ETL PROJECT

Are certain news outlets more biased against Trump?

Jessica Hills

March 24,2019

PROCESS OVERVIEW

- I wanted to generate my raw data from web scraping several websites
- I then wanted to store all of that information into a SQL database
- I also wanted to use an API to generate data
- The API chosen was used to evaluate the sentiment of both the headline and summary statements of the news articles that were scraped
- The analysis was then done to evaluate the results

WEB SCRAPING

- For this part of the project I wanted to scrape 3 different news outlets
- I chose 3 different news outlets that have different reputations for their biases
- For the neutral bias I chose CNN
- For the “left” or “Democratic” bias I chose MSNBC
- For the “right” or “Republican” bias I chose Fox News

WEB SCRAPING PROCESS

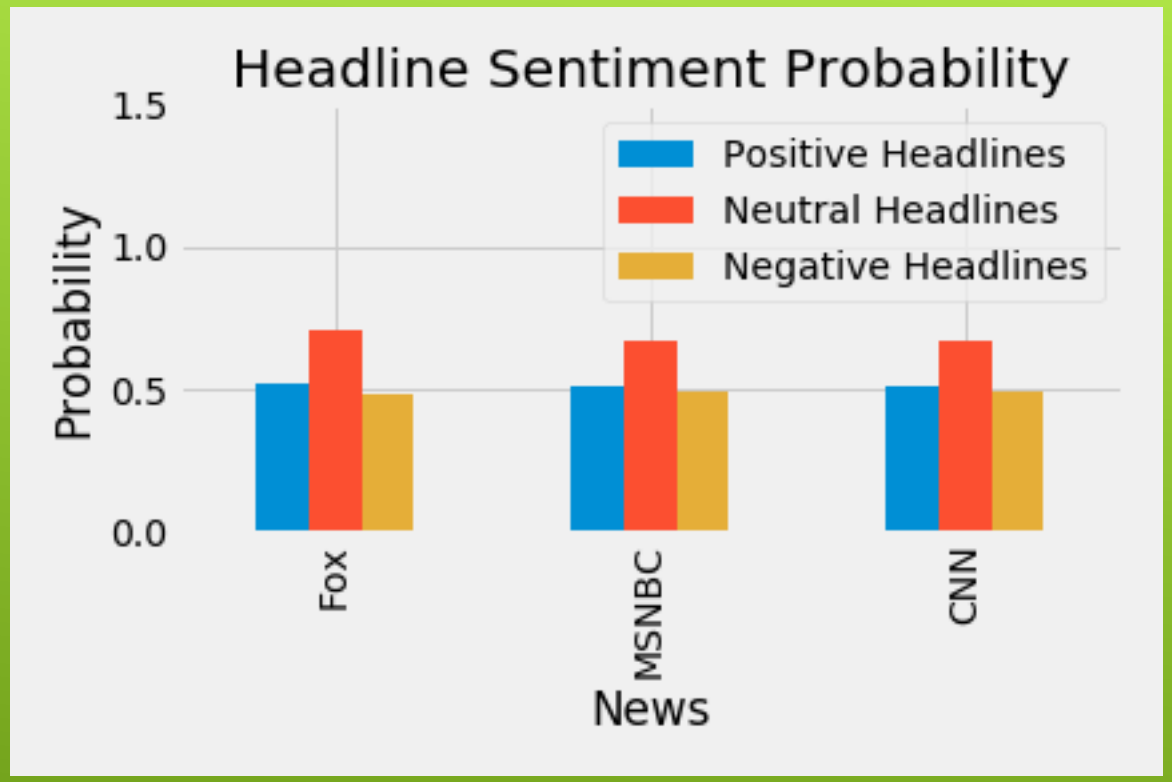
- To create my web scraping code I first inspected all three of the websites for their similarities
- Fortunately all 3 websites were structured in a very similar way so I was able to write one base script which then only required slight modifications for each news outlet
- I wanted a substantial amount of data because proving bias can be very different based on the sample size, too small and that could yield inconsistent results, so I chose to try and capture around 7500 articles
- Overall the data was rather clean due to the effort put into writing the code that scrapped and collected the information.
- There was only a small amount of transformation required to convert the probability numbers into actual numbers to build the bar graphs

