**Average Accuracy Scores per Classifier:**

*Based on 100 runs of the classifier, each time sampling 2000 chunks at random (even distributed between author and synthetic/authentic)*

|  |  |  |
| --- | --- | --- |
|  | **Mean Accuracy** | **Max Accuracy** |
| **Random Forest** | 0.99 | 1.0 |
| **Decision Tree** | 0.95 | 0.98 |
| **Support Vector Machine** | 0.91 | 0.95 |
| **Logistic Regression** | 0.85 | 0.89 |
| **Nearest Shrunken Centroid** | 0.73 | 0.8 |

**Features:**

* Mean length
* VADER sentiment score
* Female pronouns
* Male pronouns
* Type-token ration
* Lexical density
* Concreteness
* Relative frequency of 100 most common stop words in the corpus overall
* Relative frequency of 100 most common content words in the corpus overall

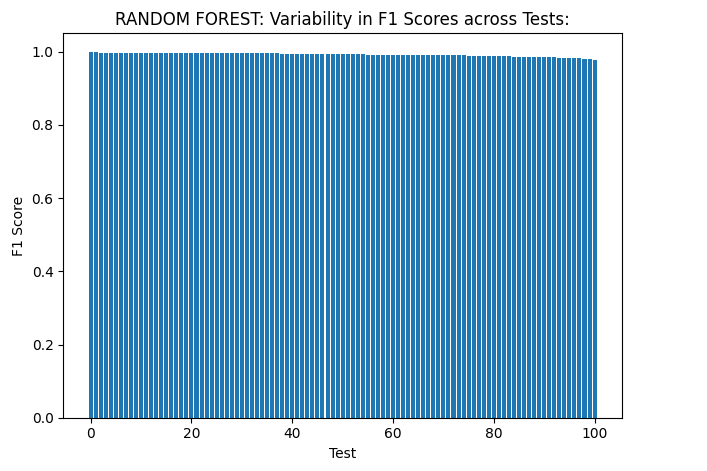
207 features total

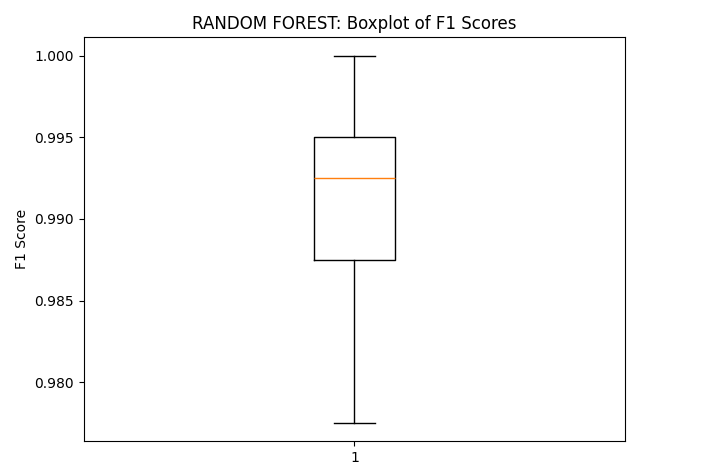
**Corpus:**

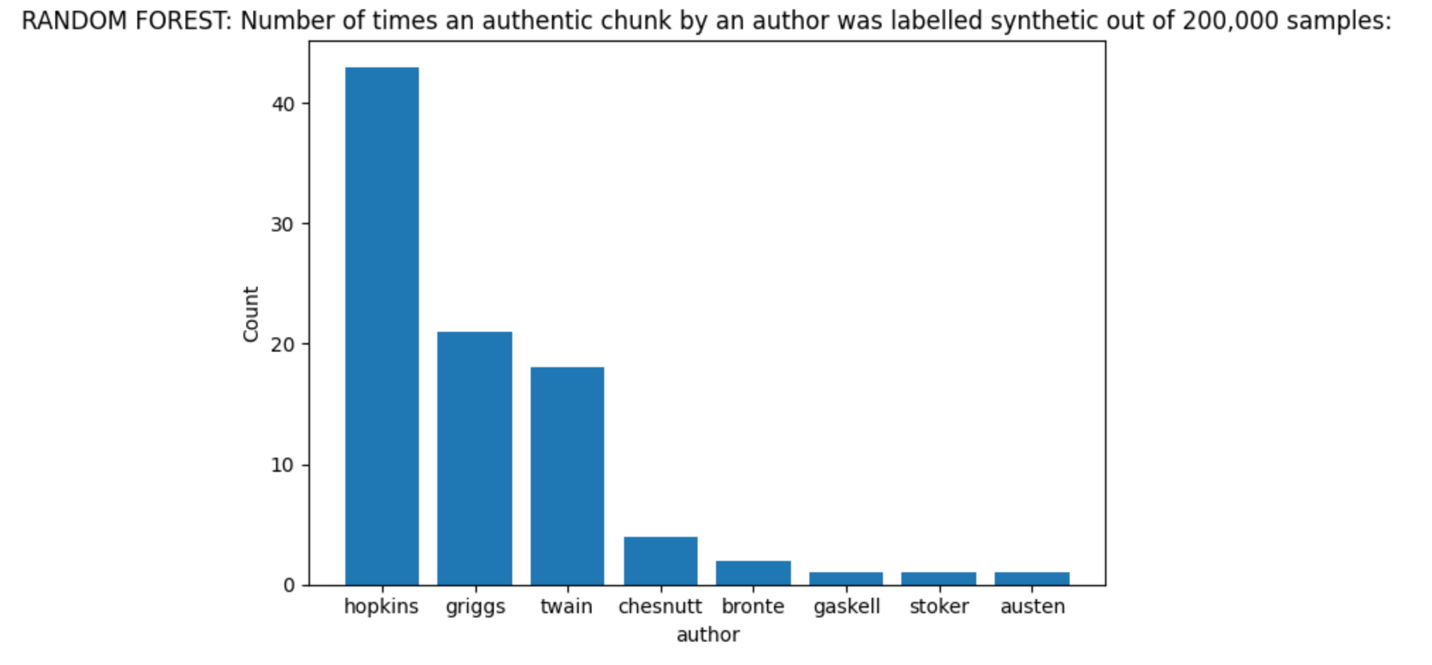
For each run of the classifier I randomly sampled:

