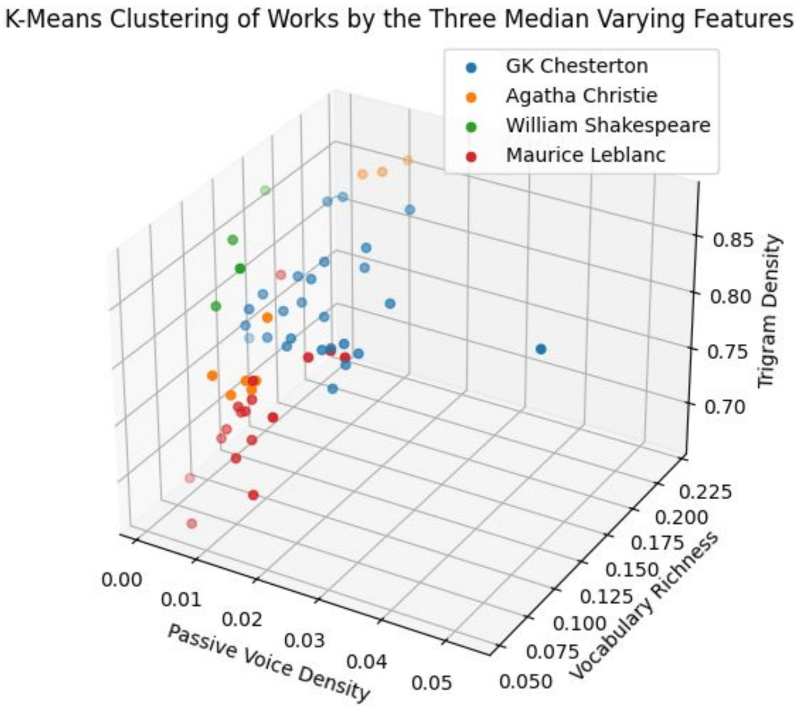
Visualization of Clusters (Most Varying Features)

- 3D Scatter Plot
  - Plotted clusters using the three features with the highest variance.
  - Colored data points by author.
- Observation
  - Some separation between authors, indicating stylistic differences.



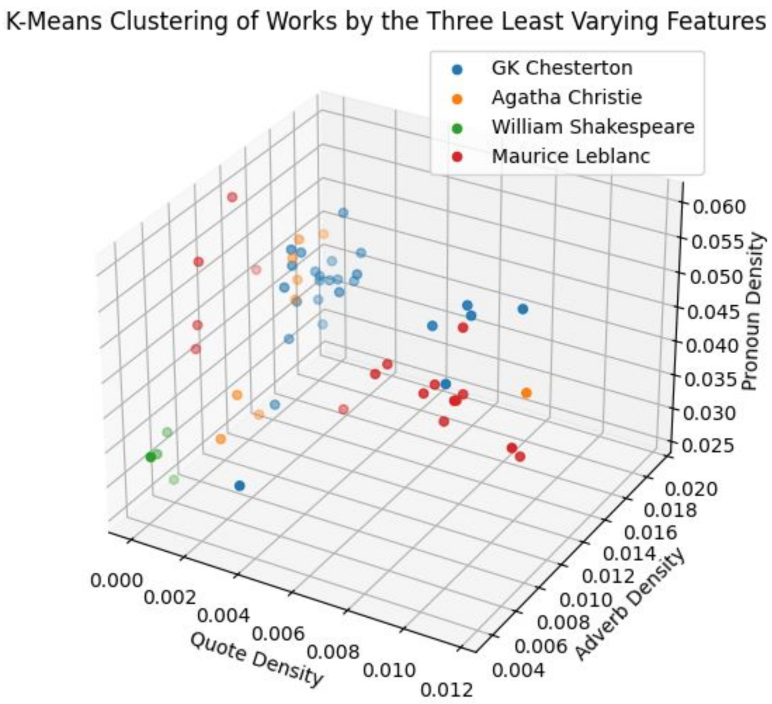
Visualization (Median Varying Features)

- 3D Scatter Plot
  - Plotted clusters using features with median variance.
- Observation
  - Less clear separation, suggesting these features are less discriminative.