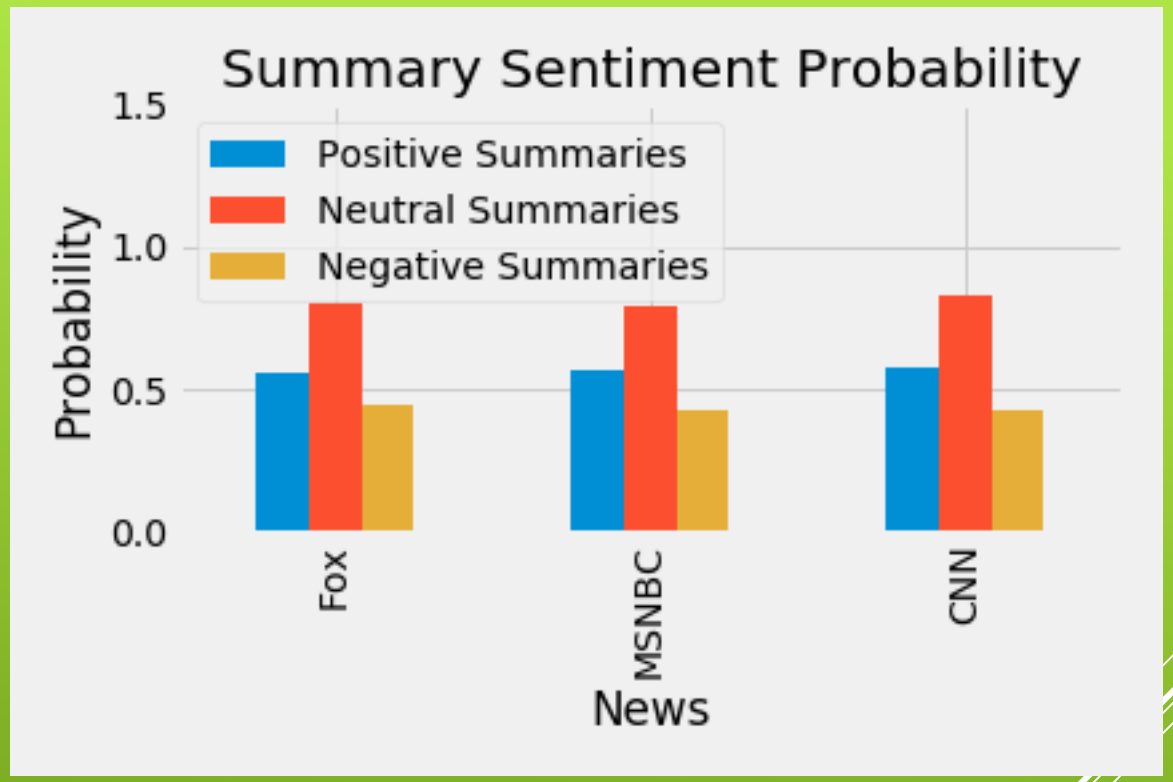
API CALLS

- To gather information on the sentiment of the headlines and summary statements I found a sentiment analysis API on RaidAPI
- After I had scraped all of the headlines and their associated summaries I then made the API calls to gather the probability information
- I chose to capture and record the actual probability numbers because I thought that would give the most flexibility for the analysis

