Political stance classification of news-articles

Political bias is increasingly prevalent in common discourse, among elites, and the media. Most attempts at classifying ideology or bias from text have used Recurrent Neural Networks (RNN), with some recent attempts using Multi-Layer Perceptrons (MLP). Here, we explore the extent to which models of different complexity can identify the political leanings from the sources of news article documents based on the articles' lexical semantics. To this end, we use the BIGNEWS dataset, which consists of 3,689,229 English news articles published on politics between January 2000 and June 2021, from 11 US media outlets covering the left, right, and central political spectrum. As a first step, we summarized documents via one-hot-encoded word-vectors and applied multinomial logistic regression to predict the news leaning of a subset of 526,500 articles which succeeded chance-performance significantly (accuracy > 70%). A simple context-insensitive feed-forward NN using pre-trained FastText word embeddings did not outperform logistic regression. We compare these results to the performance of two ANNs. Both networks are based on pre-trained context-sensitive text embeddings that differ in input-token-length: (1) BERT (max. length 512) and (2) Longformer (max. length 4096). We found that Longformer was the most effective model for classifying political leaning. Based on recall precision and accuracy, and model confusion-matrices, we will determine which model is most effective at predicting news outlets' political leaning. These results can help quantify the characteristics of political biases in news-articles, and potentially uncover unwanted or hidden biases in ostensibly neutral news reports.