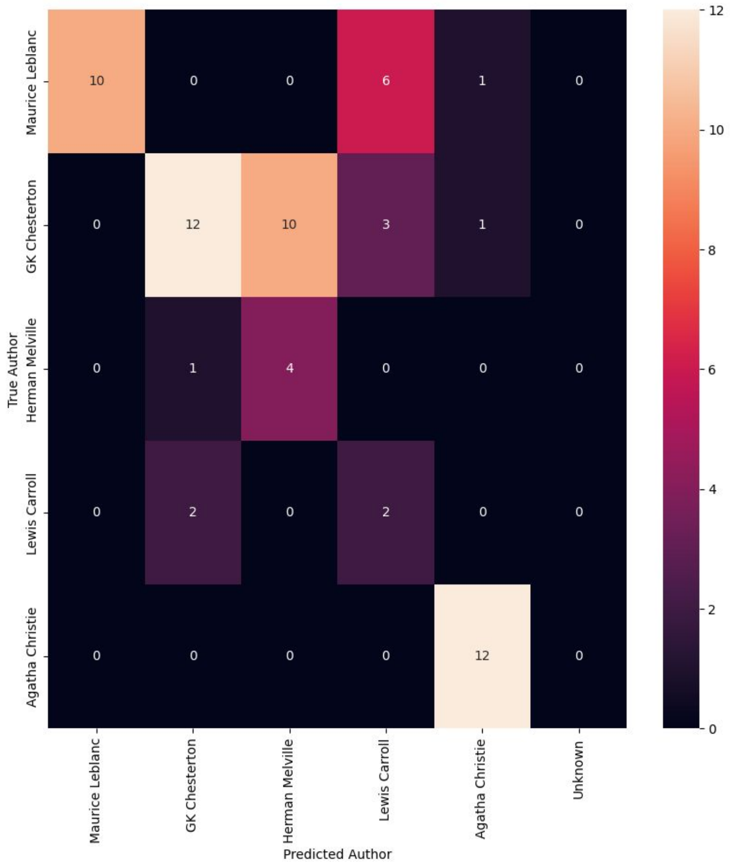
Visualization (Least Varying Features)

- 3D Scatter Plot
  - Plotted clusters using the least varying features.
- Observation
  - Overlapping clusters, indicating minimal discriminative power.



Unsupervised Clustering

- First, we used K-means clustering to group together novels based on the extracted features.
- Clusters represent predicted authors: authors are attributed to clusters greedily after training by looking at the number of works by a given author in each cluster.
- Achieved an accuracy of 62.50% using unsupervised learning, grouping novels with no author labels.
- This approach was inefficient and did not yield satisfactory precision and accuracy.



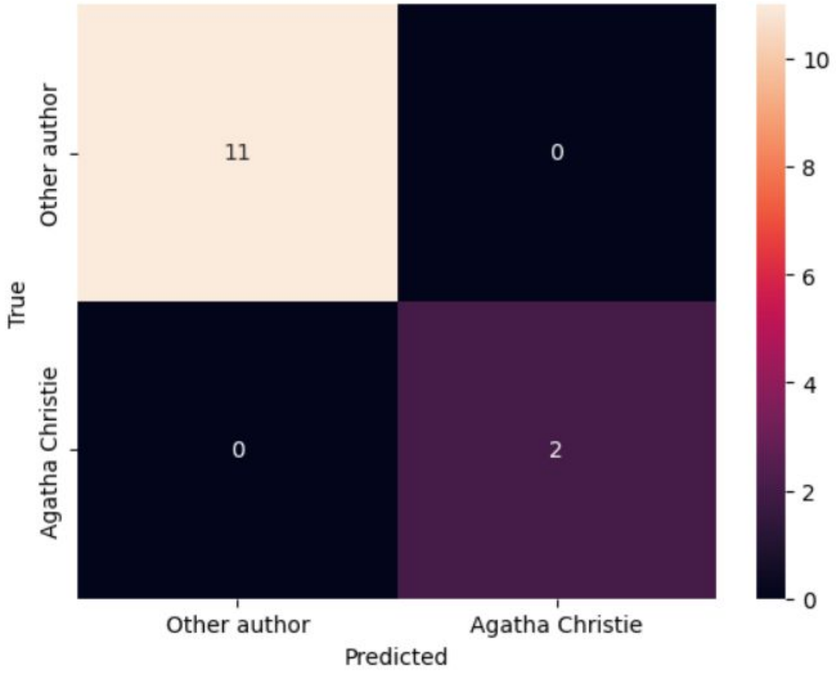
Analysis with First N and Last M Chapters

