News Bias Analysis

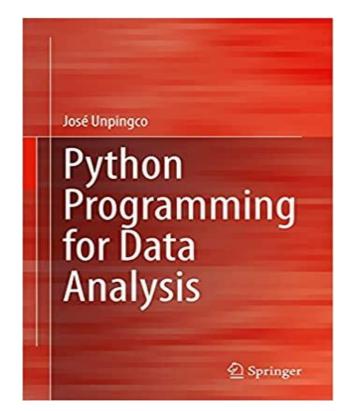
By Alex Paskal, Jeff Wirojwatanakul, Jianyu Tao, Zheng Li

Motivation and Objective

- Some people read only from certain news sources
- Each news source has its own biases
- We hope to reveal such biases for notable news outlets such as CNN and Fox News
 - Generalizable: Don't take the session as Fox vs CNN!
- Goal: To make readers appreciate the biases in the different news articles and to be appropriately skeptical

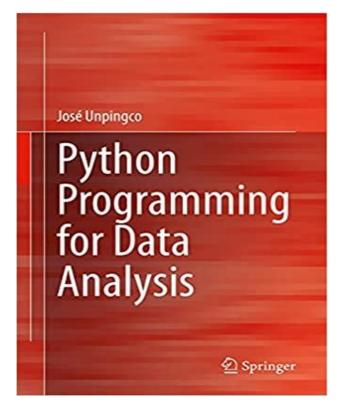
Methodology

- Preprocessing with NLTK:
 - Punctuation and non-alphanumeric characters were stripped
 - Common words known as "stopwords" were removed
 - Text was *lemmatized* to reduce extraneous features
- Sentiment Analysis with TextBlob:
 - Prior to preprocessing, three categories were formed:
 - Full text
 - Sentences
 - Trigrams
 - Each category was preprocessed
 - Subjectivity and Polarity evaluated for a variety of conditions



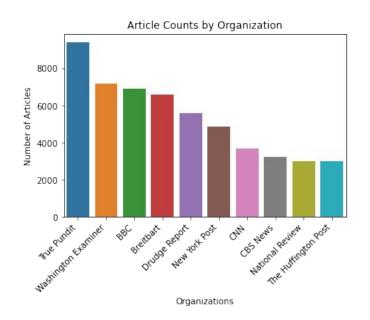
Methodology

- Readability Features Analysis:
 - Linguistic analysis of the difficulty of a text
 - Text evaluated according to several metrics, including TTK, SMOG,
 FKE, average word length
- BONUS! Biden + Covid with BeautifulSoup
 - Scraped and processed html from newsapi
 - Performed high level sentiment analysis



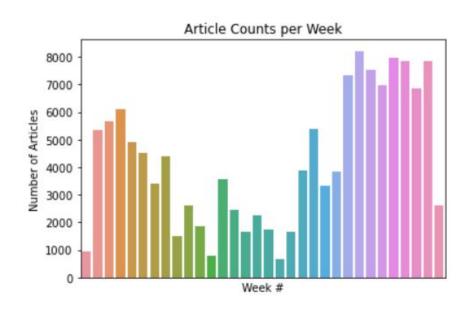
Dataset

Dataset



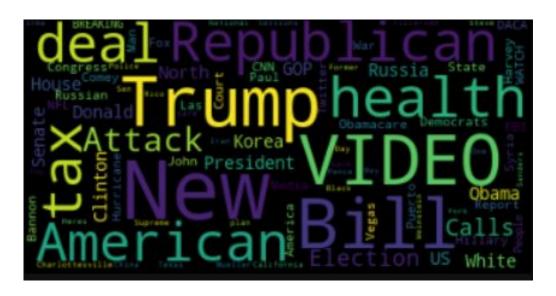
- NELA2017 dataset, presented by Nørregaard et.
 al.
 - Article: https://ojs.aaai.org//index.php/ICWSM/article/view/3261
 - Dataset:
 https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Z
 CXSKG
- Articles were scraped from RSS feeds of various news organizations
- 135559 articles in total
 - JSON files
 - 99 different organizations
 - o 149002 bytes average file size