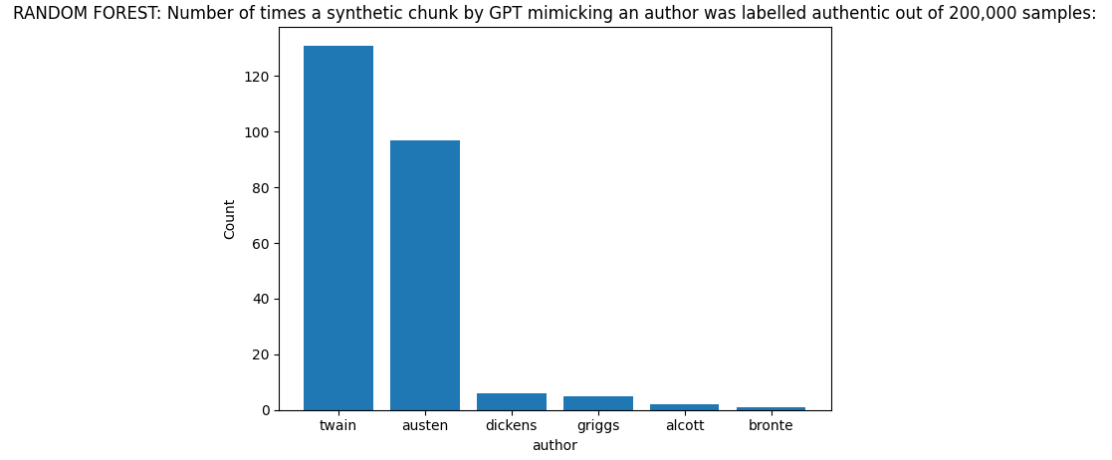
* 2000 ‘chunks’ of text
* 100 chunks for each author, for both synthetic and authentic
* Chunks contain 500 tokens