- Beginnings of novels are more distinctive for authorship attribution than endings.
    - Authors establish unique voice in opening chapters.
    - Conclusions may follow more standardized narrative structure.

First N	Last M	Multiclass Classification Logistic Regression	Unsupervised Clustering K-means
1	1	83.33%	28.33%
4	4	79.17%	58.33%
4	0	93.75%	63.33%
0	4	75.0%	50.0%
6	6	83.33%	50.0%
10	10	83.33%	50.0%

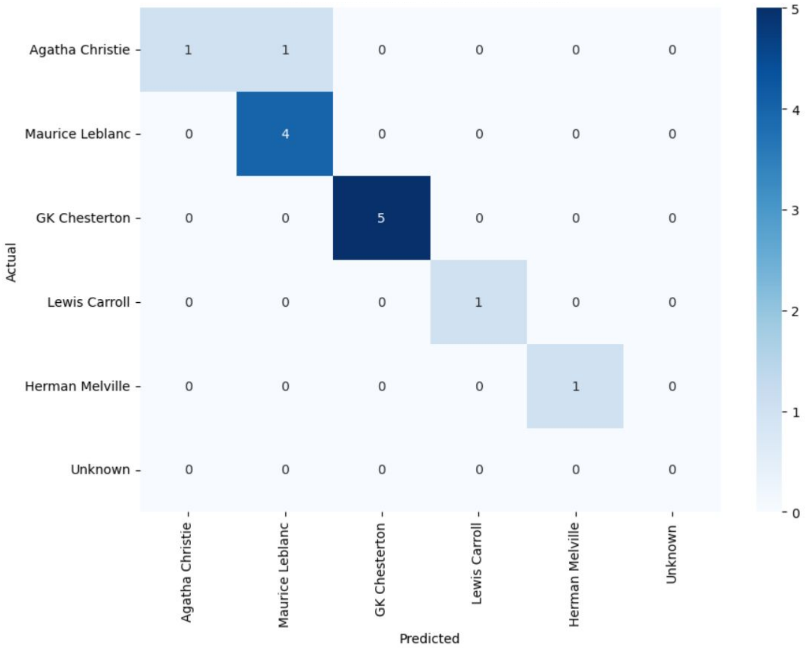
Initial Supervised Model Results

- Using a custom validation dataset, which is 20% of the total corpus we collected, we evaluated the performance of logistic regression.
- The entire text was featurized and classified at once, not in chunks.
- Achieved 100% accuracy on our custom validation dataset.
- This model is less explainable and achieves the same accuracy compared to our later approaches, however.



