SENTIMENT ANALYSIS

PROJECT PROPOSAL

IST 652 Scripting for Data Analysis
Katherine Hurtado-da Silva

How are political affinities influencing news reporting via Twitter?

Source: https://www.allsides.com/media-bias/media-bias-chart

Friday, May 13, 2022

DATA DESCRIPTION

Fox News, BBC, and CNN are considered far-right, centrist, and far-left news outlets, respectively. To gather the data needed for the analysis, each news outlet's twitter account was scraped for the last 1,500 tweets, as of May 13, 2022. Three dataframes containing the content of the tweets were made and will be used to conduct a sentiment analysis for each.

RESEARCH QUESTIONS

The analysis will attempt to address the following questions in hopes of generalizing the tweet-based reporting differences between Fox, BBC, and CNN news:

1. What do the "like" counts for tweets indicate about each news outlet's audience?
2. How did the reported topics for each news outlet differ?
3. How did negative, positive, and neutral comments vary according to each news outlet?

DATA PREPROCESSING

The tweet content will need to be turned to lower-case letters and tokenized. The links, hashtags, punctuations, and stop words will be removed via Re and NLTK and used for a sentiment analysis. Any entries with missing content will also be removed.

METHOD OF ANALYSIS

WordCloud(s)- I will attempt to make a word cloud that is based on a token's frequency at minimum. Ideally, it would be neat to be able to get the top 50 frequent tokens, but have its size be based on the number of "likes" its respective tweet had.

K-Means Clustering- This approach will be attempted to be used to group the tweets into groups that reflect its general topic. Hopefully, this will be effective and I'll be able to show the visualization of the clusters.

Sentiment Analysis- After looking up different sentiment analysis methods, TextBlob seemed to be the best approach because it is able to provide subjectivity in addition to sentiment scores. Since the nature of Fox, BBC, and CNN news is being compared to see the influence of its political inclinations, the subjectivity would be helpful to measure the amount of personal opinion and factual information contained in the tweets.

SCOPE & CHALLENGES

The analysis is going to involve K-Means Clustering, Sentiment Analysis, and visualization methods not addressed in class. This, in addition to scraping tweets, goes beyond what is expected in the homework assignments. However, I will need to search for code related to each of these methods to be able to execute the analysis.