SQL DATABASE

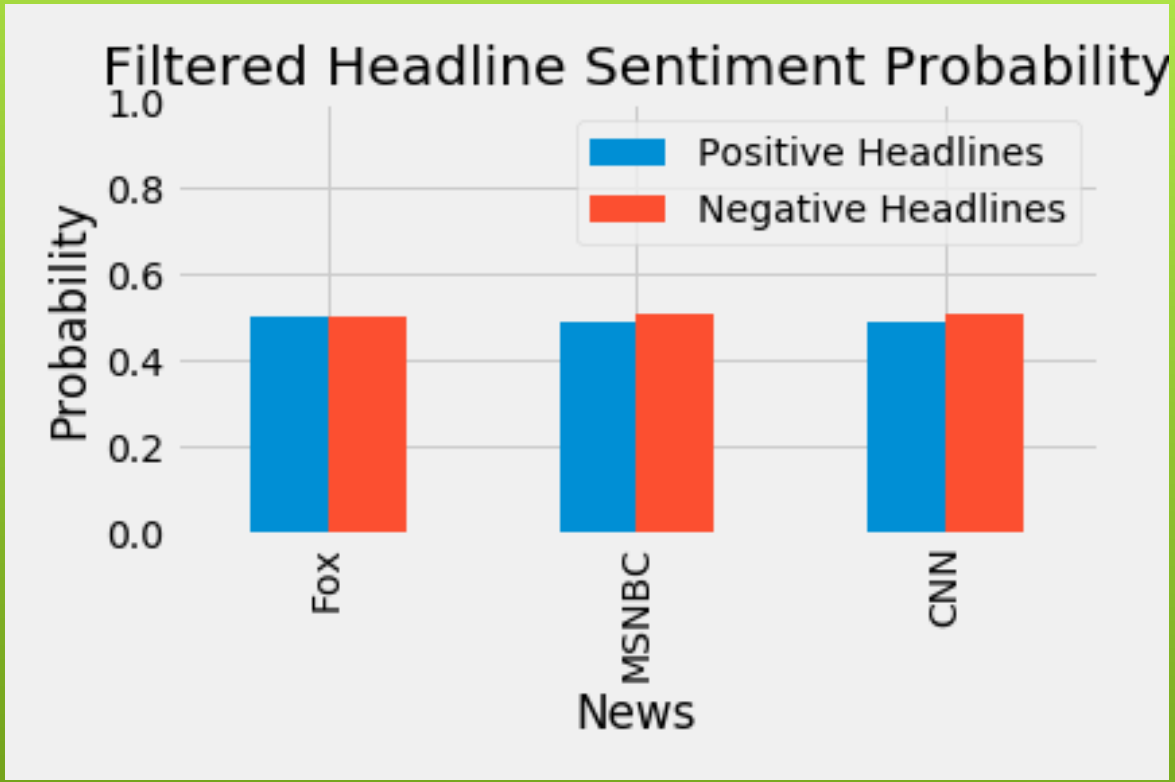
- To store my information I created 3 tables, one for each news outlet
- After I scraped the data and made the API calls to gather the sentiment data I committed all of the data to the correct table in Postgres