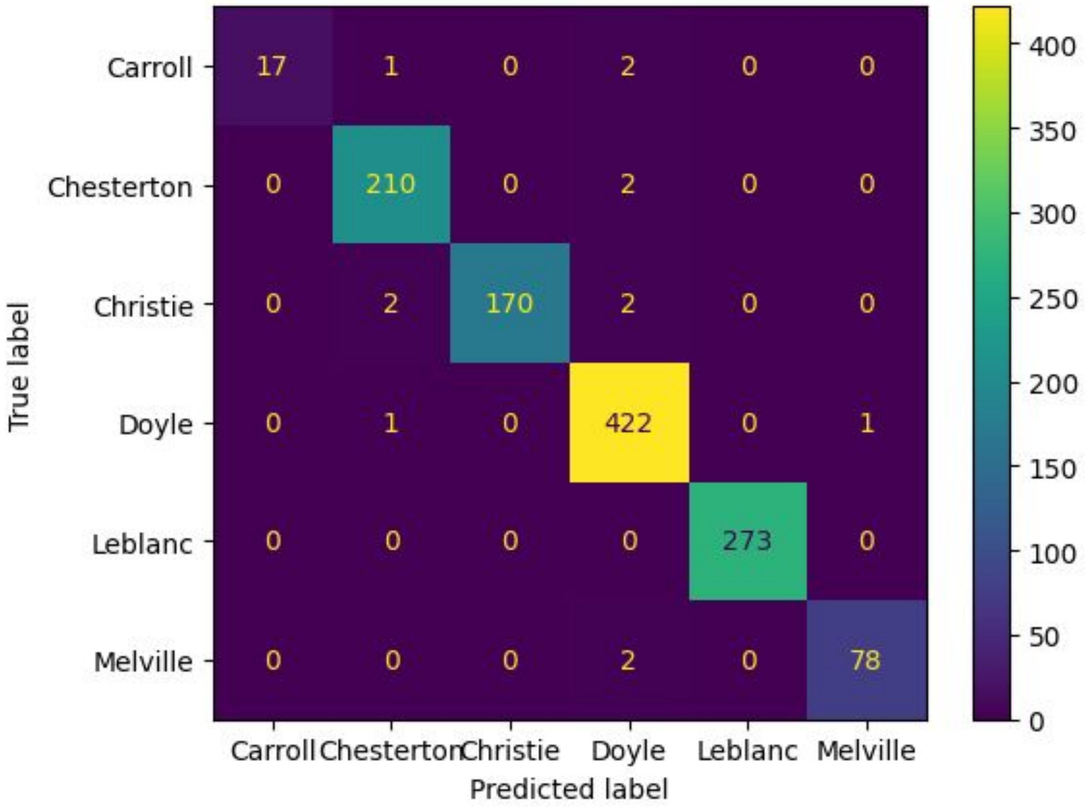
Extended Supervised Model (Multiclass Classification)

- We also attempted to used logistic regression to perform multiclass classification, instead of a binary approach.
- Predicted the exact author from a set of authors.
- Results
    - Achieved 91.84% accuracy on our validation set.
- This approach was better than the unsupervised approach, but there is still room for improvement.



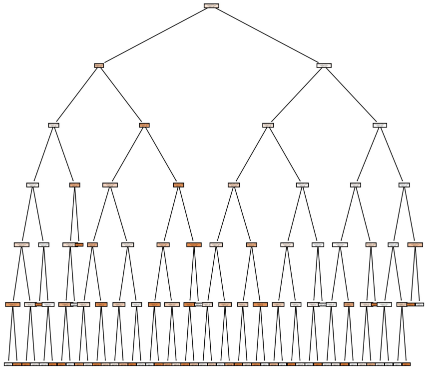
Multiclass Classification on 100-Sentence Chunks

