Analysis of Word Choices in News Articles

Brandon Wong and Johnny He

Table Of Contents

1. Summary of Research Questions

- Is Fox News more anti-mask than CNN?

  For this question, we will explore different libraries than we were exposed to in class to sort out relevant words in a news article and generate a word cloud for the most popular words in the article. We plug 30+ COVID-related articles each from both Fox and CNN into the code and answer the question by comparing the resulting Word Cloud.

- Which source is the article?

  We pass in an article and ask the ML classification model to assess whether the article is more likely from Fox or CNN. We will use Random Forest to extract important information from words and sentences to provide necessary features to predict the label(origin of new source).

- Is there an inherent difference in writing style between CNN and Fox News?

  We conducted statistical analysis to find if there is a significant difference in the means of the number of times a term is mentioned in a new sources' article content by developing a null hypothesis and an alternative hypothesis. We first observed the variability of the statistics through graphing the frequency of terms to notice any outlier articles and if the observations are independent between the samples. Then we compute the test statistic versus the null statistic to determine whether we reject the null hypothesis or not.

2. Motivation

With the ongoing Covid-19 pandemic, news sources report updated public safety guidelines, interview the public for their opinion, and follow relevant storylines. However, over time, the pandemic itself has become highly politicized as people have differing views on how public safety should be addressed. Furthermore, news sites report information using different diction based on their political affiliation which caters to their specific audience. We noticed this during previous instances during the most recent presidential election where the two most prevalent news sources were CNN (left-wing) and Fox News (right-wing). The significance behind these differences in reporting is its power to influence people and their views. The danger of this is that potential misinformation can spread which creates tension among the public. We are interested in exploring exactly how this process occurs and hopefully making journalists more accountable for the articles they write.

3. Dataset

Our original dataset consists of articles we found from the Fox and CNN websites. Before scraping the content of these articles, it was best to access them via their link so created two datasets. Each data set consists of over 30 articles. We also added additional columns such as article ID, if we needed to find a particular article, article content, where planned to place the scraped text, and the news source, which will help us keep track of the article's origin when we later combined the datasets to train the machine learning model. Below, you can access each original dataset via the links.

Fox:
https://docs.google.com/spreadsheets/d/1RHGUDfBY_BKG0Akcu2TBmJ2p8sYy4kU-JwBMFc-2vvw/edit?usp=sharing

CNN:
https://docs.google.com/spreadsheets/d/1sSfQqVf0xSb2hpkM0IBYQGn_oEIKdPbyqg3lCrW9rbo/edit?usp=sharing

4. Method

To analyze the sentiment of news articles, we first apply **web scraping** to obtain the text contents of the **CNN and Fox sites** that are about COVID. 30+ articles' content from each news source will be placed in two ".txt" files. Then using code to develop visualization techniques, we get the most relevant words in both text files and generate a word cloud. By comparing the word clouds from two different news sources, we can see the most frequently occurring words or phrases, therefore understanding each side's attitude about COVD. This will help us answer our first research question about which news source has more anti-mask sentiment.

Next, we used sklearn to create a RandomForestClassifier to understand the preferred "writing style" of each news source. Insights from their writing patterns act as features that help us identify which news source the article is coming from, which acts as our label.

