Bias in News: A Data Driven Approach

Devanshi Deswal, Samar Dikshit, Connor Higgins, Kartheek Karnati, Oliver Spohngellert

Introduction

Different news organizations of different political leans not only have biases, but actually report on completely different topics. To prove this, we collected and analyzed data on what is appearing on news sites of left, right, and centre leans. We hypothesized that there would be clear differences in what is appearing in both the headlines, and the text of news articles from different news organizations. We also hypothesized that a classifier could be trained to recognize the political lean of an article given only its text.

Data Collection

Data was collected through custom scrapers of eight online news websites. For left leaning sources, we scraped CNN, Vox, Mother Jones, and FiveThirtyEight. For right leaning sources, we scraped Fox News and Breitbart. For the centre, we scraped BBC and Reuters. All these labels were determined through All Sides' media bias ratings. In total, there were 8,992 articles from August 1 to November 20th, after filtering out non-political articles.

Exploratory Analysis

We focused first on headlines because many people do not read the content of articles and just read the headline. As can be seen in Figure 1, the words used in headlines differ greatly depending on the political lean of the reporting organization. It is clear that the left is reporting on impeachment much more, though "trump" appears in similar amounts on both sides. Further, the word "biden" appears much more in right leaning headlines, likely due to reporting on both Joe Biden and his son Hunter Biden. It is also of note that as impeachment has ramped up, it has become much more common to see in headlines, almost rivalling even "trump" during November.

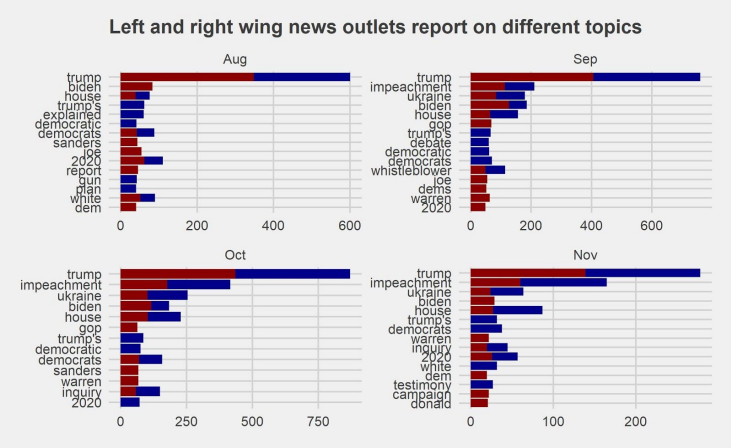

Figure 1: Top words used in headlines on the right and left

We also look at which 2020 presidential candidates are being mentioned the most in each news outlet. As can be seen in Figure 2, almost every organization had the top 5 polling candidates (including trump) as the top 5 most mentioned. The most notable outlier here is FiveThirtyEight, which has a clearly much more balanced amount of reporting on the top candidates, with Trump not even being the most mentioned candidate.

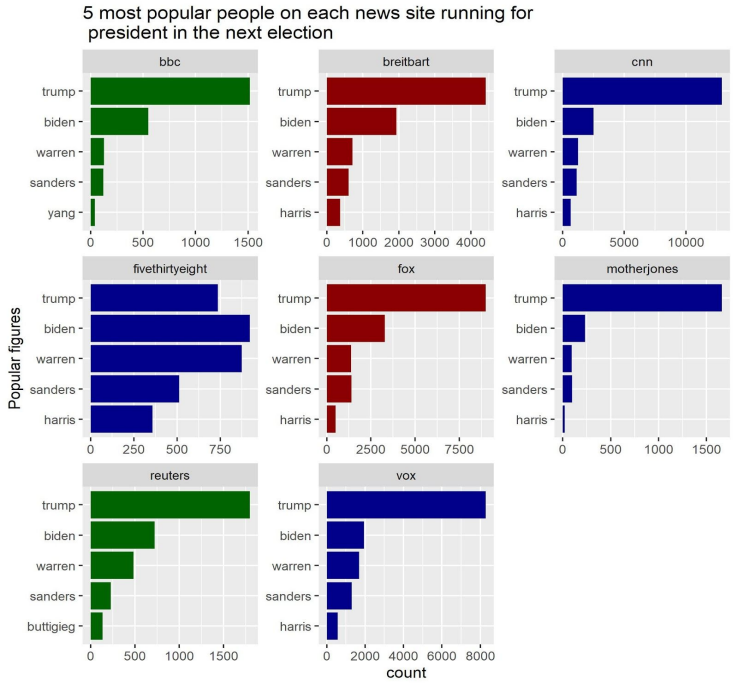

Figure 2: Number of mentions of top candidates per news site

There is deeper analysis beyond this, including readability scores, and analysis of top bigrams, but we decided not to include these for brevity.

Modeling

Another area of interest was the extent to which a machine learning model could predict an article's political lean given its text. We used fasttext's pretrained word embeddings to transform the data to a numerical format, using only the first 200 words of each article. We then trained a LSTM model on the data. This produced promising results, with a top line accuracy of 88%, however this is not the best metric as we do not have many articles labeled centre. Below is the ROC curve for this model. Based on this, all models are performing fairly well on the test data.

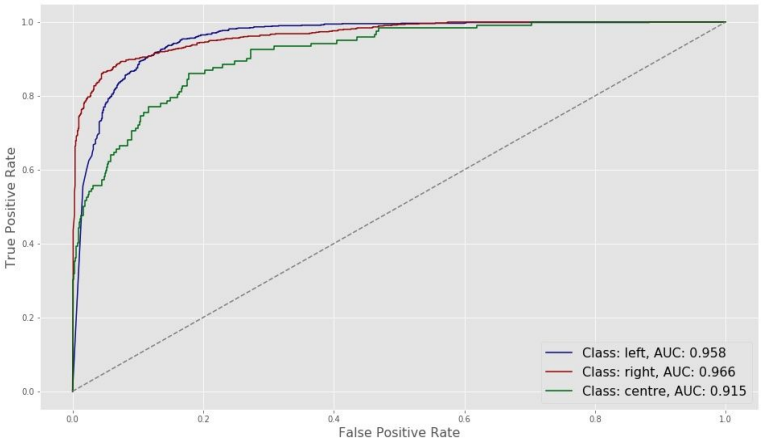

Figure 3: ROC curve for LSTM model

This shows the text of the article contains clear indicators as to what side of the political spectrum the article comes from.

Extensions

There is insight to be gained from analyzing this type of data. For example, we would like to analyze differences between this data and data during Obama's presidency to see if organizations have changed the way they report since then. Further, we would like to gather from more sources to create a more robust classification system. Finally, we would like to collect from extremist sources, to build a classification system for such content, though this is a long term goal.

Conclusions

There is strong evidence in the data that different news organizations have distinct ways of reporting the news. Organizations on every side of the political spectrum have different focuses and writing styles, and that shows in this data. This type of analysis is important, as it informs people on how to view the news they are reading.