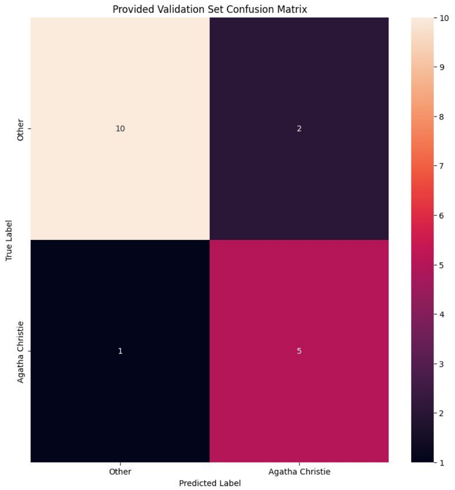
- The works were broken into chunks of N sentences each before being featurized.
- Results
  - Achieved 99% accuracy for classifying each individual chunk in our custom validation dataset.
- This approach was adapted to perform prediction for a whole text. The work is chunked into blocks of N sentences, featurized, and a majority vote is taken based on the predicted classification of each chunk.
- Using sentence chunks yielded much higher accuracy than our previous methods: our final result uses this strategy.

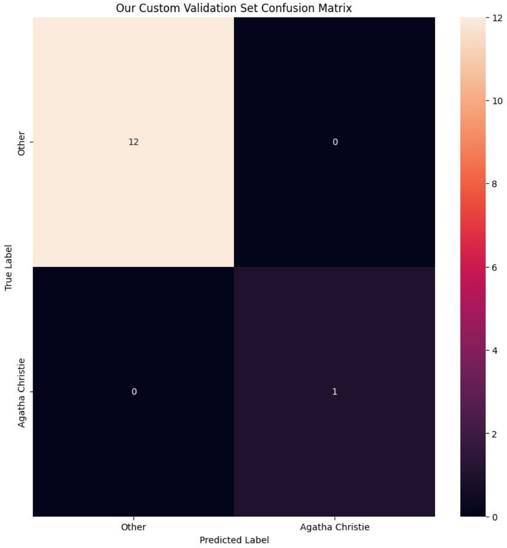


Our Final Model: Explainable Decision Trees

- A decision tree model was used to perform binary classification of each chunk in the text based on their features.
- A majority vote of all the chunks predictions is performed to obtain the final prediction for the work.
- 15/18 samples from the provided validation set were predicted correctly.
    - 83.33% accuracy on the validation set provided to us, along with 100% accuracy on our own validation dataset.
- The final decision tree model was chosen due to its explainability.