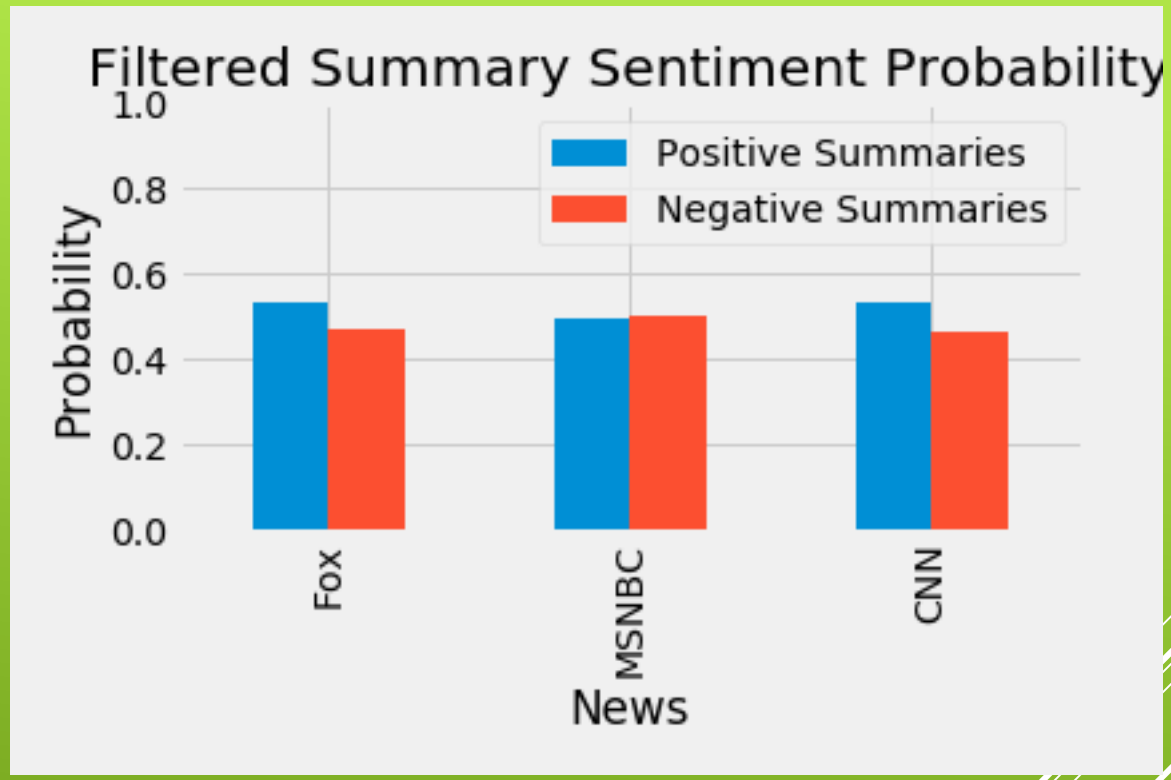




- For my first analysis I looked at all of the headlines and summaries as a function of the average of all of the probability values
- For the headlines all of the news outlets seemed very similar
- For the summary data, all of the news outlets also seemed very similar