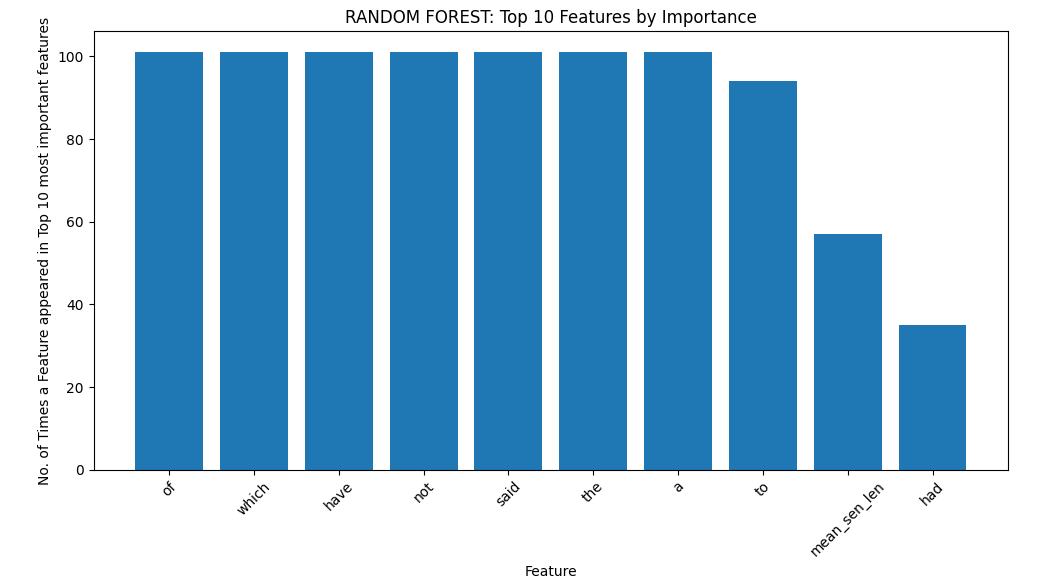
**RANDOM FOREST VISUALIZATIONS**

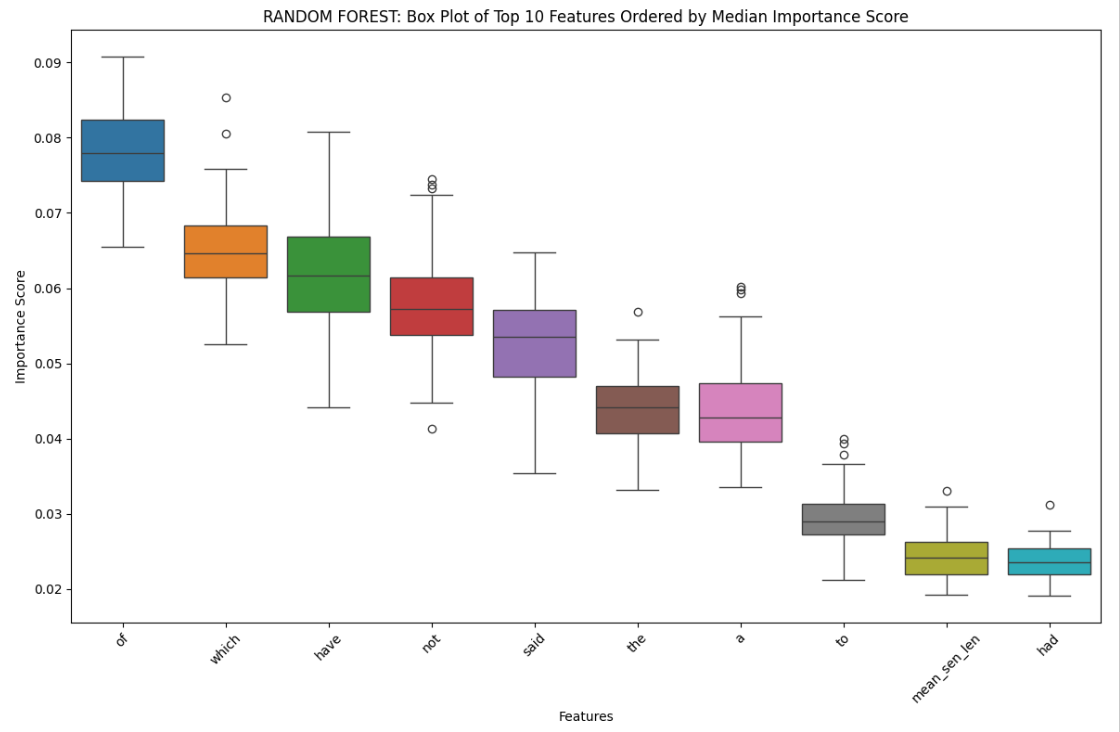




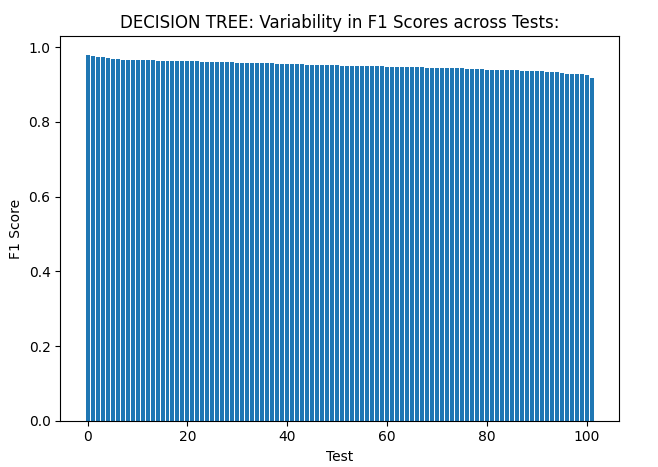


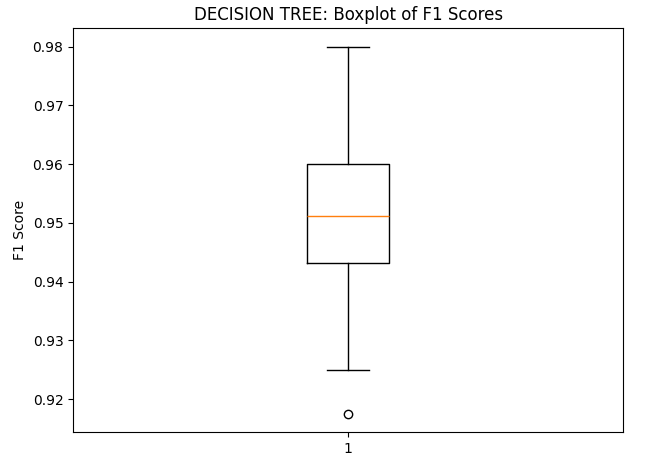


