DS 5001 Final Project
Anoop Nath (nux9aq)
12/9/2022

Text Analytics of News Articles from Power Line and Politico from 2013 to 2020

The news that Americans consume has become a controversial topic over the last 5-10 years. There have been studies that have shown that misinformation and polarization in the news are on the rise. For this project, I was interested in performing text analytics on news articles from both left-wing and right-wing outlets. I performed text analytics on the Newzy data set, which is a collection of news feeds and stories from conservative and progressive sources from October 2013 to February 2020.

In particular, I focused on articles from two sources: Power Line and Politico. In total there were approximately 9,700 articles from PowerLine and 3,700 from Politico. Going into the project, I knew very little about these two news outlets. Ideally, I would have liked to incorporate more outlets, however, the dataset only had full articles for these two companies. I will be using principal component analysis, topic modeling, word clouds, and sentiment analysis to get insights into these two sources and see if either or both have a political bias. Knowing nothing about these two sources when I first started the project, it was fascinating for me to explore, compare, and contrast without any preconceived notions.

Word Clouds

After cleaning the data, which included things like lowercasing all text, removing digits, removing stop words, etc. the first analysis I performed was to create word clouds for both outlets to see if there is a significant difference in the top words.

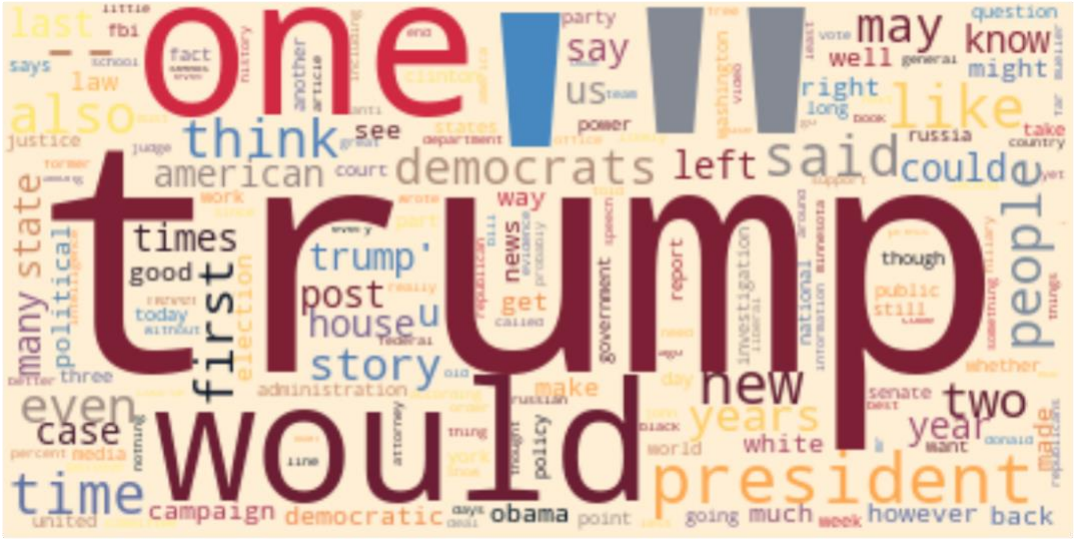


Figure 1 Power Line

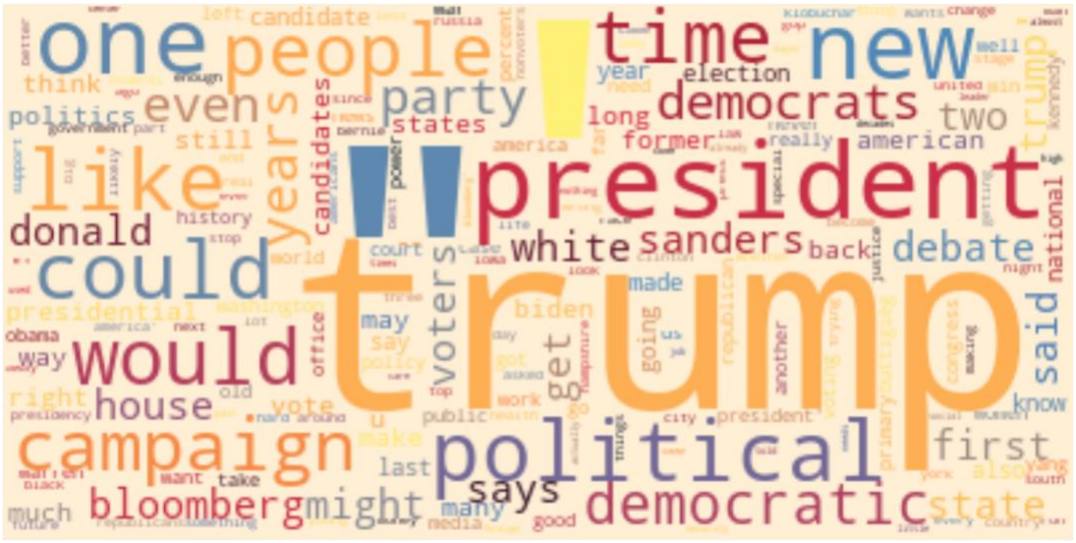


Figure 2 Politico

However, the word clouds for both look very similar with the largest words being related to politics such as Trump, Obama, president, campaign, etc. It's hard to tell if there's any political bias for these two outlets from these.

PCA

The next analysis I performed was principal component analysis. Using the Text Hero package in python, I was able to calculate and plot the first and second components for both sources.