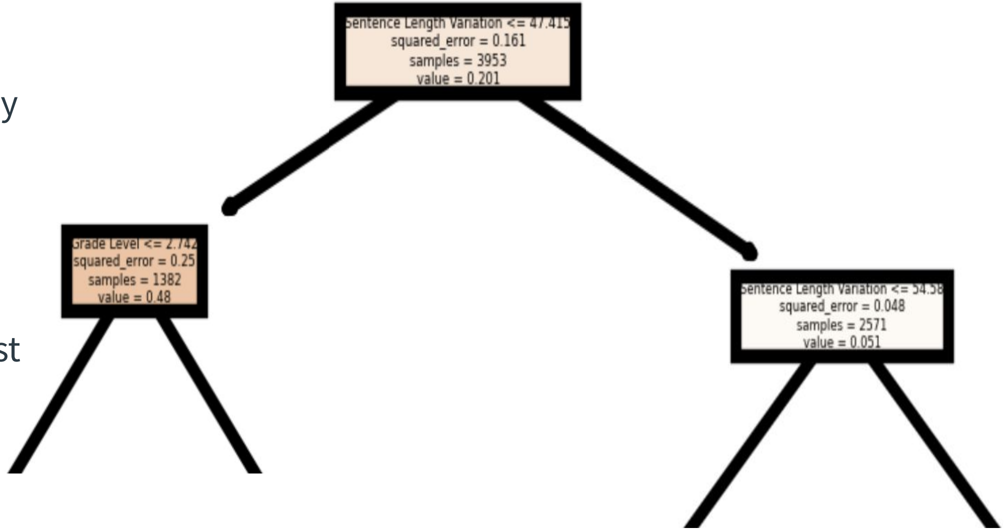
Validation Set Predictions

Work	Prediction	Work	Prediction
Text #1	**Not A.C.**	*Text #10*	**Not A.C.**
Text #2	**A.C.**	*Text #11*	**A.C.**
Text #3	**A.C.**	*Text #12*	**A.C.**
Text #4	**A.C.**	*Text #13*	**Not A.C.**
Text #5	**A.C.**	*Text #14*	**Not A.C.**
Text #6	**A.C.**	*Text #15*	**Not A.C.**
Text #7	**Not A.C.**	*Text #16*	**Not A.C.**
Text #8	**Not A.C.**	*Text #17*	**Not A.C.**
Text #9	**Not A.C.**	*Text #18*	**Not A.C.**

Decision Tree Investigation And Analysis

- The decision tree learned that sentence length variation was the most important feature to distinguish Agatha Christie’s works, followed by the estimated grade level.

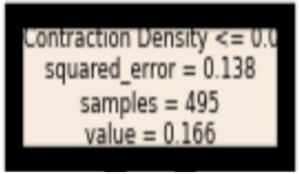
- This is expected, as these features were identified early on as the features with the most variances.

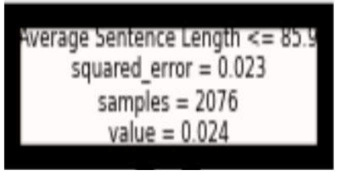


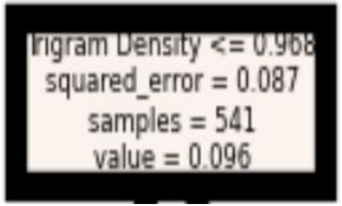
Decision Tree Investigation And Analysis

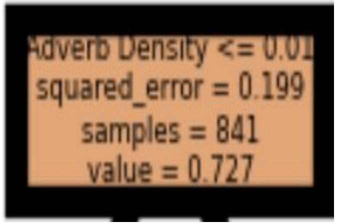
- At the middle layers of the tree, features like trigram density, adverb density and contractions become more prominent.

- Sentence length also remains relevant throughout the entire tree. This might be due to the tree learning to analyze and associate the sentence length in light of other feature values.





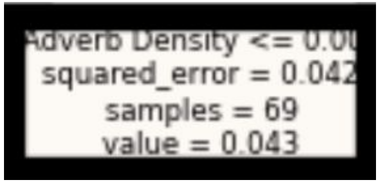


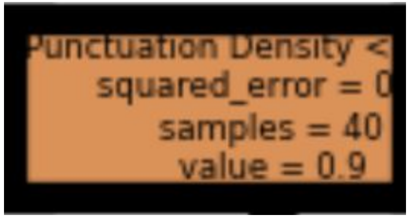


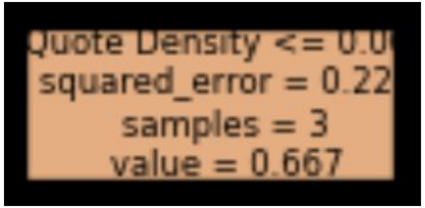
Decision Tree Investigation And Analysis

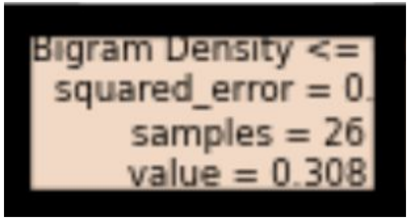
- At the lowest levels of the tree, quote density, adverb density, and bigram density are most prominent.

- This is expected, as these features vary little from author to author. So, their ability to discriminate authorship is small.









Conclusion

- We identified 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.
- We revealed that unsupervised clustering may not be ideal for authorship attribution, as it yields lower precision and more confusion between authors.
- We demonstrated the effectiveness and explainability of binary decision tree classification on featurized sentence chunks for authorship attribution.
- *We present an explainable, reasonably accurate, and reproducible decision tree model for classifying Agatha Christie's novels.*

Questions?