A NLP Approach to Understanding Variation in Political News

Stephen M. Lee

July 24, 2019

Contents

1 Introduction

The 2016 United States Presidential Election, to many, raised questions as to the reliability of their news sources [SOURCE NEEDED]. Since then, many social media companies and other institutions have begun public campaigns to combat the perceived threat from fake news, with arguably limited results [SOURCE NEEDED]. With a goal of news source classification, I scraped several thousand news articles from Fox, Vox, and PBS News. By some estimations [SOURCE NEEDED], these three represent distinct categories of news: Fox is often considered extreme right opinion (i.e. conservative); Vox is considered extreme left opinion (i.e. liberal); and PBS is considered center primary source news. I first calculate the top word, 2-gram, and 3-gram frequencies to better understand the dataset, and then train a Bidirectional LSTM neural network with pretrained embeddings to classify the news source.

1.1 Literature Review

2 Data

I mined political news articles from the websites of Fox News, Vox News, and PBS News. Importantly, I restricted focus to only URLs that contained an explicit reference to politics i.e. from their respective political sections. This allowed me to collect articles that were as similar to each other as possible to try and limit the chances of spurious predictive results. Intuitively, the motivation behind this is to facilitate classifiacation based only on sentiment or semantics, rather than subject matter differences. A summary table of my data is shown in Table 1.

Source	Documents	Avg word count	Pct Adjectives	Pct Adverbs	Avg words per sentence
Fox	661	686.2	0.066	0.034	20.1
PBS	1739	654.3	0.066	0.032	18.0
Vox	1027	1332.8	0.073	0.046	21.3

Table 1: Summary statistics by data by source.

To better understand how similar the content of these articles are, we can construct n-gram tokens and count the frequency of their occurrences. In Table 2, we see the most frequent one, two, and three-gram phrases.

Word Frequencies			
	Vox	PBS	Fox
1	trump	trump	trump
2	tax	said	said
3	will	president	president
4	people	house	house
5	health	will	new
6	bill	new	will
7	republicans	white	democratic
8	one	senate	democrats
9	new	democrats	told
10	care	campaign	border
2-gram Frequencies			
	Vox	PBS	Fox
1	health care	white house	white house
2	white house	president donald	new york
3	trump administration	donald trump	president trump
4	donald trump	special counsel	green new
5	tax cuts	supreme court	health care
6	health insurance	attorney general	new deal
7	new york	new york	united states
8	affordable care	justice department	border security
9	tax bill	counsel robert	donald trump
10	federal government	trump said	state union
3-gram Frequencies			
	Vox	PBS	Fox
1	affordable care act	president donald trump	green new deal
2	president donald trump	special counsel robert	house speaker nancy
3	congressional budget office	majority leader mitch	special counsel robert
4	health care bill	attorney general jeff	partial government shutdown
5	new york times	senate judiciary committee	speaker nancy pelosi
6	majority leader mitch	sarah huckabee sanders	state union address
7	american health care	senate majority leader	new york times
8	leader mitch mcconnell	counsel robert mueller	majority leader mitch
9	corporate tax rate	leader mitch mcconnell	president donald trump
10	senate majority leader	secretary sarah huckabee	senate majority leader

Table 2: Word Frequencies

2.1 Challenges

There are several limitations to discuss. The most obvious is the difference in corpus size from each source. In particular, Fox News has fewer documents than either PBS or Vox by quite a large number. Fortunately, however, there are many well established best practices for dealing with imbalanced data [SOURCES]. In this work, I bootstrap the data for balance. Second, due to the variability of online formatting, it's worth noting the possibility that, even after cleaning, each source exhibits some subtle idiosyncratic standards that could allow a neural network to detect those instead of pure sentiment and semantic differences. To mitigate this, I removed any mention of their own organization, or any other unique affiliations or locations.[1] Finally, each news source shows a significant difference in the average length of each article. To overcome this, I limited the article input size to the first 500 words to ensure that no single source was consistently shorter.

3 Models

4 Results

5 Implementation

6 Conclusion and Discussion

[1] For example, for every Fox News article, I removed any mention of 'Fox'. Additionally, Fox News cited the 'Associated Press' disproportionately often, so I also removed that string.