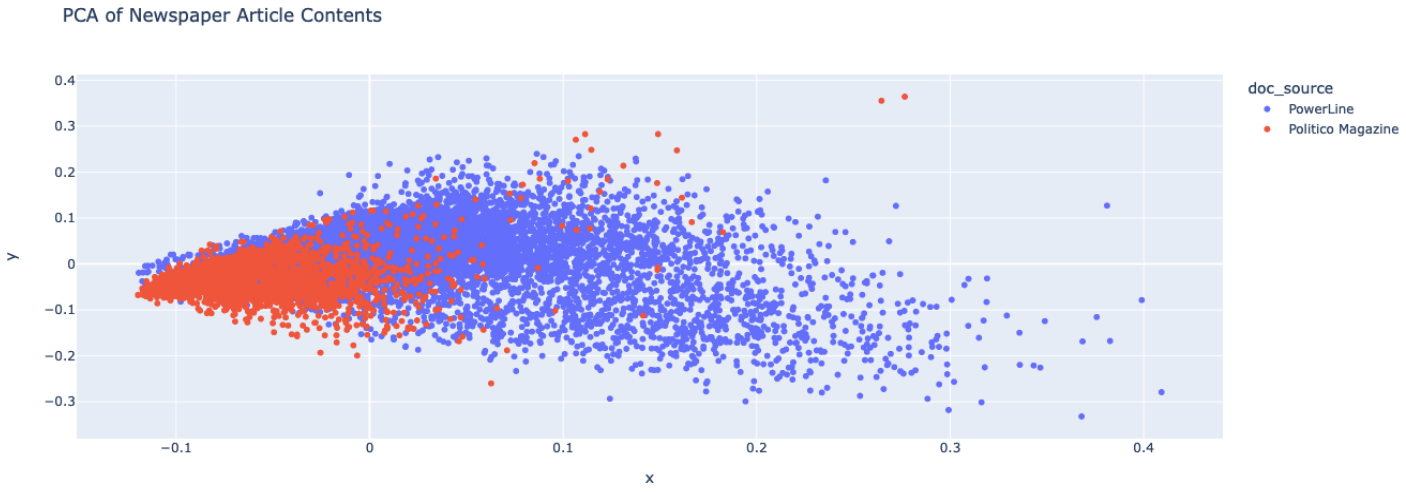
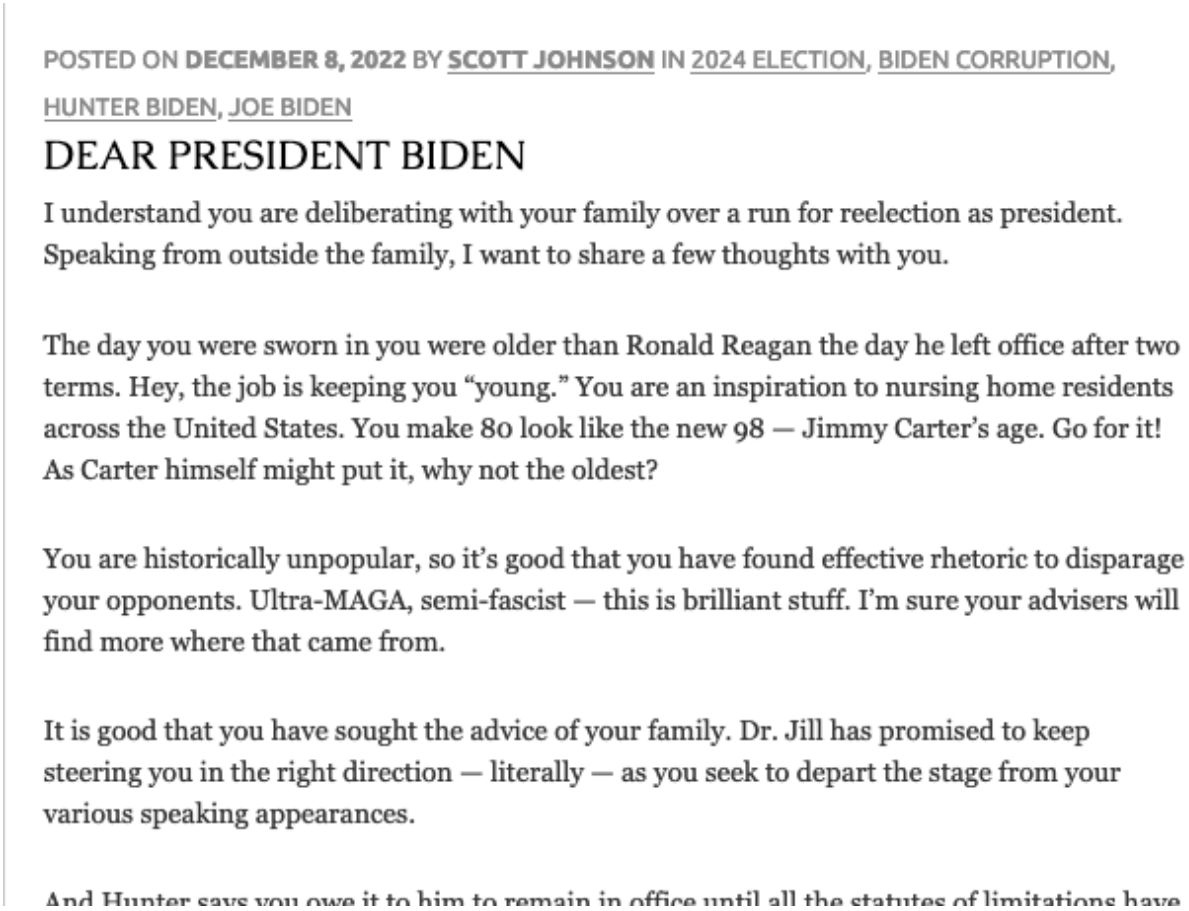


Figure 3 PCA analysis

From the PCA chart above, it looks like Politico has very little divergence in terms of the second component (y-axis). Meanwhile, Powerline has divergence in both the first and second components. My hypothesis is that since Power Line is a far-right new outlet, it tends to put out more sensationalist stories than Politico, which is a more established, left-leaning, news outlet.

Browsing the Power Line website, only adds confirmation to this theory as this was one of the "articles" on the front page.



POSTED ON **DECEMBER 8, 2022** BY **SCOTT JOHNSON** IN 2024 ELECTION, BIDEN CORRUPTION, HUNTER BIDEN, JOE BIDEN

DEAR PRESIDENT BIDEN

I understand you are deliberating with your family over a run for reelection as president. Speaking from outside the family, I want to share a few thoughts with you.

The day you were sworn in you were older than Ronald Reagan the day he left office after two terms. Hey, the job is keeping you "young." You are an inspiration to nursing home residents across the United States. You make 80 look like the new 98 — Jimmy Carter's age. Go for it! As Carter himself might put it, why not the oldest?