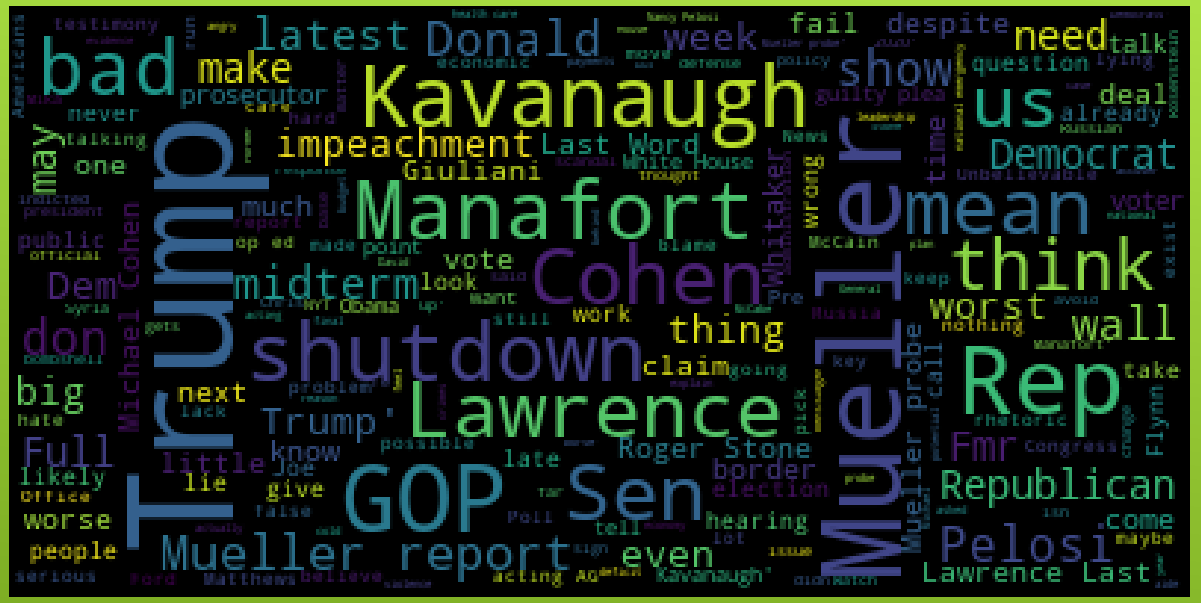
DATA ANALYSIS

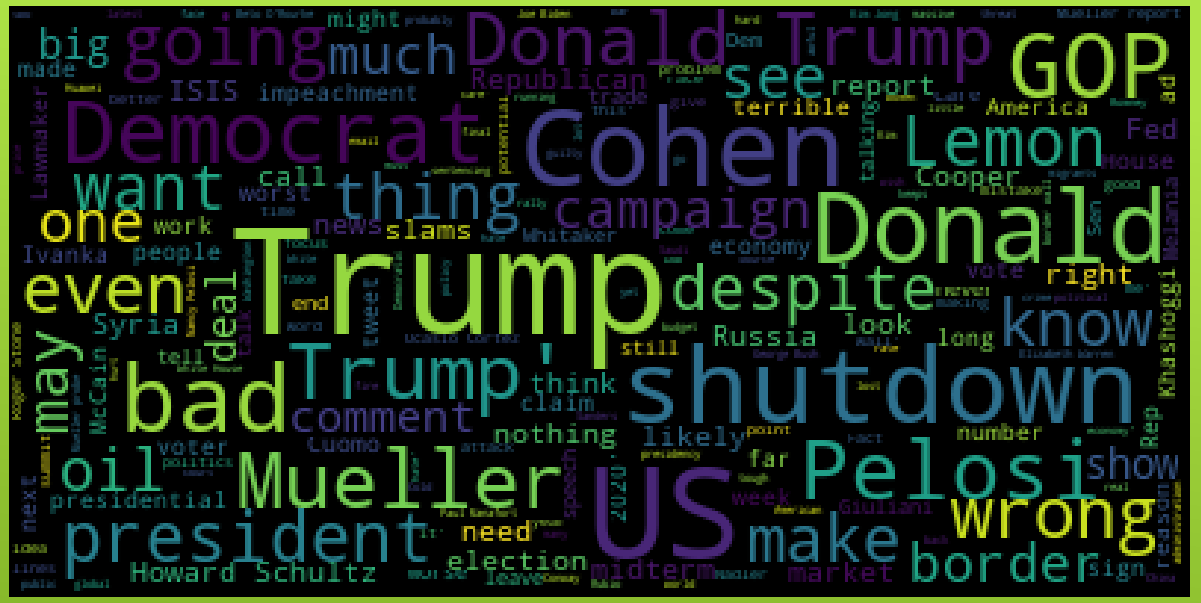




- For my next analysis, I filtered out all of the neutral results and looked just at the statements that had a positive or negative sentiment
- For the headlines, MSNBC and CNN are almost the same and both have a slightly higher likelihood of having a negative sentiment
- For the summary data, Fox News and CNN are very similar and have a slightly higher likelihood of having a positive sentiment