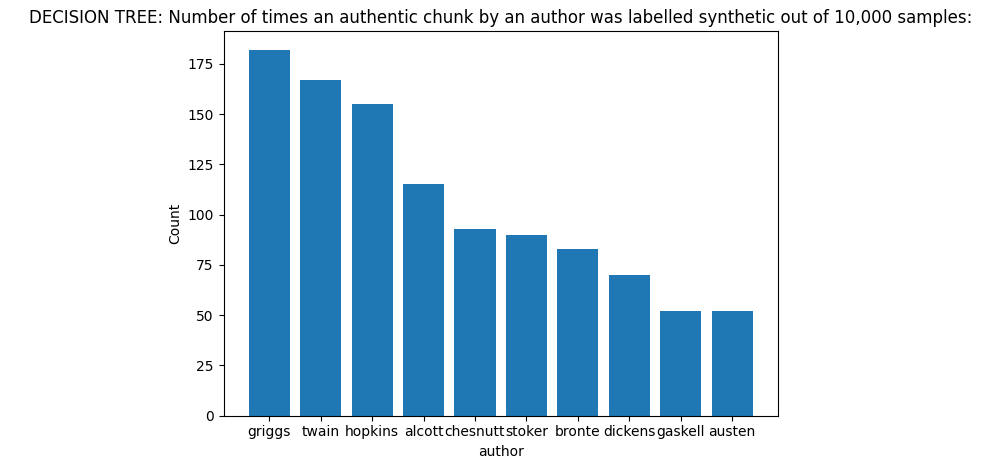


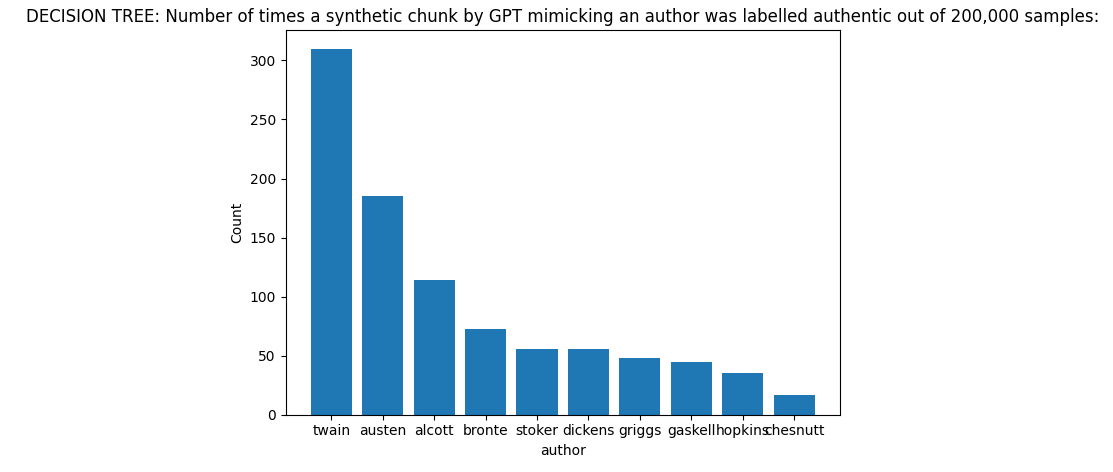


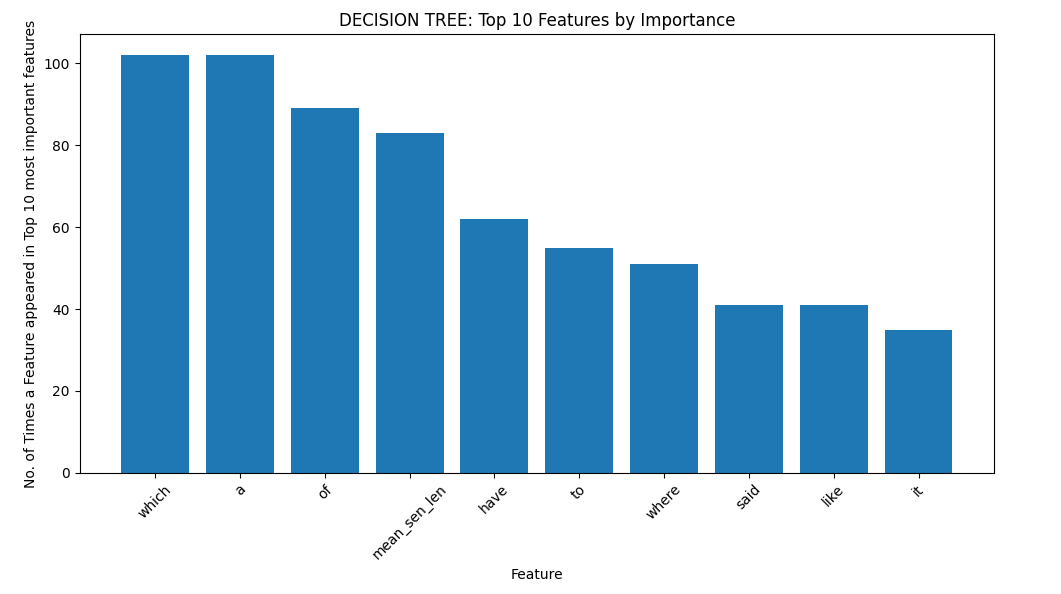
DECISION TREE VISUALIZATIONS

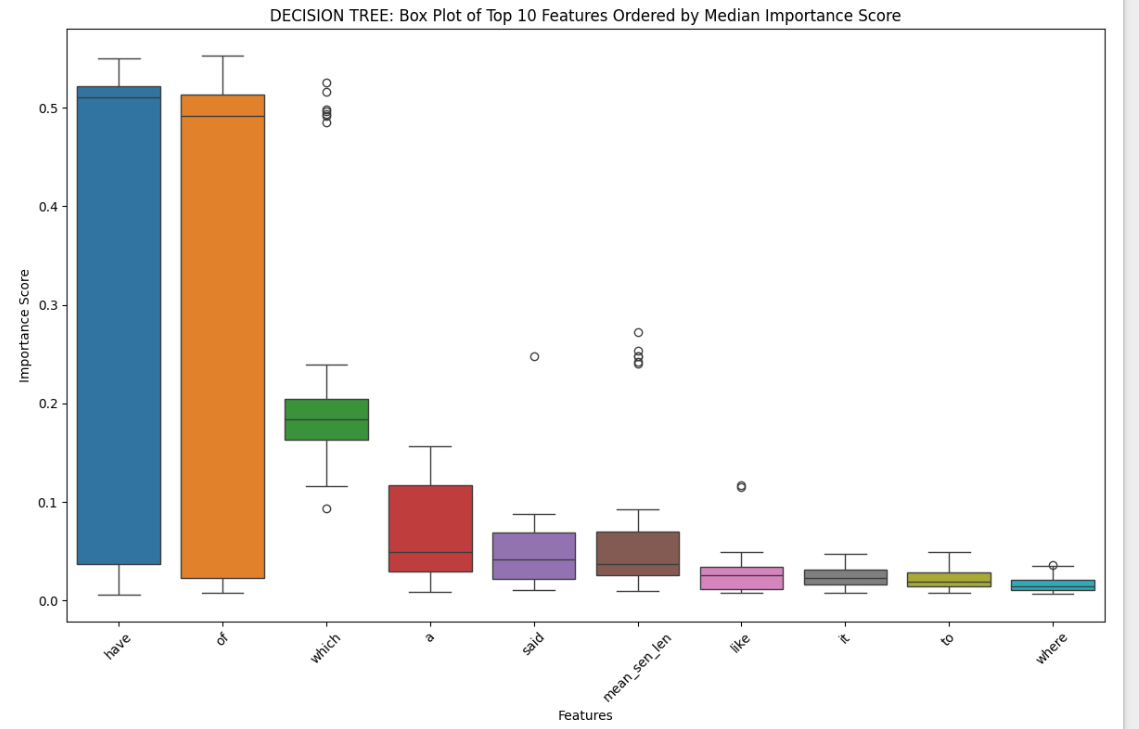