You are historically unpopular, so it's good that you have found effective rhetoric to disparage your opponents. Ultra-MAGA, semi-fascist — this is brilliant stuff. I'm sure your advisers will find more where that came from.

It is good that you have sought the advice of your family. Dr. Jill has promised to keep steering you in the right direction — literally — as you seek to depart the stage from your various speaking appearances.

And Hunter says you owe it to him to remain in office until all the statutes of limitations have

Figure 4 Recent article from Power Line

This reads more like a blog post someone wrote from their basement than a news article. There are other articles that are news related but this variety of content could explain why the PCA chart shows more variation in both components for Power Line.

Topic Modeling (LDA)

Next analysis I did was to create a topic model using Latent Dirichlet Allocation (LDA). The gensim package in python makes this very simple to calculate and visualize.

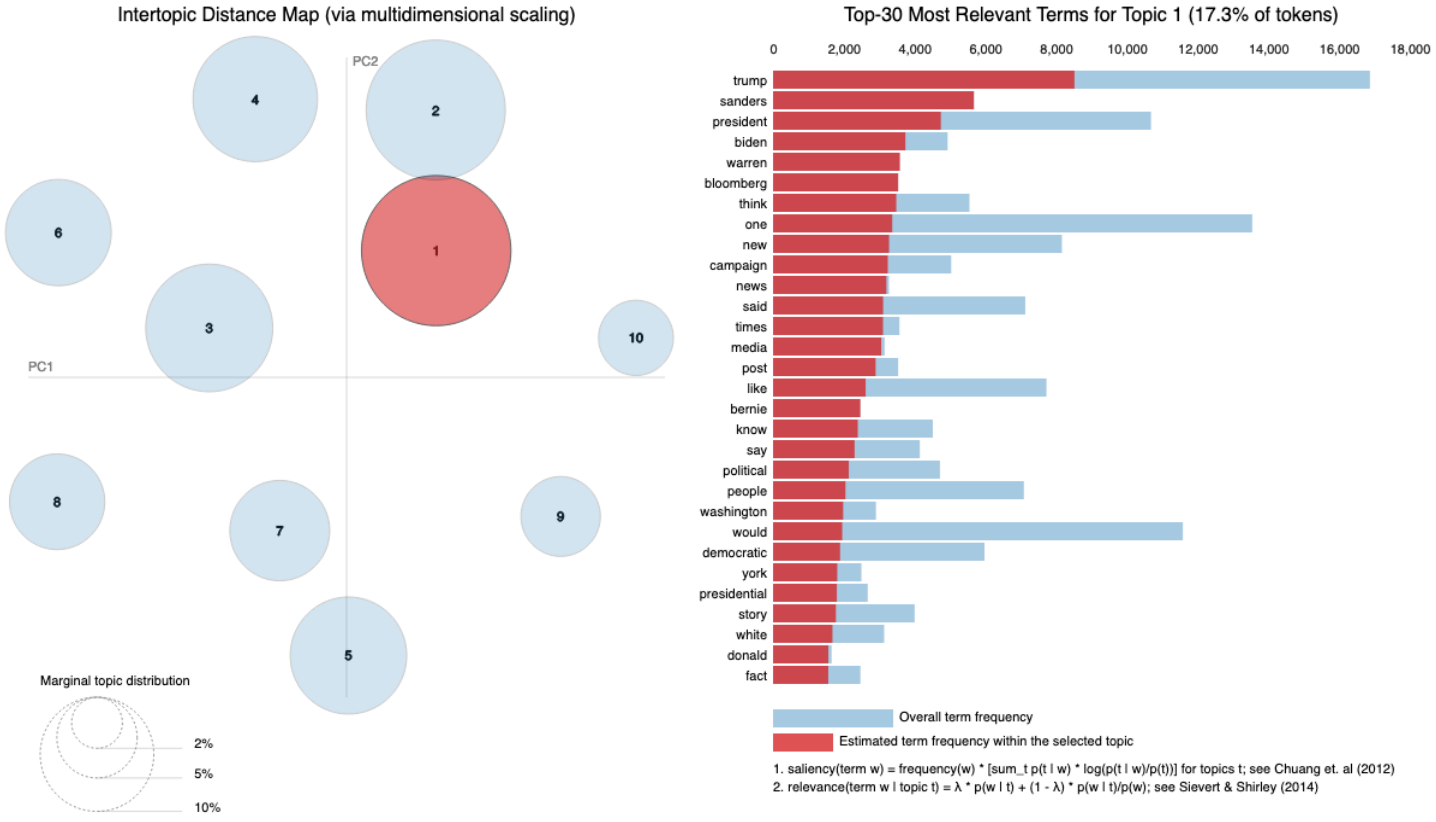


Figure 5 LDA topic models