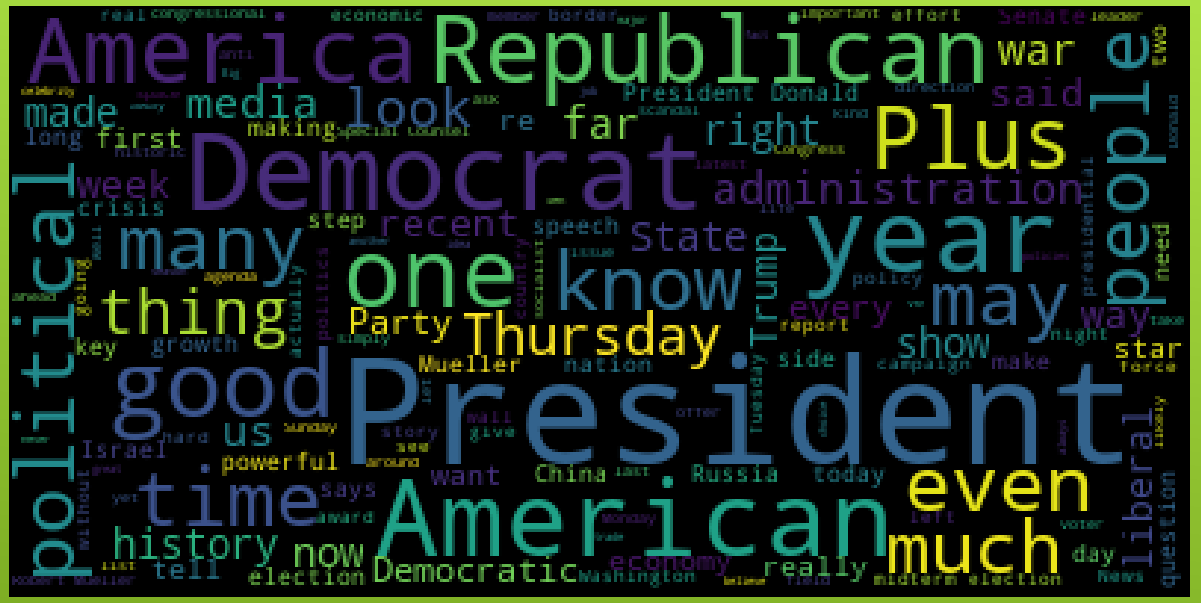
DATA ANALYSIS

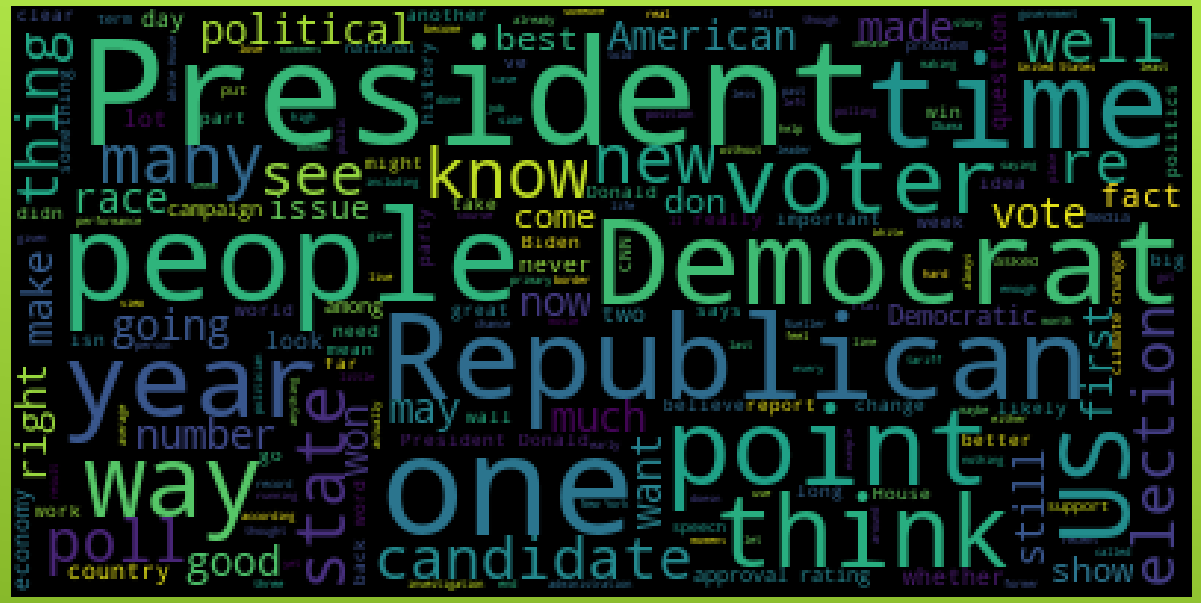




- For my next analysis, I looked at the results from the previous charts and I wondered if there were common words between the negative headlines from MSNBC and CNN
- I made word clouds from the text that was scraped, on the left is MSNBC's negative headlines and on the right is CNN's negative headlines
- There are several word in common here but what I think is most significant is the number of names that appear, I think this is due to the fact that the names are the most common thing that is associated with a negative headline

DATA ANALYSIS





- For my last analysis, I looked at the positive summaries from Fox News and CNN
- I made word clouds from the text that was scraped, on the left is Fox News' positive summaries and on the right is CNN's positive summaries
- There are also several similar words here, however unlike the previous ones the types of words are very different. We see many more words like president, democrat, republican etc. rather than names of people

DATA ANALYSIS

FINAL OBSERVATIONS