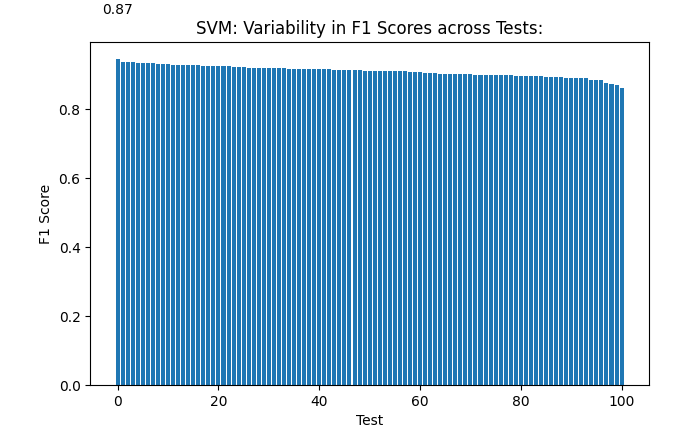


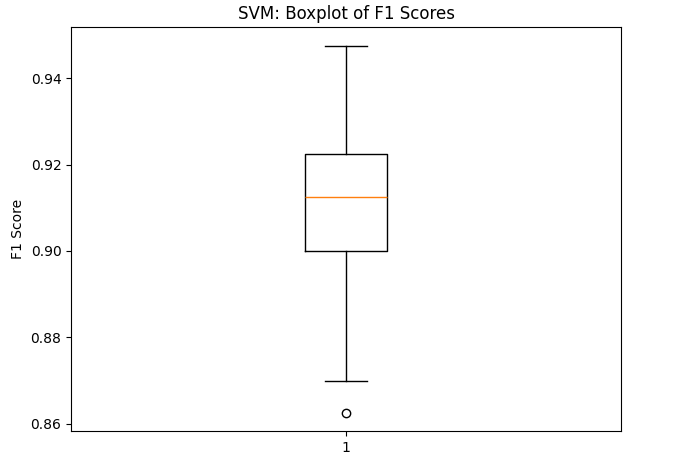


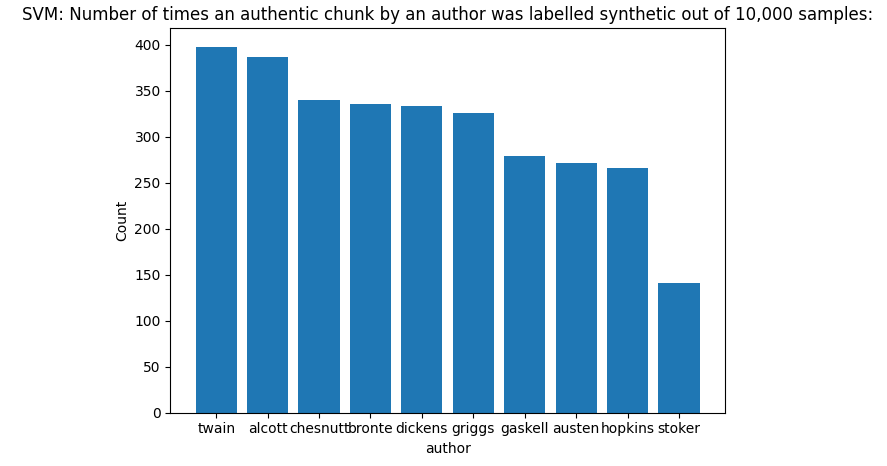


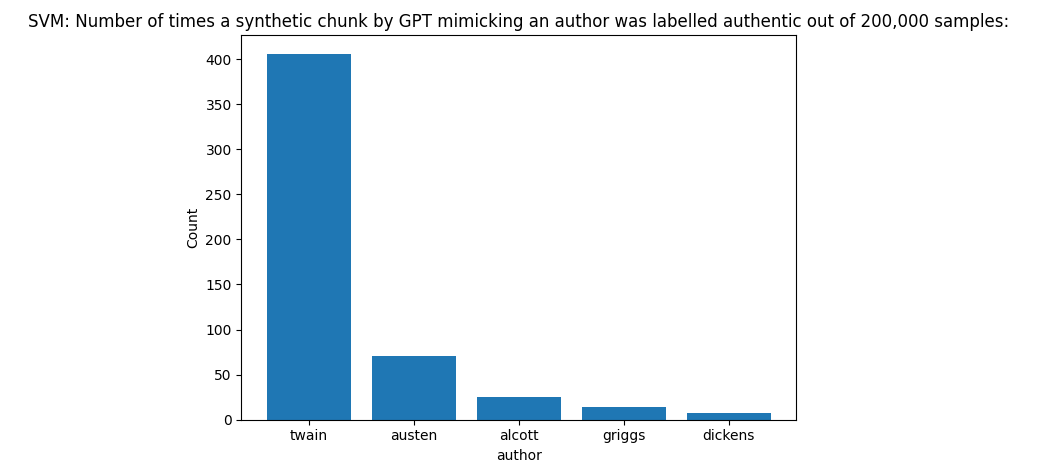


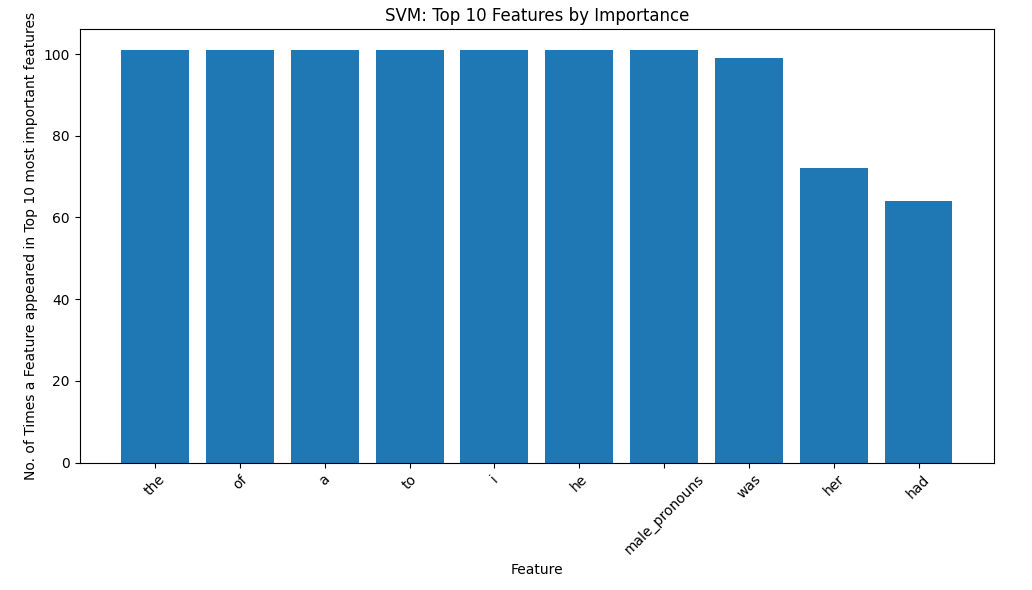