We can see some very clear topics. For example, topic one (shown above), has words such as Trump, Biden, Sanders, president, Warren, and campaign so this seems to be related to the presidential election.

Topic eight has words such as Iran, Ukraine, Russia, and China so this topic seems to be related to foreign affairs.

Meanwhile topic ten has words such as Omar, Ilhan, and Minnesota. This topic is clearly about Ilhan Omar, who is a U.S. representative for Minnesota's 5th congressional district. I'm surprised to see that there is an entire topic for her because though polarizing, she is not at a high level of power within the U.S. government. However, it is worth keeping in mind that there are nearly three times as many Power Line articles than Politico, and the right-wing media has made it a hobby to attack minority women in Congress.

Sentiment Analysis

The final analysis I performed was sentiment analysis on both sources of articles. I used the NRC Emotion Lexicon, which associates each word with 8 different emotions.

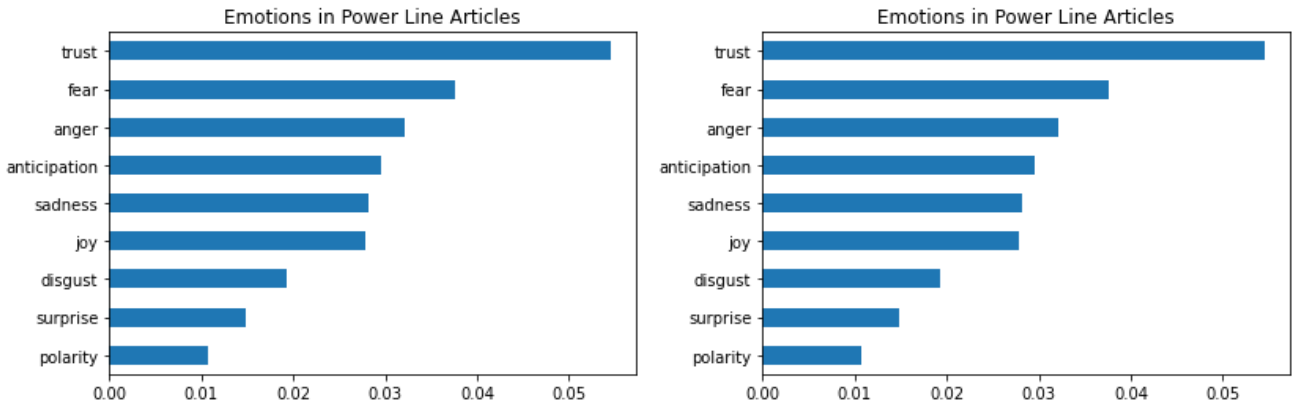
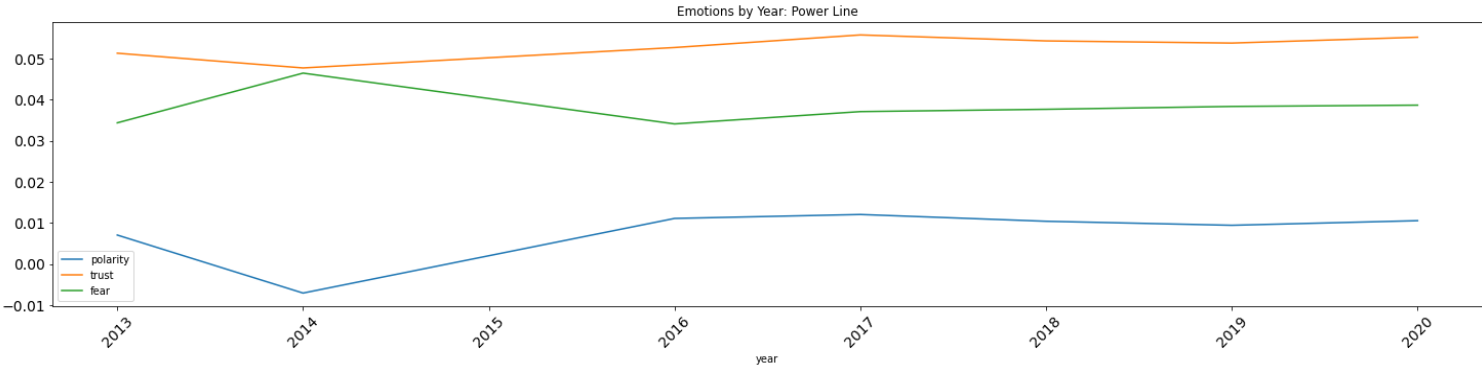


