A NLP Approach to Understanding Variation in Political News

Stephen M. Lee

July 25, 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. articles from their respective political sections. This allowed me to collect articles that were as similar as possible to each other in order to try and limit the chances of spurious predictive results. Intuitively, the motivation behind this is to facilitate classification based only on sentiment or semantics, rather than subject matter differences. 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 1, we see the most frequent one, two, and three-gram phrases.[1]

[1]The associated counts of each n-gram phrase are shown in Table [NEED REF] in the Appendix.

Table 1: Most frequent words and phrases, by news source.

Most Common Words

	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

Most Common 2-gram Phrases

	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

Most Common 3-gram Phrases

	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

Looking carefully at the most common words and phrases we see substantial similarity in terms of topic. Regardless of the source, "Trump" is the most used word. Perhaps surprisingly, the top four most used words for PBS and Fox news are exactly the same, and in the same order. Beyond that, we see phrases "white house", "president Donald Trump", and "senate majority leader" appear in the top ten most frequent phrases for each news source. Together, this suggests that, topic wise, the corpus for each source are comparable.

In addition to subject, wording, and phrasing, I also check for grammatical and structural differences with a simple lexical analysis. Using the University of Pennsylvania tagset, I count the percent of adjectives and adverbs in each article and calculate the average for each news source. I find very similar use of adjectives and adverbs for Fox and PBS, and a slightly more frequent use with Vox. Intuitively, the goal is to understand, on average, how descriptive the language is for each source. Toward this end, I included comparatives (e.g. better, worse, greater) and superlatives (e.g. best, worst, greatest) in the count. One may be interested to separate them out further to see if any news source provides more "dramatic" descriptions i.e. has a higher relative percent of superlatives. Unfortunately, given the size of my dataset, I was unable to find anything statistically meaningful.

Finally, I calculate the average number of words per article by source, and the average number of words per sentence, again grouped by news source. Here I find that PBS writes the shortest sentences, while Vox writes the longest. This measure is relevant when considering the average number of words per sentence as a proxy for complexity, as shown in [NEED SOURCE]. Table 2 summarizes these descriptive statistics.

Table 2: Summary statistics by news source.

Source	*Documents*	*Average word count*	*Percent adjectives*	*Percent adverbs*	*Average words / 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

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, any other common and unique affiliations, and other identifying characters.[2] Finally, each news source shows a significant difference in the average article length. To overcome this, I limited the article length to a maximum of the first 500 words to ensure that no single source was consistently shorter when fed into the neural network.

3 Models

I train and compare a recurrent, bidirectional, long-term short-term memory (LSTM) neural network against a baseline unidirectional LSTM model. The key advantage of the recurrent LSTM architecture is the ability for neural cell to "remember" relevant lagged values. Mathematically, each unit is described in Figure [NEED REFERENCE] below.

I train unidirectionally (i.e. forward-only), where a given word only "knows" of words that precede it through the LSTM architecture, and bidirectionally (i.e. forwards and backwards), where a given word can "know" about what came before it *and* what follows. Intuitively, we can think about this through the following example. Consider the sentence, "The man sat to eat an orange, which, oddly, matched the color of his beard with surprising accuracy." When we as humans read that sentence, we can retroactively modify our understanding: this is to say that we can update our image of the man even after reading about him. In this example, it's possible to first imagine a cleanly shaved man with short brown hair, and *later* update your mental image to a man with long orange hair and a shaggy beard. Similarly, training the neural network both forward and backward allows for additional context.

3.1 Embeddings

To encode the

4 Results

5 Implementation

6 Conclusion and Discussion

[2]For example, I removed any mention of 'Fox' from every Fox News article. Similarly, Fox News cited the "Associated Press" disproportionately often, so I also removed that string. Additionally, PBS News begins each article with location information in the following format: "LOCATION — Start of article...". In this case, I removed the names of the most frequently referenced cities and the following "—" character.