SUPPORT VECTOR MACHINE VISUALIZATIONS

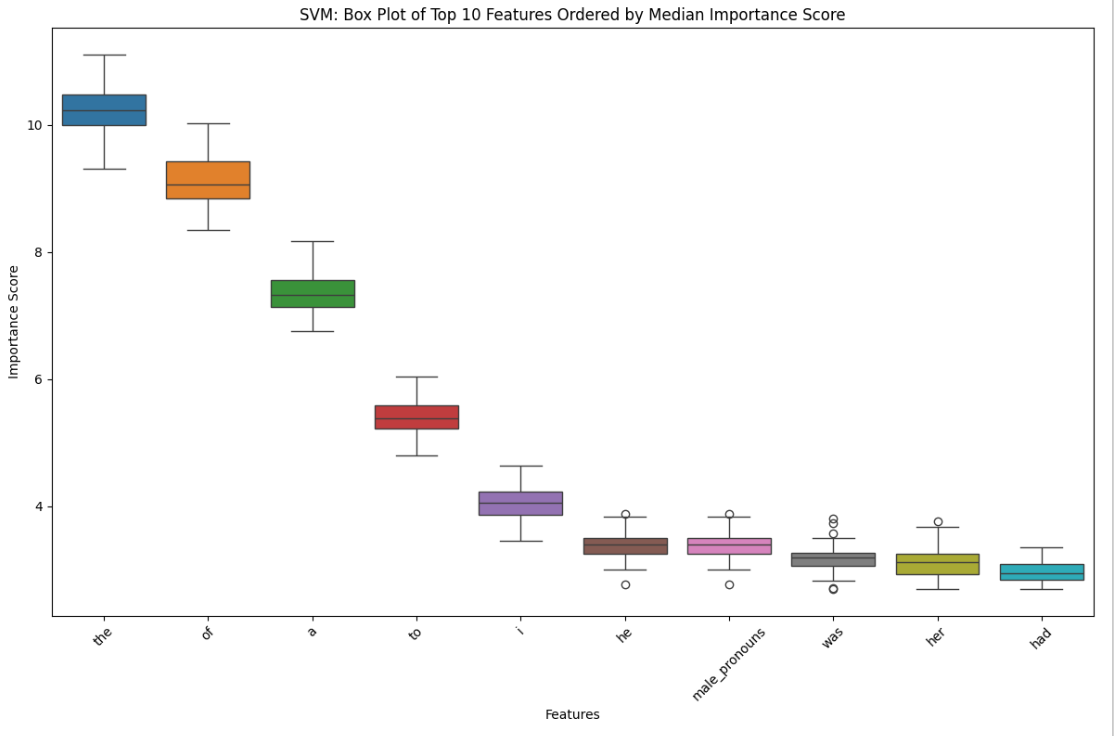




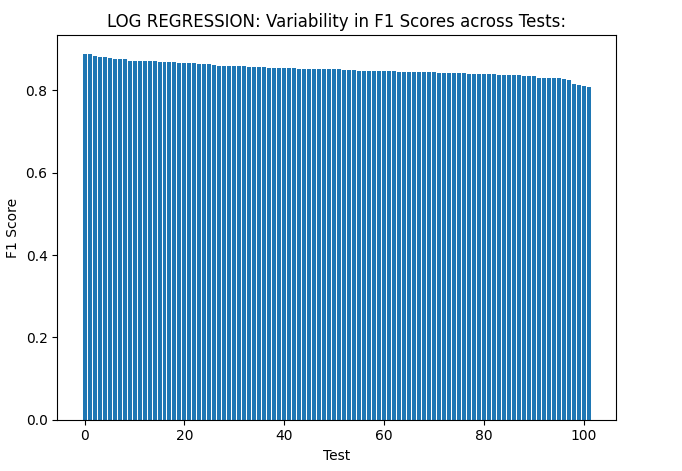


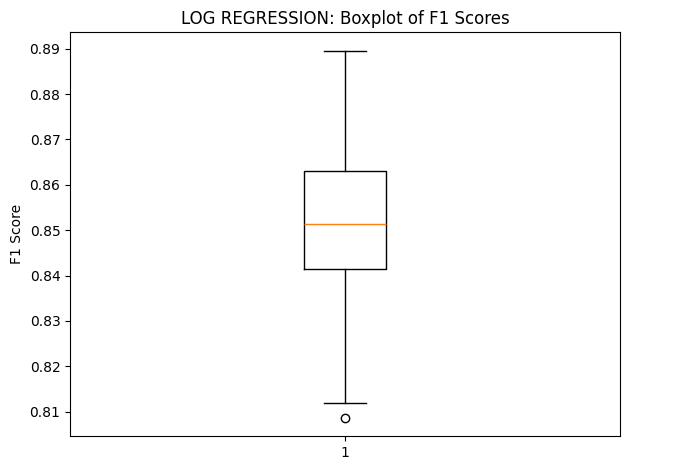


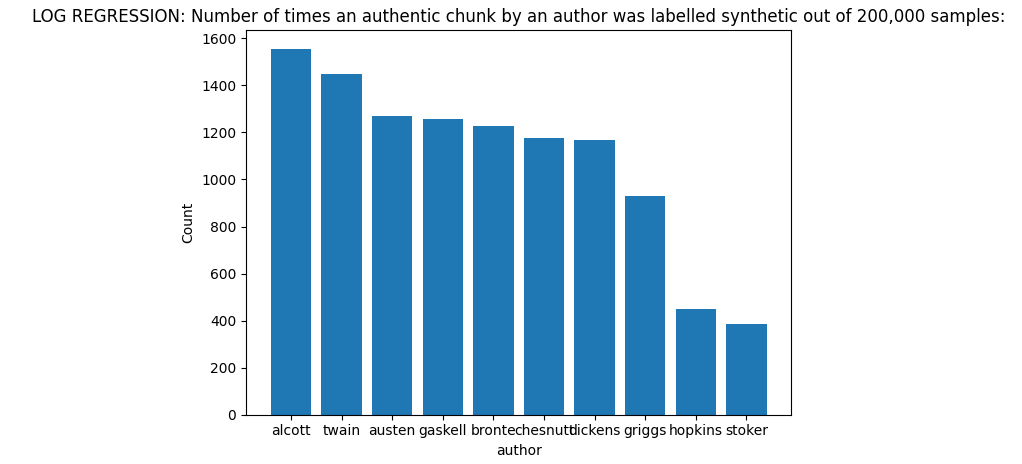