Figure 6 Emotions in Power Line and Politico

The charts above that both Politico and Power Line have similar sentiment in their articles. The top four emotions in both are trust, fear, anger, and anticipation. However, the polarity for Politico is slightly higher. Meanwhile, negative emotions like anger, fear, and disgust are all higher for Power Line compared to Politico. This lines up with my experience of browsing the two sites as well.

The texts for both sources span from late 2013 to early 2020. I wanted to explore if there has been changes in sentiment over the years for each news outlet.



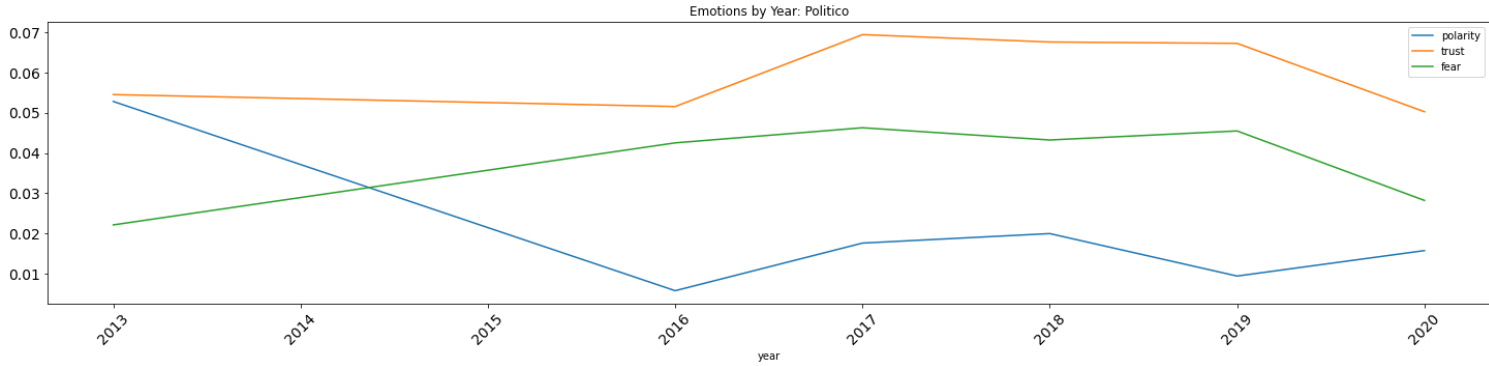


Figure 7 Emotions by year

There is a very clear trend in the two charts above. For Power Line, we can see that polarity was very low from 2013 to 2015, then increases from 2016 onwards. Meanwhile, there is a dip in fear and an increase in trust starting in 2016.

We see the opposite trend for Politico, with polarity dropping and fear increasing from 2016 onwards. Although you do see an increase in trust in 2017.

It's no coincidence that we see such sudden shifts in sentiment in 2016, which was the year Donald Trump won the presidential election. From these charts, we can infer that Power Line is a right-leaning source while Politico is left-leaning.

Conclusion

I really enjoyed using all the tools I have learned throughout this class into one final project. I started this project with no idea what the dataset or articles looked like but finished with a very good sense of what it contains. I was also able to compare and contrast the contents of articles from both Power Line and Politioc. I was able to use word clouds, PCA, topic modeling, and sentiment analysis to accomplish this.