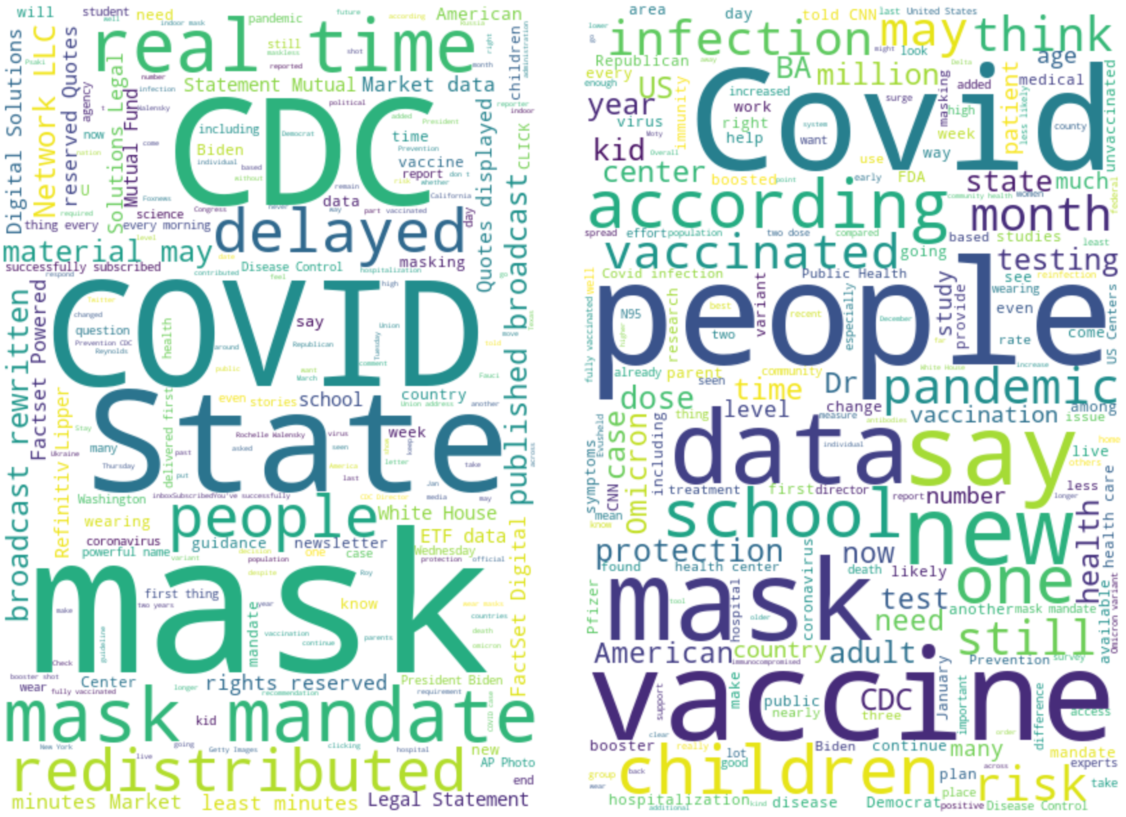
Lastly, we plan to conduct a statistical analysis to compare the two news sources' and their most frequent terms. This will help us to learn whether there are significant differences in CNN and Fox writing styles.

5. Result

- Is Fox News more anti-mask than CNN?

  To answer this question, we used the news_scrape.py file, to reformat our original dataset of URLs into formatted columns in a data frame and add additional columns. This process was executed by using a function named get_links that turns the original news source datasets into a list to be passed into word_scrape_fox and word_scrape_cnn where we found the article content by specifying the paragraphs we wanted from each article which would be all the text from the article. Then, using to_txt we converted these data frames into separate text files for CNN and FOX. The make_word_cloud file has separate functions to create a FOX and CNN word cloud that finds the most prominent words from each sources' Covid-related articles.

Figure 1. FOX (left), CNN (right)