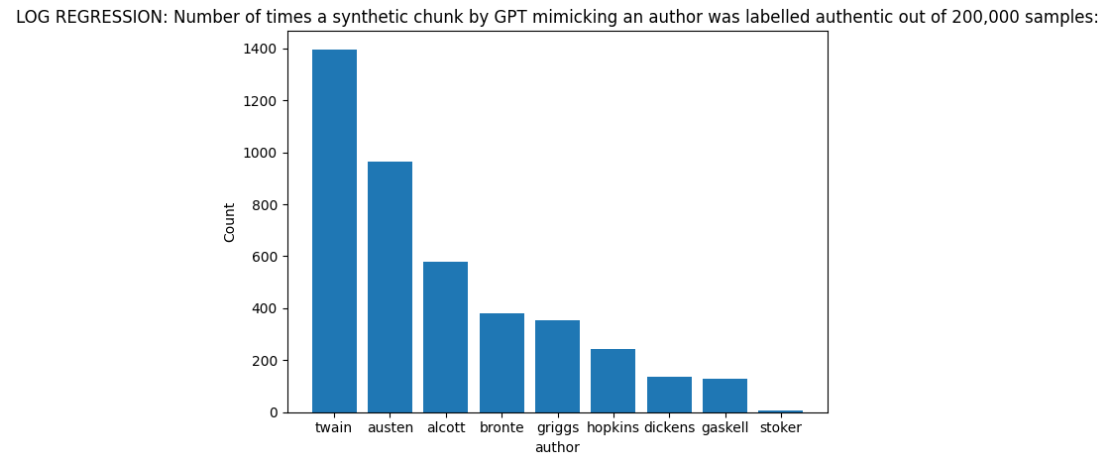


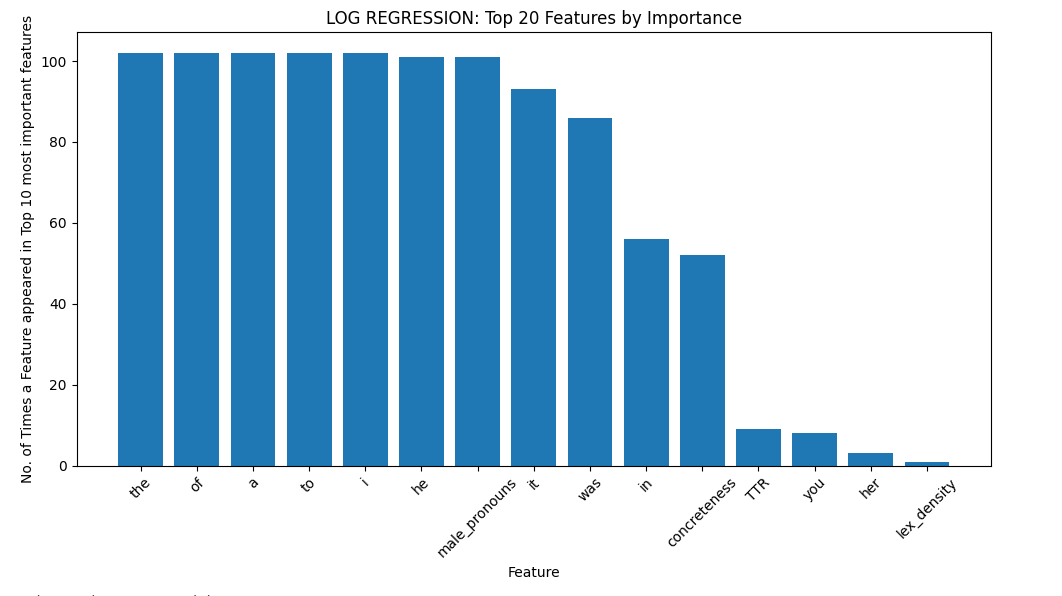
LOGISTIC REGRESSION VISUALIZATIONS

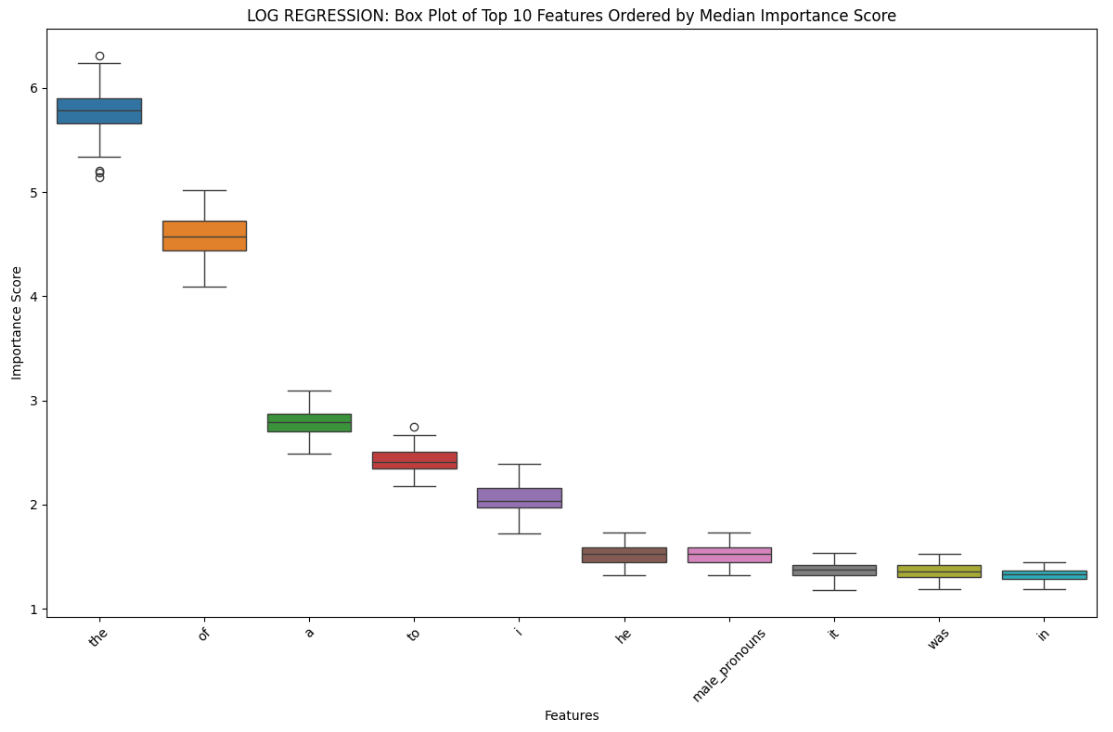


