Mercuries and the Machine: London Newsbooks in 1649

Qiyun Chen, Eric Huang, Brandon Fafata, Anupam Pokharel
Advised by Christopher Warren (Department of English) and Peter Freeman (Department of Statistics)

Carnegie Mellon University
Statistics & Data Science

Introduction

Political strife brewing around the time of the English Civil War (circa 1642 - 1651) gave rise to three primary factions:

- the Royalists who supported the Monarch;
- the Parliamentarians, who were opposed to the monarch's ambitions of expansive rulership in Scotland, Ireland, and England; and
- the Levellers (who advocated for populist ideologies).

The unifying objective of the research in the project was to understand how the dynamics of this background influenced the content of newsbooks and vice-versa. A CMU-in-house NLP application, DocuScope, had already been used to generate some features on transcripts of newsbooks (17th-century analogs to the newspapers of today) from the time. Starting with these features and newsbook texts (that have notable spelling and grammatical variations compared to modern English, which can be made more analytically usable with spelling normalization), we formulated three research questions to guide our analyses in fulfilling the objective.

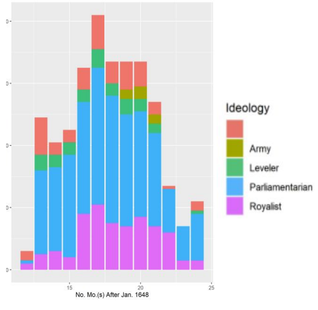
1. ***Can a refined feature selection process that optimizes for ideology prediction help create reduced-dimension plots with more conspicuous boundaries?***
2. ***Are we able to detect texts that are likely to be forgeries?***
3. ***Will spelling normalization significantly alter the feature values generated by DocuScope?***

Data

The dataset contains 607 newsbooks authored by people affiliated with all three factions (a small subset of which could be forgeries) and features (76 of which are numerical) generated by our client and faculty advisor, Professor Christopher Warren of the English Department at CMU, using DocuScope.

Ideology	Parliamentarian	Royalist	Leveller	Army	null
Counts	349	134	40	10	74

Newsbook counts per Ideology group



Distribution of Newsbooks (and the breakdown of ideology) across dataset's timespan

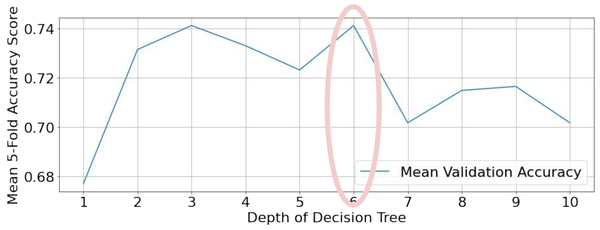
Results and Conclusion

1. Using decision tree models, we are able to perform feature selection. The dimension-reduced t-SNE features constructed from the selected features show better clusterings than the t-SNE features constructed from the original features.
2. We are able to detect potential forgery candidates by noticing outliers in the dimension-reduced plots.
3. We find that normalized and unnormalized English texts cluster similarly (i.e. normalization is not worthwhile for clustering improvement).

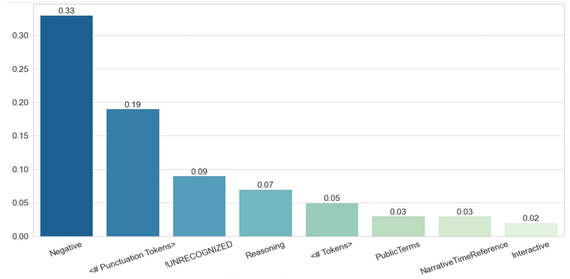
Analysis

Feature Selection

- We select 68 out of original 76 numerical features, and decide not to use the four categorical features since they do not bring predictive power by hypothesis testing. We also perform scaling and fill in missing numerical values (denoted null) with zero.
- We construct decision tree model using these features to predict a newsbook's ideology, performing 5-fold Cross Validation to choose the best tree depth (depth = 6) as our hyperparameter.
- The top most important features are: negative terms (terms conveying a negative sentiment), number of punctuation tokens, unrecognized tokens, reasoning phrases, number of tokens, and public terms.



