Analysis of Word Choices in News Articles

Brandon Wong & Johnny He

Research Background

News media are important sources of information on the Covid-19 pandemic as they report safety guidelines, health data, government actions, and the public opinion

Yet, News media sources may cover different topics or change their writing style based on their political affiliation to cater to their audience.

The goal of our project is to discover if news sources truly differ in their reporting for the purpose of furthering their political affiliations' agenda. The ramifications of this type of reporting is an increased social divide among the public

Dataset: 30+ each of Fox & CNN news articles

Research Question

1. What do Fox and CNN like to bring up in terms of COVID?

2. Are their writing styles distinct enough that we can train a classification model to correctly identify where an article is coming from?

3. Is there a statistically significant difference in frequency of certain word usages between the two news sources?
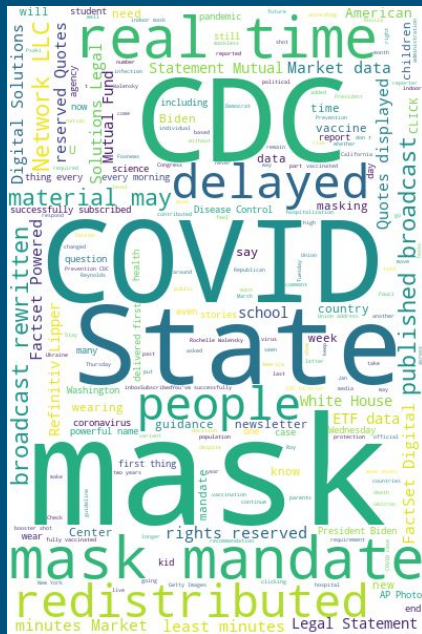
Strategy

Web Scraping

Word Cloud Visualization

TF-IDF Vectorizer and Random Forest for Classification

Statistical Analysis of Word Frequency

Research Result

Fox

CNN

Research Result


```
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()>
```
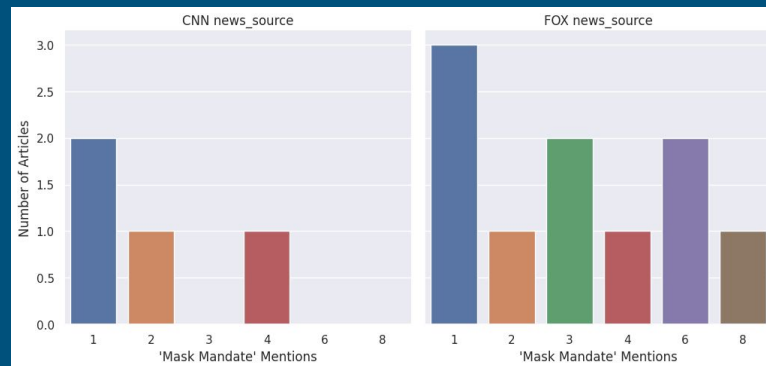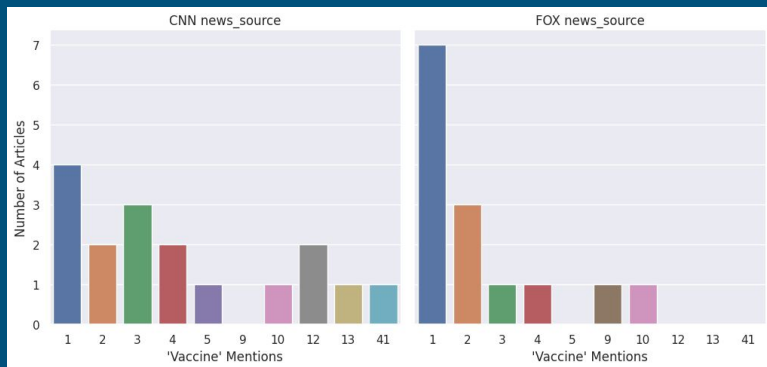
Research Result

Term	T-test	Significance level	Degrees of freedom	P-value
Vaccine	2.2	0.05	38	0.034
Mask Mandate	4.78	0.05	9	0.001

Thank You!