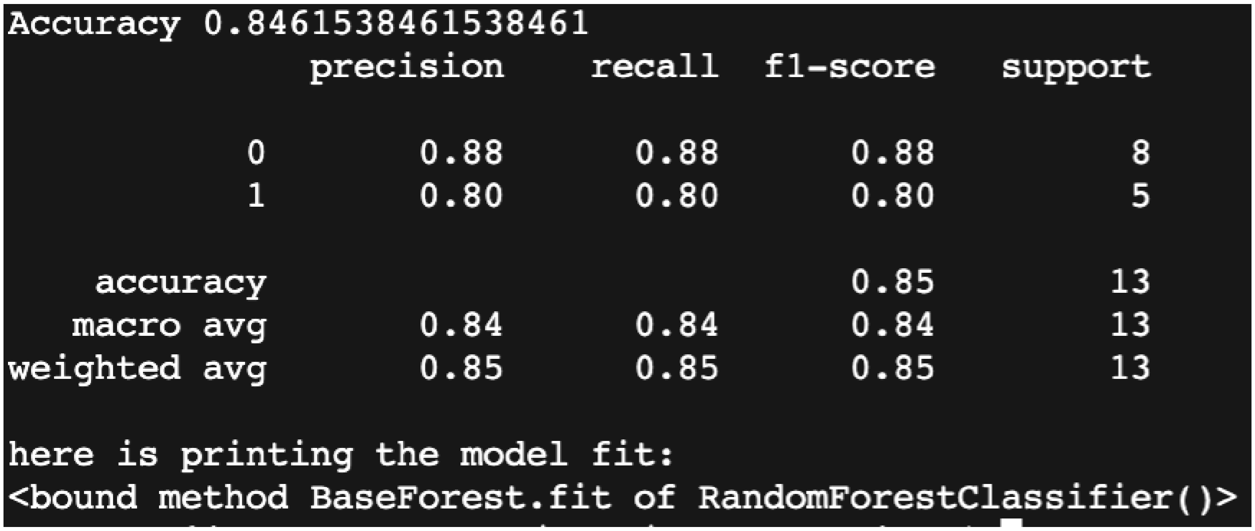


Some of the most prominent words from both news sources are ‘Covid’, ‘people’, and ‘mask’. Yet one significant finding is that Fox tended to have ‘mask mandate’, ‘CDC’, and ‘State’ be a more frequent term to appear while ‘vaccine’, ‘children’, ‘data’, ‘school’ was more prevalent with CNN articles. While these visualizations tell us the most prominent words, it's difficult to concretely determine which of the two is more anti-mask. Both news sources have a high frequency of the term ‘mask’ but there is insufficient information as to the context ‘mask’ was used. We see that ‘mask mandate’ appears more frequently in FOX which most likely refers to the situation where mask mandates were repetitively lifted and reinstated across multiple States between March 2020 and the present. We may also infer that FOX is more anti-mask as another prominent term is ‘delay’ which may refer to mask mandates and the CDC’s constant revisions to safety guidelines. Conversely, we may infer that CNN relies more on covering vaccine information with its high frequency of that term and of the term ‘data’. While those terms are directly anti-mask, CNN tends to focus more of its Covid coverage on vaccine information. This indicates a slight more anti-mask coverage by FOX based on the inferences from earlier.

- Which source is the article?

  For this research question, we built and trained a Machine Learning model in our analysis file using a Random Forest Classifier model. We were able to successfully train the model to distinguish between CNN and FOX articles based on a joined dataset. This model involves using a TF-IDF vectorizer that works similarly to the Search Engine. Initially, we had a 100% accuracy score when we set the max_feature parameter to 300. Our TA Esteban explains the max feature specifies the number of most relevant words we include in our features. Each news article includes phrases like “CNN reporter discovers,” “Fox News journalists observe.” As these are the only words included in our features since they are the most relevant ones, it makes sense that we were overfitting and obtained that perfect accuracy. To fix this, we first adjusted the max_feature to “None” so that the training algorithm accounts for the top most relevant words. We also imported a library named “newspaper,” used it to summarize every article’s content, and saved them in a new column. We then used the content of the summary of each article as new features to train our algorithm. This eliminates keywords like “Fox” or “CNN” and only focuses on the article content itself. Now our model has an accuracy score of 84%. One remaining flaw of the model is that since we used the summary of each article to train the algorithm, we lose focus of each article’s unique writing style like diction, grammar, etc.

The inherent bias of the summary algorithm also influences the training results.



```
Accuracy 0.8461538461538461
              precision    recall  f1-score   support

           0       0.88      0.88      0.88         8
           1       0.80      0.80      0.80         5

    accuracy                           0.85        13
   macro avg       0.84      0.84      0.84        13
weighted avg       0.85      0.85      0.85        13

here is printing the model fit:
<bound method BaseForest.fit of RandomForestClassifier()>
```

- <u>Is there an inherent difference in writing style between CNN and Fox News?</u>