Articles were not evenly distributed over time

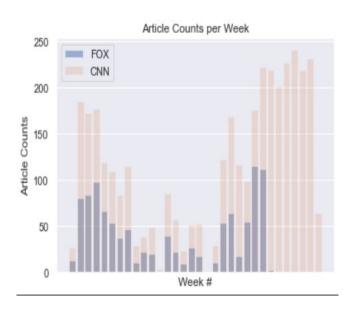


 Articles range from April 1, 2017 to October 31, 2017

Politics dominated the conversation

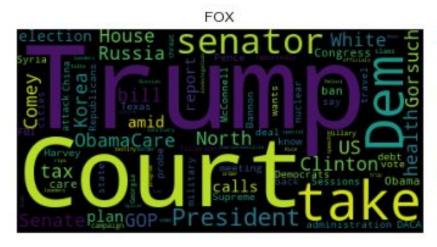


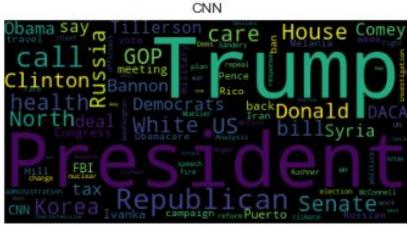
CNN was more heavily represented than FOX



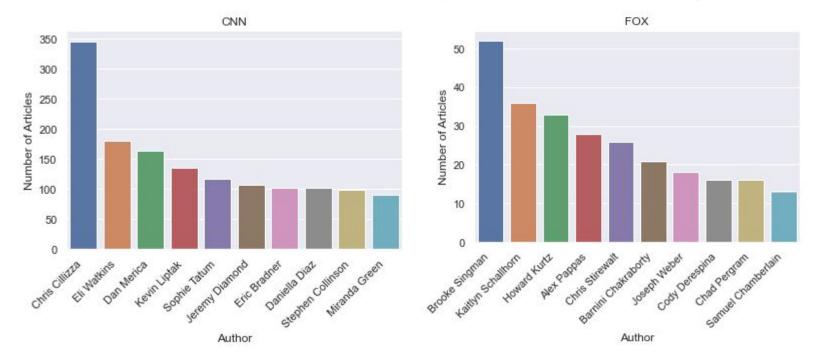
- 1069 FOX NEWS articles
- 3707 CNN articles

Both Fox and CNN kept talking about El Presidente





Chris Cillizza and Brooke Singman were the superstars



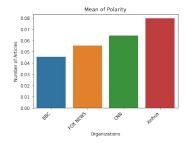
Sentiment Analysis

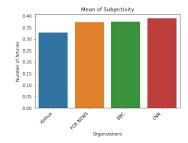
Subjectivity and Polarity of Articles

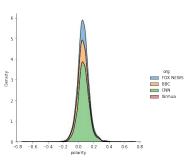
- Subjectivity
 - Degree to which text represents an opinion, rather than factual information
 - Measured from 0 to 1
- Polarity
 - Degree to which text expresses positivity or negativity
 - Measured from -1 to 1
- Text analyzed in three forms:
 - Prose
 - Sentences
 - Trigrams

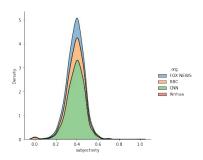
[&]quot;Nela 2017", Benjamin et al.

