Exploring Theme Recognition in Historical Documents through Machine Learning

Savannah Scott[1], Dr. Randall Cone[2]

[1]Bridgewater College, [2]Salisbury University

ABSTRACT

Text classification, a field in natural language processing, is an automated technique used to assign topics to unstructured text. Themes are a central topic, message, or subject found within a written body of work. While text classification seems to be a useful tool for the humanities, little implementation has been seen in the field of historical analysis. I aim to show the benefits of computational analysis in the field of history by using text classification to identify the themes found in *The Federalist Papers*. By using classification algorithms and neural networks, I aim to use machine learning to identify the themes in the text.

BACKGROUND

The field of the digital humanities contains enormous potential yet is often understudied and undervalued. This project's goal was to embrace the digital humanities by showing how the field of artificial intelligence can be applied to history by using machine learning to identify themes found in *The Federalist Papers*.

The Federalist Papers are a series of 85 essays written anonymously by Alexander Hamilton, James Madison, and John Jay between October 1787 and May 1788 to advocate for the ratification of the U.S. Constitution in New York. They have been cited in Supreme Court cases due to the belief that they communicate the thoughts of the original framers of the Constitution. Due to their historical and modern relevance, a thorough understanding of their themes is essential for discussion surrounding the Constitution.

Machine learning is a sub-field of artificial intelligence that focuses on algorithms that improve with exposure to data, a method like human learning. By training different models on a portion of the essays, the goal was to use the best model to analyze and evaluate the other *Federalist Papers*.

METHOD

The raw corpus of *The Federalist Papers* was obtained through the open-source website Project Gutenberg and was then divided into the individual essays. The training data was constructed by selecting 16 essays, breaking them down into sentences or phrases (documents), and then assigning each document a theme. The list of themes was previously identified through an extensive literature review on *The Federalist Papers*. The final training dataset contained 1519 documents and 15 themes. This dataset was limited and unbalanced. During model implementation, the documents were preprocessed and then vectorized, and the labels were converted to numerical representations.

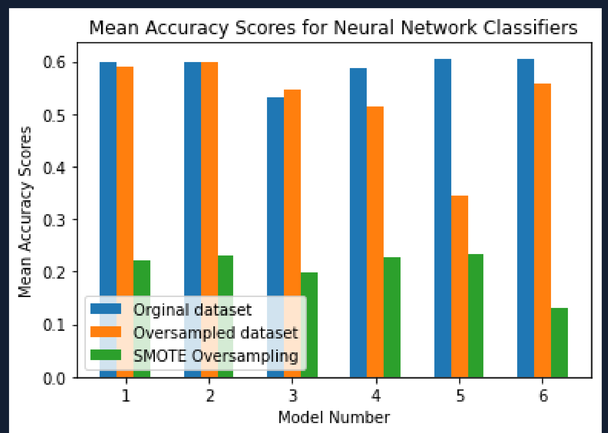
The two main methods of classification employed in these experiments were pre-existing classification algorithms from the Python library sklearn and neural networks build with the Python library keras. In total there were 4 sklearn models and 6 neural network models used throughout this project. Each model was built using a different algorithm or architecture to find the best model for classification. To increase model accuracy, 4 different ensemble classification models from sklearn were also used.

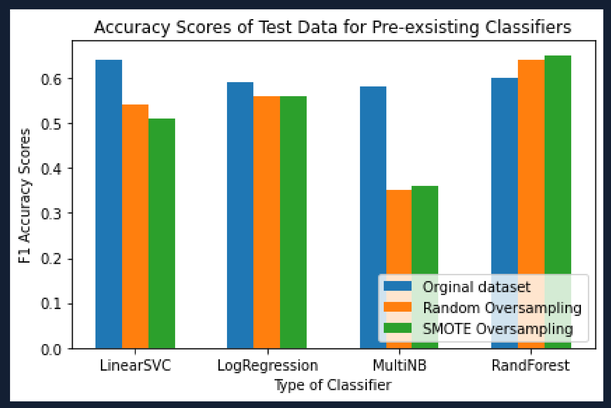
To combat the limited and unbalanced nature of the dataset, different methods of oversampling were explored. Experiments were conducted with the random oversampling and SMOTE oversampling techniques.

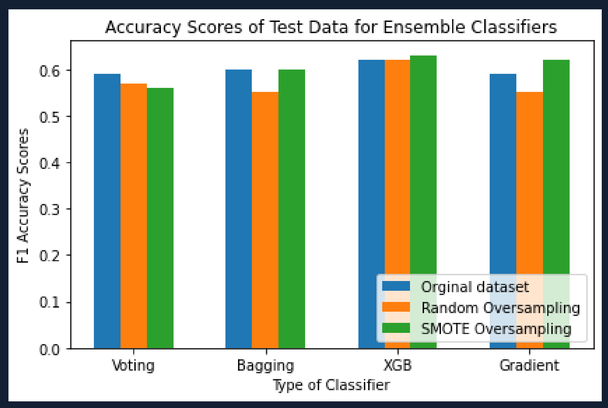
RESULTS

The sklearn models had an average accuracy score of 60.25% on the original dataset with the LinearSVC algorithm performing the best with 64% accuracy. All the classifiers, except the Random Forest algorithm, performed worse with both types of oversampling. The strategies of random oversampling and SMOTE increased the accuracy of Random Forest from 60% to 64% and 65%, respectively. The ensemble classifiers mostly followed the pattern of not being improved by oversampling, and they had an average accuracy of 60%

The neural network models had an average accuracy score of 58.77% on the original dataset, and the model with an architecture of a one-dimensional convolutional layer performed the best with an accuracy of 60.5%. For most of these models, oversampling resulted in worse accuracy scores except for model 3, which increased by about 1.5%.



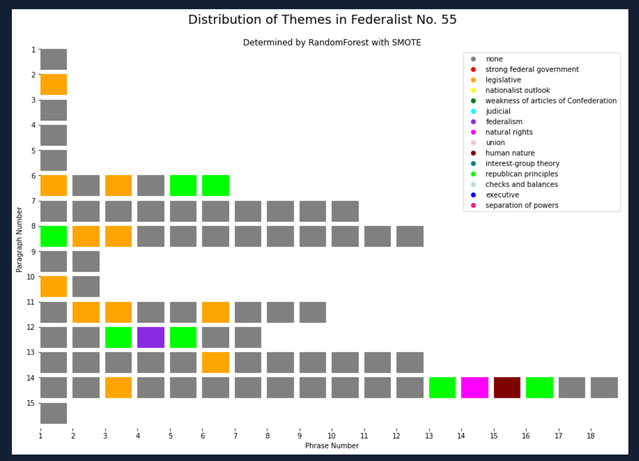




CONCLUSIONS AND FUTURE WORK

While the differences between the types of models are small, the sklearn classifiers consistently performed better than the neural network models throughout the experiment. It can also be concluded that for most of the models, oversampling did not improve the accuracy as expected.

Future work on this project would include exploring other methods to improve model accuracy. Parallelizing this project could also be possible through running each model concurrently or through parallelizing the k-fold validation process. Once the accuracy of the models has reached an adequate level, the models could be deployed to analyze other *Federalist Papers* or works from this time period, such as in the visual below.



ACKNOWLEDGEMENTS

This work was funded by NSF CCF-1757017 for the Explore Emerging Computing in Science and Engineering (EXERCISE) Research Experience for Undergraduates (REU) at Salisbury University.

BRIDGEWATER COLLEGE