Andrew Berry - NLP & Sentiment Analysis to Determine Media Bias

---

Abstract:

Using Natural Language Processing methodologies and sentiment analysis to determine/ classify media bias of major news outlets. Specifically on the topic of political views, the bias (polarity) may be determined as left leaning, neutral, or right leaning. The key analysis is done on observing the adverbs and adjectives of headlines/texts to determine such outcome.

Research Question:
Is it possible to determine/classify the type of media bias based on headlines and the body of text of articles using sentimental analysis?

Datasets:
I will mainly be using the New York Times API, as my key source in obtaining the data I will need for this project.

In addition I will be using the python library Newspaper, to scrape article metadata of other news outlets:

- CNN
- BBC
- FOX NEWS
- BREITBART

I will narrow down my scope of the project to recent US politics, and ignore other types of news.

A bulk of my time will be in the collection and cleaning of the datasets. Removing unnecessary data (Removing special characters) , and preprocessing into a nice dataset. For example, the NYT API produces JSON extracts that will be needed to convert nicely into a DataFrame or CSV. The Newspaper python library will also extract unnecessary data that will need to be thoroughly looked through, at the same time modified versions of the web scraper will be used for each media outlet.

I may also have to proceed with stemming and Lemmatization the text data, which will assist in the classification later on. Also, I will be removing the stop-words(the, we, their, etc) (words that have little significance to the context of the text) out of the text too, since stop words are usually the bulk of the words.

Project Flow:

The first big challenge I will need to tackle it trying to label or identify certain topics in my datasets. This is an unsupervised learning challenge, and I will be using clustering in determine to identify certain collections of topics in my datasets. In doing so, I will be required to create word vectors for each article. Then proceed by using a k-means algorithm for the clustering.

This part will take a significant time to fine tune the clustering as well as figure out the best way to feed and structure the data for the clustering algorithms.

Before proceeding with the sentimental analysis, I will need to train the model to identify the various aspects of a text, such as the nouns, verbs, adjective, and adverbs. This will be able to completed using NLTK or SPACY.

For the sentimental analysis I will be using third party lexicons (these are dictionaries that have been created by the community to assist in analyzing sentiments, they already have built in labels of positive or negative words, etc) such as the AFINN lexicon. A big obstacle will be taking the time to analyze the sentiment, understand it, and find tuning it. Below is an example of a quick sentiment analysis of positive vs negative article text data from the NYT API sample I was able to get. This is a simple start of determining sentiment of an article, however the challenge is comparing the sentiment of an article/topic to that of other outlet's article/topic, of the same topics.