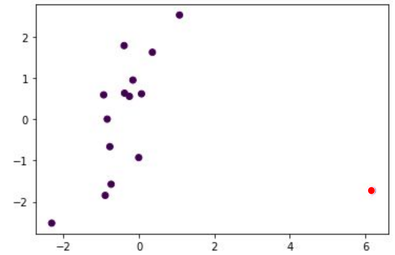
Depth of Decision Tree versus Validation Accuracy



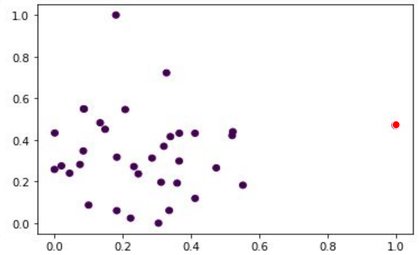
Feature Importances

Feature Outlier Detection

- We assume that newsbooks with similar titles have similar writing styles; if their styles differ, they may be forgeries.
- DocuScope gives us the ability to quantitatively examine features of the text.
- Feature outliers are newsbooks we suspect were forged.
- Graphs show dimensionality reduction of features for newsbooks of a given title.



Newsbooks titled "Severall proceedings in Parliament", Outlier: newsbook 389.

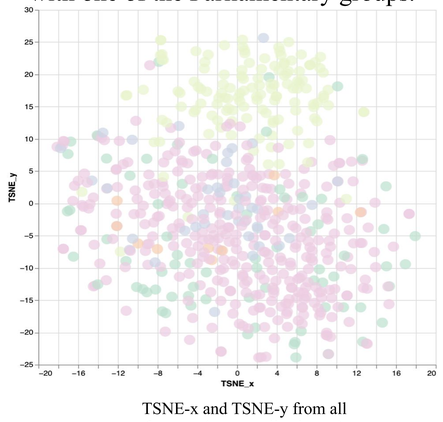


Newsbooks titled "A perfect summary of exact passages" reduced according to feature selection. Outlier: newsbook 396.

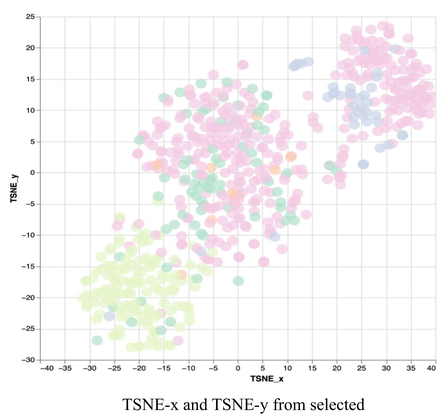
Note: x and y axes are the values in each of the two dimensions after PCA-enabled dimension reduction

Dimension Reduction

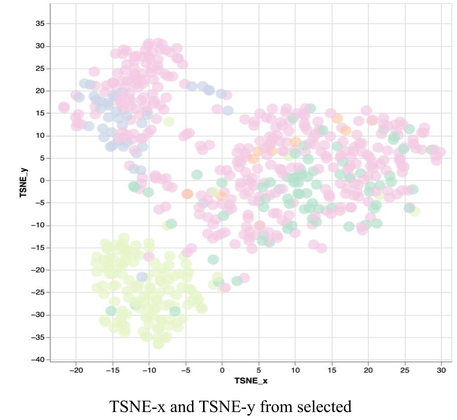
- We use t-distributed Stochastic Neighbor Embedding (t-SNE), a visualization tool that maps high-dimensional data to two dimensions, to see how texts from different factions cluster.
- Comparison among t-SNE plots shows improvement among predicting political groups after dimension reduction.
- Most ideologically unclassified (denoted "null" in legend) data points seem to be Parliamentarians while Levellers seem to group together with one of the Parliamentary groups.



TSNE-x and TSNE-y from all features (unnormalized texts)



TSNE-x and TSNE-y from selected features (unnormalized texts)



TSNE-x and TSNE-y from selected features (normalized texts)