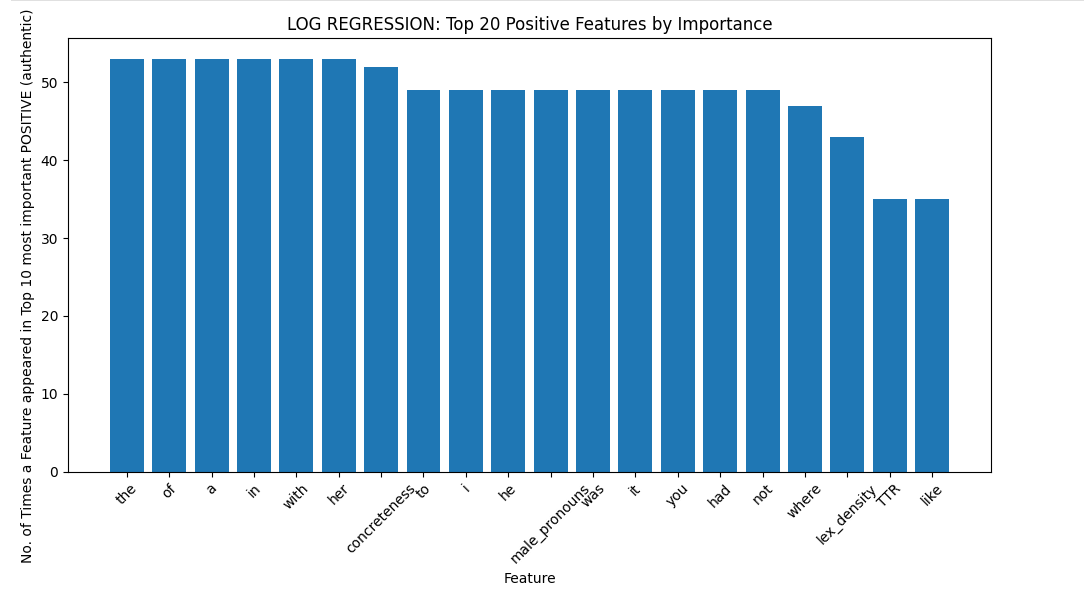


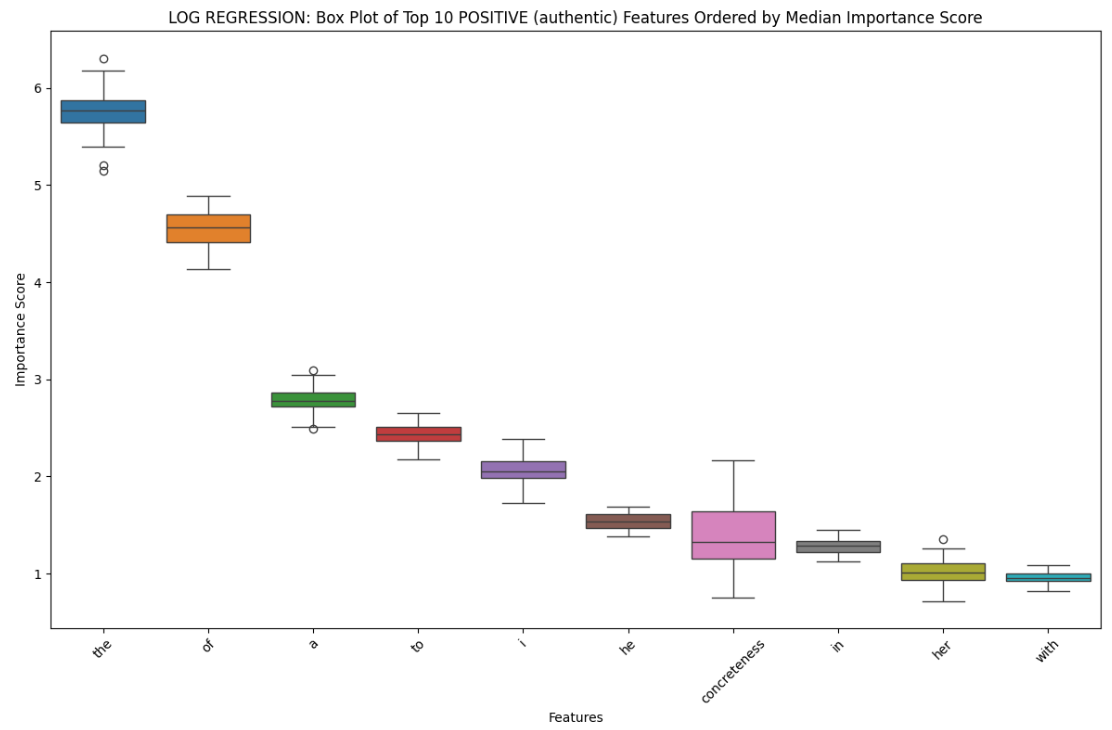


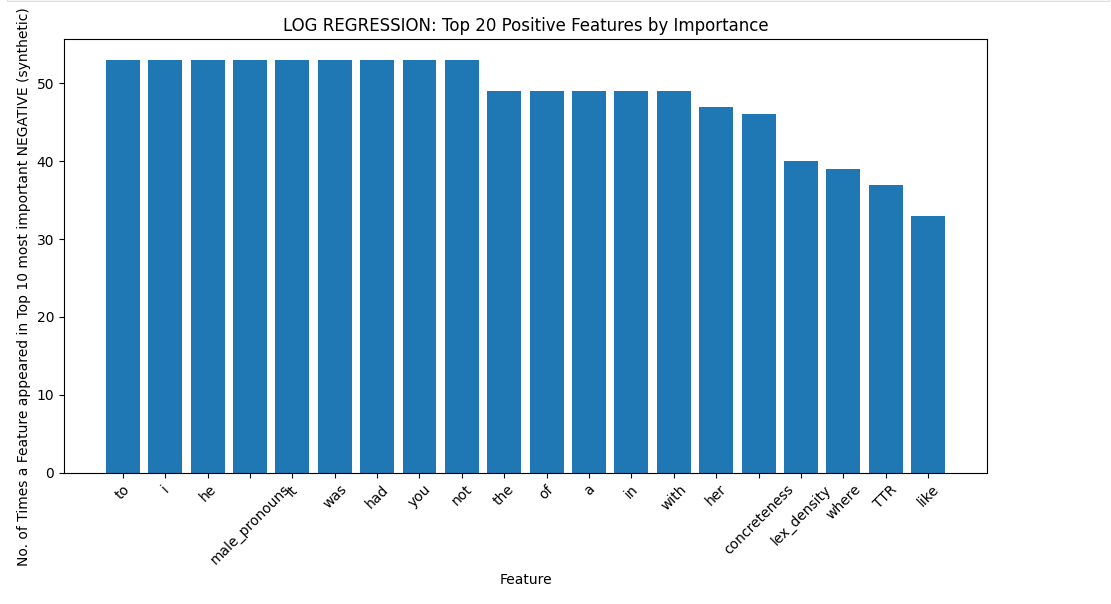


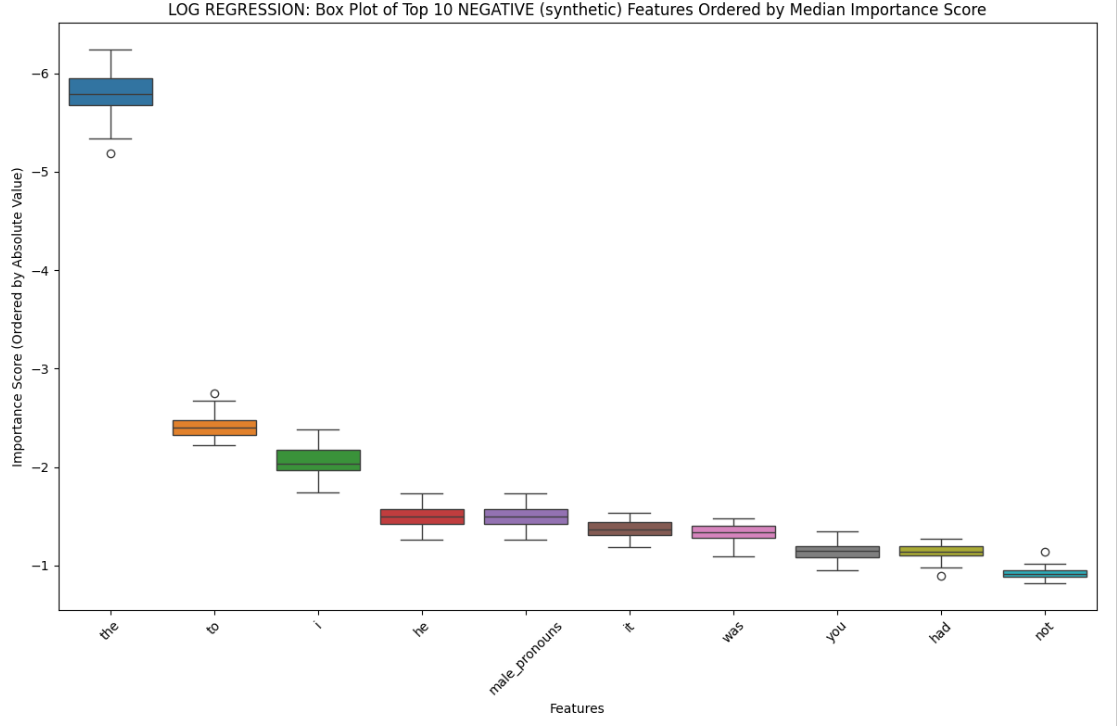












NEAREST SHRUNKEN CENTROID VISUALIZATIONS