Text Sentiment - Trump

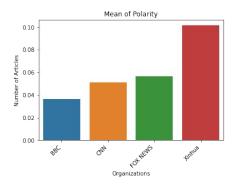


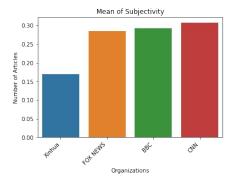


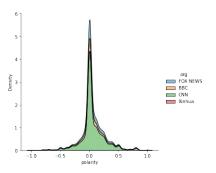


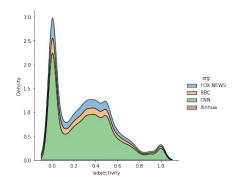


Sentence Sentiment- Trump

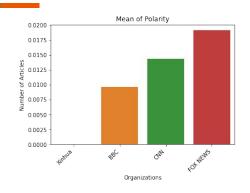


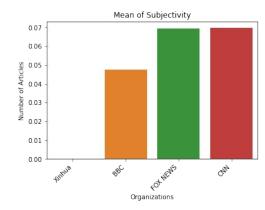


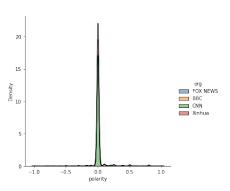


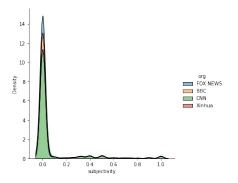


Trigram Sentiment-Trump

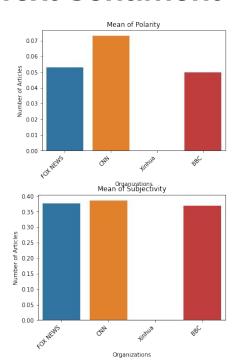


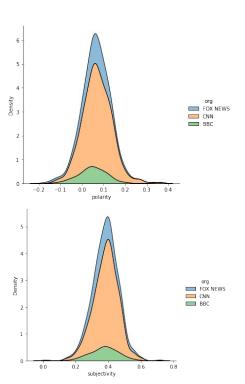




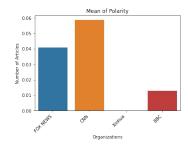


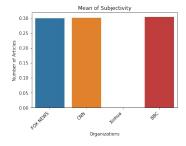
Text Sentiment - Clinton

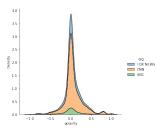


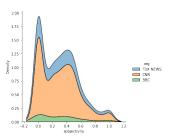


Sentence Sentiment - Clinton

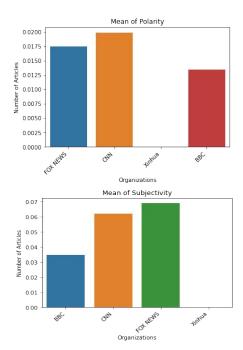


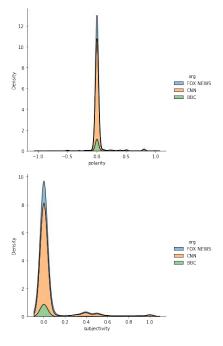




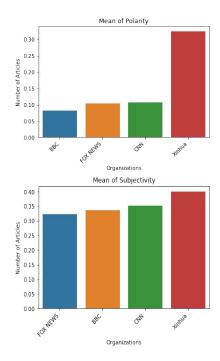


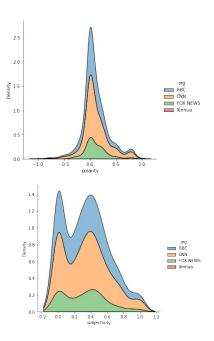
Trigram Sentiment - Clinton



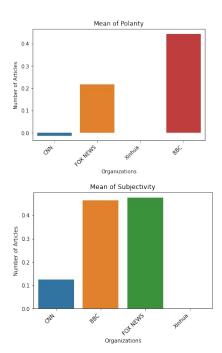


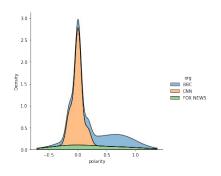
Sentence Sentiment - America

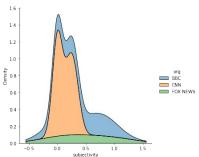




