Chris Washer

NLP Individual Project Report

**Introduction:**

News reporting often is influenced by the viewpoint of the author or media organization which affects the language in the story. Changes in the way the public consume news media and the accessibility of news from non-traditional sources have increased the prevalence of political bias in media. While all media should be expected to have bias, hidden bias in news media, and reporting opinions like facts, is an increasing challenge to the public good of open information.

These changes require individuals to become equipped with ways to understand and identify political bias in news media. Awareness and education are vital to spotting media bias in news reporting. But in this project, our team set out to develop a pipeline to help media consumers detect the bias in news they read.

The goal of our project was to compare several text classification models, from classical methods to state-of-the-art methods, to classify news articles as left, center, or right political bias. After identifying the best model to perform the classification, our team created a pipeline to ingest a URL of a news article and determine the article’s media bias. After determining the bias, we leveraged a transformer model to create center-bias abstractive summaries of the article to help resolve the media bias.

Our team leveraged prior work done by Baly et al. (2020), by using the left, center, right labeled dataset of news articles. Using the dataset they generated, we built the following models to detect political bias: Logistic Regression, Naïve Bayes, Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) Recurrent Neural Network, and a fine-tuned RoBERTa Transformer. In addition, we deployed a pretrained Pegasus model to perform abstractive text summarization on news articles.

In this report, I will talk about my contributions to the effort. Namely, I will discuss the MLP, CNN, the RoBerta model, and the Pegasus Summarization (and failed attempt to pretrain).

**Description of Your Individual Work**

Initially, I created the MLP model and trained on our news bias dataset. I created a custom data loader, word2vec embedding, and a simple 3-layer architecture.

After, I created the CNN model using a similar code framework as the MLP model. I used the same dataset class and word2vec embedding but added convolutions into the neural network module.

The third model created was the fine-tuned RoBERTA transformer model. I again used the same dataset class to load the data. However, the preprocessing was done using a RoBERTa Tokenizer with a max length of 512.

The last model I deployed was the Pegasus abstractive summarization model. I initially tried to fine tune the generative model but our data set did not have unbiased summaries that could be used for training. I tried to use an unsupervised training method but after the training the model failed to abstractively summarize the text.

Luckily, I found a Pegasus model that was pretrained on a large, multi-news corpus.