Based on the most prominent words we found in the word cloud, we decided to find the average number of mentions of the terms ‘vaccine’ and ‘mask mandate’. But, to assess if there is a relationship between these averages, we developed a null and alternative hypothesis. The null hypothesis is that there is no difference in an average term mentioned from CNN versus Fox. The alternative is some difference in the average of a term’s mentions from CNN and Fox. This was important for us to determine given that ‘vaccine’ had appeared more often in CNN articles while ‘mask mandate’ appeared more in Fox. So, the below graphs in Figure 2 display the number of times a given term was mentioned in an article, and the y-axis shows the number of articles that included x number of mentions of a given term which was created in our analysis.py file using the frequency_catplot function. This applies the count_vaccine and count_mask_mandate functions to create new columns of the total number of times a term was mentioned in a given article. We then calculated the sum, mean, and standard deviation of each term for each news source in the summary_stat function in analysis.py by printing out these statistics of the ‘vaccine count’ and ‘mask mandate’ columns. The result of these findings is shown in Table 1 and Table 2. One aspect to consider is the standard deviation for CNN with the term ‘vaccine’ is quite large relative to Fox which could be partially influenced by outliers. In Figure 2, there is one CNN article that had 41 mentions of ‘vaccine’ and the articles overall had more variability in the number of mentions which could explain the large standard deviation. Moving forward, we used the equation below to calculate the T-test to compare the means of a given term between CNN and Fox.

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

So, the observed difference in the means for the term ‘vaccine’ was 3.07 (numerator) while the standard error was 0.86. This gave a T-test of 2.2. We also determined that the appropriate significance level was 0.05 and to do a single tail test with the degrees of freedom the minimum n minus 1 which was 38. Then, using an online calculator, the p-value was 0.034. Given that the p-value is lower than the significance level, we reject the null hypothesis, meaning there is convincing evidence of the difference in average mentions of the term ‘vaccine’. This means that there is likely more coverage of the topic of vaccines for consumers of CNN content which may mean that their audience is more well informed of the topic as compared to Fox. This also may factor into how many people of each audience are vaccinated. If one news source is covering the topic more, more information about vaccines may convince these people to take one.

Conversely, we also applied the same methods for measuring the difference in average mentions for mask mandates which are displayed in Table 2. The observed difference between the means for the term ‘mask mandate’ was 1.96 and a standard error of 0.41. This produced a t-test value of 4.78. The degrees of freedom we used was 9 given that we are following the same procedures of using a significance level of 0.05 and conducting a single tail test. The calculator then outputted a p-value of 0.001 which also meant that the average ‘mask mandate ’ mentions were statistically significant. While we now know that Fox appears to have a higher average of the term ‘mask mandate’ this means that they have more coverage of the topic. Their audience is more informed of these policies that

State governments have enacted which could mean they are anticipating when these mandates are over. From prior knowledge, generally, people who identify as right-wing are more against the mask mandates as it harms the economy. What we may have found is that using different terms means that different topics of the pandemic are covered which caters to their respective audiences.

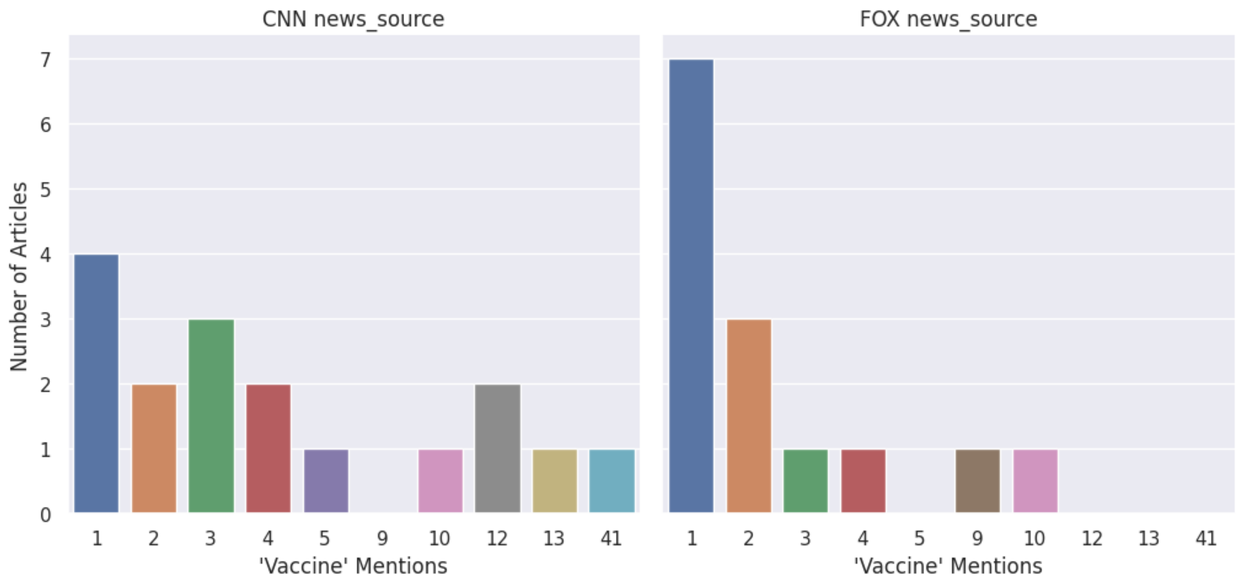
Table 1. ('Vaccine')

