Media Bias Analysis Through Text Mining

Members - Rakesh Ravi K U, Boda Ye, Sakshi Jawarani

The goal of the project was to uncover media bias through textual analysis. For the corpus, we used news headlines and google news summaries captured through RSS feeds on ontolligent.com/newzy. We then used information from media bias check to confer labels (left wing, center and right wing) to each news article based on the media house that published them. For pre processing our data we used text processing techniques using the NLTK libraries to tokenize, lemmatize and remove stop words from the corpus. As the majority of news were from left wing news houses, we balanced classes through random but representative downsampling to ensure no bias seeped through our analysis.

We performed topic modelling, sentiment analysis, word frequency analysis and data visualization to uncover hidden patterns in our data. To identify the right number of topics, we plotted perplexity and coherence score curves and this analysis yielded four as the appropriate number of topics. Based on the most frequent keywords occuring in each topic, we were able to identify the topics as Politics, Crime & Accidents, Business & Technology and others. We used TextBlob to annotate the data set with polarity and subjectivity scores for each news summary in the corpus.

We used the news from the center as a reference and compared left wing and right wing news. Comparing the yearly trend of sentiment polarity of news from different media houses showed us that peaks and troughs alternate for left wing and right wing news. Neither of left or right wing news showed any similarity in pattern with the news from the center, which we consider the least biased in our analysis. This provided some basis for the bias that we were trying to uncover.

Delving deeper into the topic modeling results we found that news from unbiased sources (center) revealed that there was substantial material in both business & technology and politics. It's interesting to note that most of the left wing news revolved mostly around business & technology with little emphasis on politics and crime. In contrast, right wing news mostly revolved around politics with the least focus laid on business & Technology.

Now that it is clear that different news tend to focus on different news material to cater to their audience, we dug deeper to identify word frequency patterns over time for two of the most frequently occuring keywords, “Trump” and “ISIS”. On a year level, the patterns were not discernible between the two keywords. It was interesting to note that there were more left wing news houses talking about President Trump than that of the right wing news houses. On a more granular month level, nuances in the patterns began to emerge. The peaks and troughs in the chart come around the same time with very little lag. But when we looked at the year 2017 in specific where President Trump was embroiled in controversy, we observed that the peaks/troughs for the left and right were out of phase. For ISIS, there was more news coverage from the right wing than the left. On a month level, it was clear that there was similarity in the pattern with a little bit of lag. There are instances where the left wing news houses suddenly publish more articles on ISIS and there will be quick but elevated response from the right wing news and vice versa. This could mean that once the left covers the news in a certain way, there is immense pressure for the right to respond to the coverage to propagate its views or vice versa.

Although we were able to extract patterns that provided an insight into the biases propagated by news houses, it is still not conclusive. More analysis can be performed by capturing more news from the right wing media outlets as there was a shortage of them in the data set. In future work, we will look to incorporate more news articles from media outlets that were not already captured in this corpus. In addition, we will attempt to build an application using flask that can work on data from newzy on a real time basis.

References

1. Media Bias Fact Check “https://mediabiasfactcheck.com”
2. Newzy “http://ontoligent.com/newzy/sources-and-items”