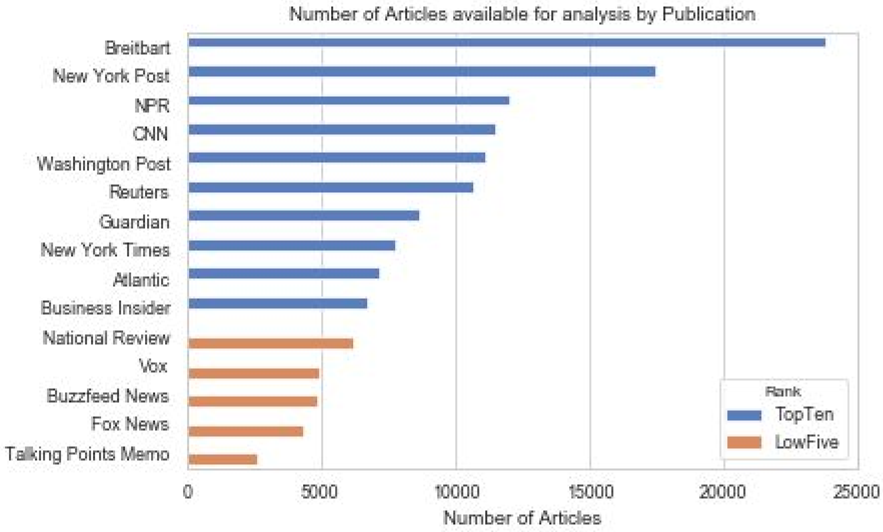
Identifying Bias in News Stories from Parts-of-Speech Frequency

Overview

With our current disparate political environment and rampant allegations of fake news and media bias, I thought it would be interesting to try to apply NLP techniques to a set of news articles to test whether one could identify bias based on writing styles. That is, can characterizations that insert opinion into a story's facts be measured as an indicator of bias?

About the data

The data source was obtained from Kaggle. It is titled "All the News" and was compiled by Andrew Thompson. The data includes articles from 15 news sources and I further filtered the data to the top 10 sources. That is, the 10 sources with the highest number of articles.



There are nearly 117 thousand news stories in the filtered data source retrieved predominantly from the years 2015 to 2017. In addition to the article itself, the data includes the publication, title of article, author, url (when available), date of article with year and month as additional separate data items.

Data cleansing

There were a series of rows with nulls in the date fields (2641 rows - less than 2%) that I determined were errors in the csv conversion. They appeared to be continuations of the prior row’s article filling the text fields. Since they represent such a small amount of the data and it is likely that the portion of the article correctly captured with the initial row of data is sufficient for purposes of this particular project, the rogue rows were dropped from the data.

There were 2 rows with missing titles and over 15,000 missing the author. The missing data was filled with “Not Available” to avoid issues with nulls and to allow these two data items to be included in the analysis. I dropped the url column.

As part of the data cleansing process, I modified and applied the text cleaner function to remove double dashes and apostrophes that were impeding the tokenizer. In addition, I broke the data into articles by each of the top ten publications to avoid memory errors during tokenization.

Initial word frequency counts that excluded punctuation and stop words had some unexpected results. The single letter “s” and the combination “nt” came up in the top 20 of several publications. Investigating what I suspected was some kind of error related to apostrophe use, disclosed that “S” was of such high frequency due to the frequency of U. S. and U. S. A. The frequency of “nt” was due to the use of a wildcard to mask inappropriate words as in “c*nt”. Therefore, I concluded that the functions were performing correctly. The data is good for analysis.

Analysis Process

I used spaCy to tokenize the article content. It was applied to all articles by publication. The basic content breakdown follows:

Brietbart has 25490 sentences with 219071 words.
New York Post has 20634 sentences with 169333 words.
NPR has 13141 sentences with 119393 words.
Washington Post has 12519 sentences with 97561 words.
Reuters has 10841 sentences with 97278 words.
Guardian has 9435 sentences with 84338 words.
New York Times has 9727 sentences with 73093 words.
Business Insider has 6961 sentences with 57544 words.
Atlantic has 7507 sentences with 67735 words.
CNN has 22087 sentences with 127629 words.

I then created a dataframe with new features. For each sentence, separate relevant words (words that are not stop words) from punctuation and stop words and capture each in a separate column. Calculate and store the length of the sentence, the number of relevant words, number of stopwords and number of punctuation items. Identify and store the parts of speech for each relevant word.   Count and store the total nouns, verbs, adjectives, adverbs, and interjections in the sentence.

After concatenating the results from the individual publications to a single dataframe, I created a training set with 75% of the data. The first test was to see if one could identify a publication based on the use of parts of speech and the number of relevant words per sentence. The scores using random forest, regression and gradient boosting were all very poor. Altering the training variables in various combinations yielded substantially the same score of roughly 30%. In all cases, the training and test scores were very similar.

The conclusion might be that there is no relation between parts of speech used and bias in the media OR there is equal bias in all sources as it is not discernible among publications.