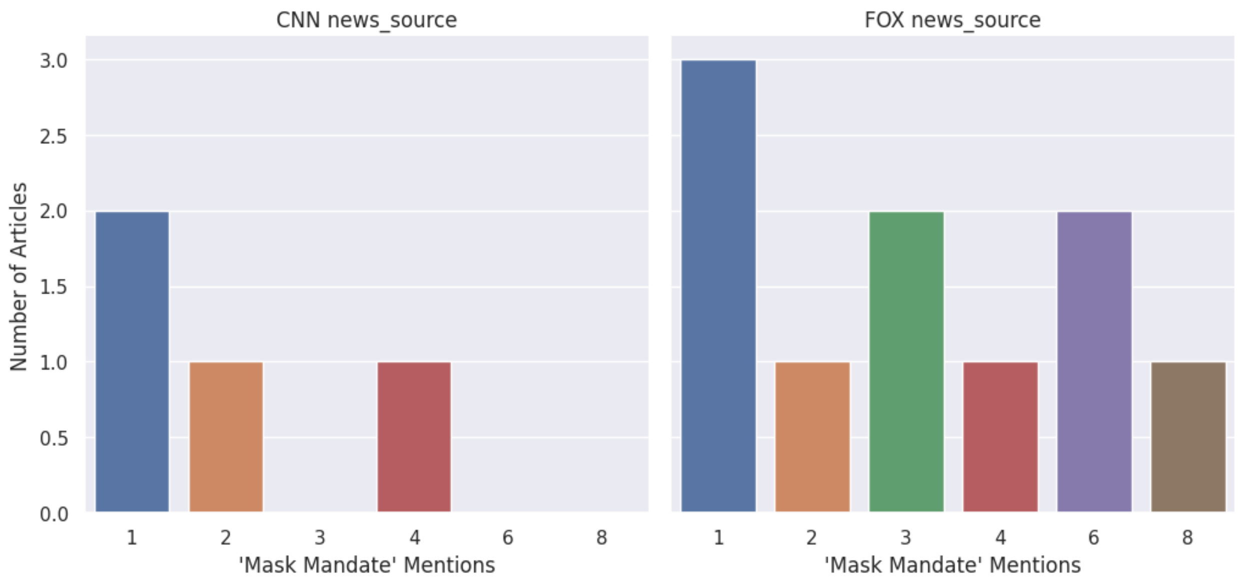
News Source	n	Mean	Standard Deviation
CNN	118	4.37	8.36
Fox	39	1.3	2.45

Table 2. ('Mask Mandate')

News Source	n	Mean	Standard Deviation
CNN	10	0.37	0.8835
Fox	67	2.33	2.4023

Figure 2.





6. Impact and Limitations

Some of the implications of our findings are that it is difficult to fully conclude any of our findings to our research questions without having a healthy skepticism. We tended to find terms that correlated with being more anti-mask but it does not necessarily cause some of the sentiments that people have towards the Covid pandemic. For instance, the word cloud does not give context as to the viewpoint of the article and how the term was used which does not tell whether a news source was talking positively or negatively about a term.

