

NLP Project Proposal:

Newspapers and online publications are often politically biased and reflect the political leaning of their publisher. Changes to the way the public consumes news media have exasperated these biases and more than ever people are consuming news from non-traditional sources. American politics have become more polarizing as well and many draw on sources which lean to their preferred political ideology.

Given the election year and the ramping up of politically biased news articles, our group will address the question: can we develop a natural language processing pipeline to detect political bias in news articles and summarize the articles using unbiased language to mitigate biased media consumption?

To address this question, we will use the Newspaper Bias Dataset (NewB) created by Jerry Wei and an Article-Bias-Prediction dataset created from allsides.com by Ramy Baly. The NewB dataset contains 200,000+ Sentences about Donald Trump labeled by news source and political bias. The Article-Bias-Prediction dataset contains a total of 37,554 articles labelled by political bias.

Our group will employ both a classical model and an advanced neural network model. The classical model will be either a Naïve Bayes or Logistic Regression model to classify articles based on their bias. This will be evaluated using F1 score to determine the performance of the model.

After an article is classified as biased, a neural network-based model will be used to generate an unbiased summary of the article. The metric to evaluate this model will be determined when the exact neural network model is determined.

A rough schedule to complete this project is as follows:

Week 1: Data ingestion, cleaning, and classification model development.

Week 2: Classification model optimizing and begin generative model development.

Week 3: Develop generative model and optimize.

Week: Finalize project write presentation.

Datasets:

<https://github.com/JerryWeiAI/NewB>

<https://github.com/ramybaly/Article-Bias-Prediction/tree/main>

<https://arxiv.org/pdf/2006.03051v2.pdf>

<https://arxiv.org/pdf/2010.05338.pdf>