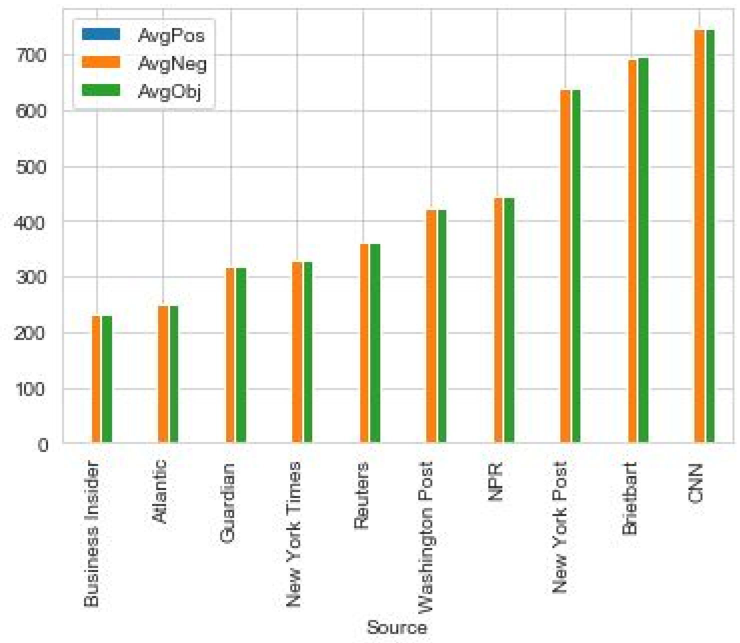
The next phase was to determine if there was any positive or negative leaning in the overall sentiment of the relevant words. I used WordNet SentiNet to assign a positive, negative and objective sentiment value to the lemma of each relevant word. In some cases, the lemma and part of speech identified with the spaCy tokenizer conflicted with available sentiment classifications in Word Net. These cases were dropped for purposes of this project.

For each tokenized sentence, each word was checked for scores and then accumulated in a temporary variable. The scores for all the relevant words in a given sentence were then summed and average for each of the 3 sentiment categories.

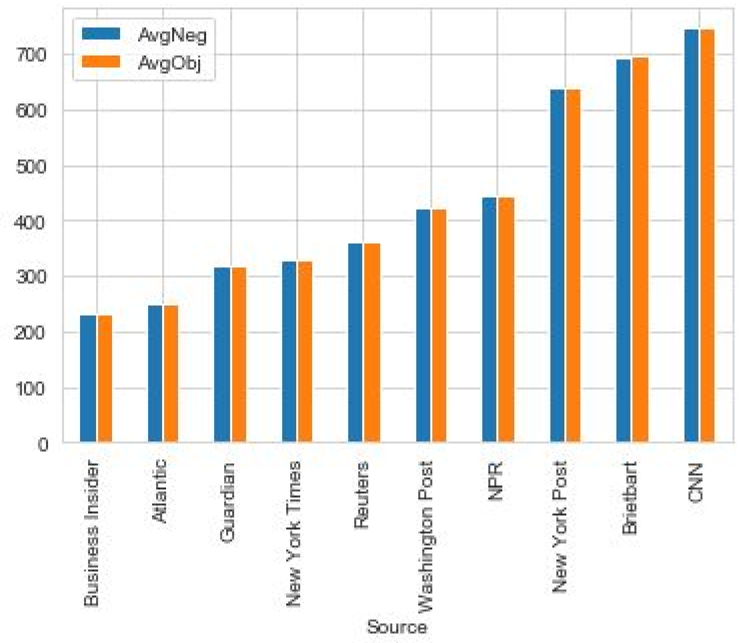
A new training data set using only word count and average sentiment in the sentence was created from 75% of the data. This time, random forest produced extremely high scores in the results with 96% for the training set and 81% for the test set.

The logistic regression results were still around 30%. The gradient boosting results were in the 70% range, substantially improved from analysis based on parts of speech.

Below is a summary of the sentiment results by publication.



It is clear that the relevant words were largely negative (AvgNeg). Further, it appears the objective factor is merely a calculation of the net negative and positive sentiment. This is as opposed to neutral sentiment. This is supported when viewing the information without the positive sentiment which is so small as to not be visible in relation to the negative sentiment.



Based on these results and findings, I reran the analyses using only the average negative sentiment per sentence, the number of relevant words per sentence, the number of punctuation per sentence, the number of adjectives per sentence and the number of interjections per sentence.

The random forest was again the best performer by far providing a score of 89% using the default 10 estimators and a negligible increase to 91% when specifying 500 estimators. Note that the increase to 500 estimators made a noticeable increase in processing time.

The logistics regression performed again in the 30% range while gradient boosting scored around 70%.

On a purely visual note, one might conclude that CNN, Brietbart and the New York Post articles contain a lot of bias based on high negative sentiment in the writing expression. Likewise, the conclusion based solely on visual inspection of sentiment is that Business Insider, Atlantic, and Guardian contain the most objective stories. That is, stories containing the least bias.

Conclusion

The random forest model produced the best results in terms of ability to predict from which publication an article may be sourced. The use of parts of speech to determine bias in an article is not feasible. The hypothesis that “flowery” speech in articles reveals bias is false. While sentiment analysis can provide an indicator of bias, the corpas are still very young and produce inconsistent results.

Potential Future Analysis

This project mainly produced suggestions for additional research that might provide insight into media bias based on article content.

Among the many additional features that could be considered are:

- The relation of average number of words per sentence by author and/or by publication
- Using combinations of sentiment and parts of speech
- Using combinations of certain punctuation with parts of speech and sentiment
- Using a layered approach that first identifies overall sentiment for an article; then uses a corpus that identifies slant or bias based on usage of specific phrases, e.g.: “slam dunk” out of sports context