An implication if our Machine Learning Model was used to predict the source is that it may overgeneralize a news sources' diction. In a scenario where more people believed that Fox is more anti-mask because 'mask mandate' appeared more frequently in their articles, this may lead to misguided stereotypes about a news source which can be, in the extreme case, considered defamation. Alternatively, based on the accuracy of our model and the inherent differences we did find between the writing styles of both news sources, this also raises concerns since there are already political affiliations with these news sources that inherently divide people. Yet, this divide is further enhanced by the writing styles these writers and companies choose to produce which also points to ethical concerns.

7. Challenge Goals

- Messy Data

  To get our text data for later analysis we first needed to scrape from online news articles. This turned out to be the more challenging aspect of our project. Locating the precise CSS selectors for each article was difficult. We wanted to scrape all the results of COVID-related articles but for some reason, CNN does not allow us to scrape article links from their search, so we had to put all the links in a CSV file and process the links through the CSV file.

- Machine Learning

  After we understand which words are most associated with each news source, we would like to predict the news source of an article. To challenge ourselves, we want to experiment with using different machine learning algorithms. In class, we learned about classification(supervised) models, but we would like to try a clustering model which is unsupervised for our third research question.

- Statistical Analysis

  We used the statistics module function in Python to calculate the standard deviation for our dataset. Additionally, conducting the t-test required extensive research as to how to calculate the necessary statistics to find any significant difference between the means.

7. Work Plan Evaluation

Our general work process was followed as a guideline rather than a strict schedule. This made it so that we met up between three to four times each week. The first week was mainly understanding how to web scrape the news sites and store them in a CSV format. This step took much longer than anticipated which comprised the time we spent later in testing. We were initially unclear how to format our dataset and web scrape their content's from multiple links until we met with our TA advisor. After this challenge, we began working on the word cloud images which required a short amount of time of about 1 day to fully develop. For the remainder of the first week (2/27 - 3/5) we spent watching video tutorials of different Classifier models so that we could pick the best-fit one to use. This step required 3 days of work as we had to solve how to correctly format the dataset and use the function Tfidf Vectorizer to properly rank terms by their frequency. As we moved to the second week, the progress was much slower due to obligations to other class projects. We drifted further from our original plan as it was more difficult to set up times to meet in person. This led to most of our interactions online where we developed further statistical analysis to answer our third research question. We anticipated to finish well before the deadline but some challenges in our preprocessing and training the model took longer than anticipated. Although we still had enough time to finish the analysis, we sacrificed some quality in testing our model extensively and relating our findings to our research question.

8. Testing

We mainly wanted to test the accuracy of our scraping results and data manipulation precision. For this, we first tested our web scraping result accuracy by testing if the length of each article matches up. We then tested if the categorical labels for CNN and Fox are converted correctly. These results are tested with python's built-in assert statements in "test.py" and can run without producing assertion errors.

9. Collaboration

The main resources we used for this project pertained to building our machine learning model.
We used this video series which can be found here.

https://www.youtube.com/watch?v=H-gE1zLWjQ8&list=PLv3ECF2BkCSNu4914GVqCOxC3N4K9-wtS&index=8
https://www.youtube.com/watch?v=nFna2s244vA&list=PLv3ECF2BkCSNu4914GVqCOxC3N4K9-wtS&index=9

https://www.youtube.com/watch?v=HcKUU5nNmrs&list=PLv3ECF2BkCSNu4914GVqCOxC3N4K9-wtS&index=5

To learn more about the TfidfVectorizer function, we used this site as a resource

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

For the catplot visualizations, we referenced seaborn documentation

https://seaborn.pydata.org/generated/seaborn.catplot.html

To calculate the p-value from the t statistic

https://statscalculator.com/pvaluefromtstatistic?x1=2.2&x4=38&x2=2&x3=1

10. Conclusion

Our goal was to conduct text analysis on Covid related news articles and see if there are differences in CNN and Fox's writing styles that can be found with data processing and machine learning. With our word cloud visualization, bar charts from statistical analysis, and classification model, we noted some interesting findings in diction and developed a model that correctly differentiates Fox and CNN articles with an 84% model accuracy.