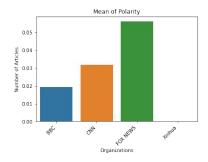
Sentence Sentiment - MAGA

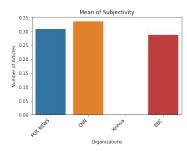


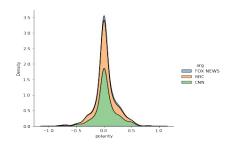


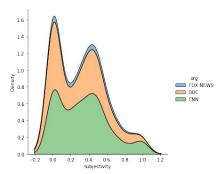


Sentence Sentiment - Gun









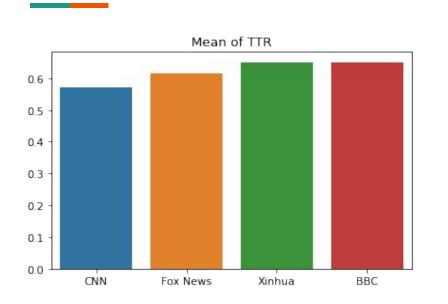
Readability Analysis

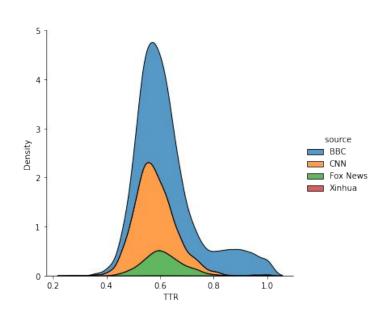
Difficulty and Readability of Articles

- 1. Type -Token Ratio(TTR): Can be a measure for lexical diversity
- 2. Simple Measure of Gobbledygook (SMOG) Readability Measure:
- 3. FKE Readability Measure:
- 4. Average Word Length

[&]quot;Nela 2017", Benjamin et al.

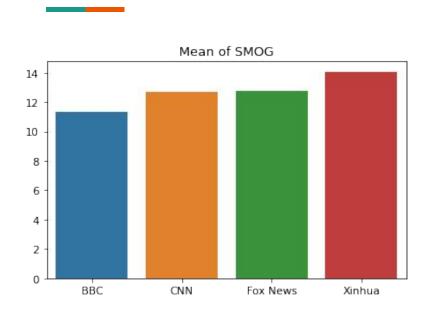
TTR Distribution and Mean Plot:

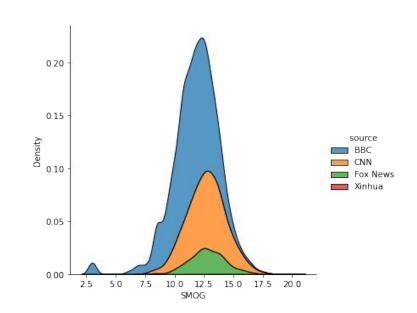




 $\frac{\#uniquewords}{totalwords}$

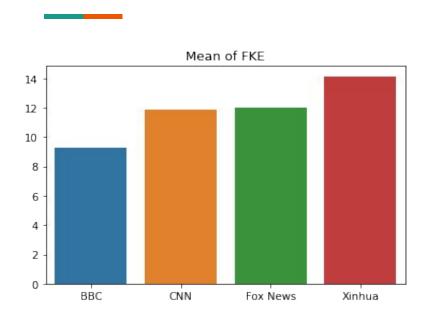
SMOG Distribution and Mean Plot

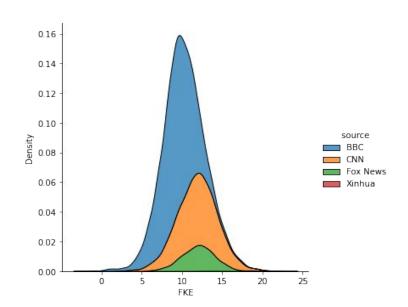




$$1.0430*\sqrt{\#polysyllables*\frac{30}{\#sentences}}+3.1291$$

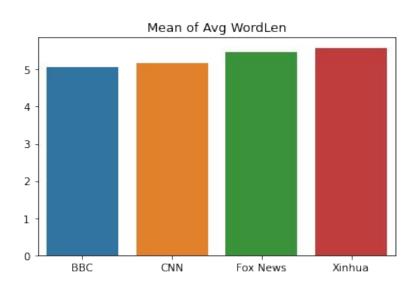
FKE Distribution and Mean Plot

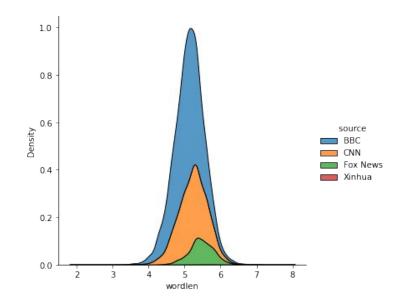




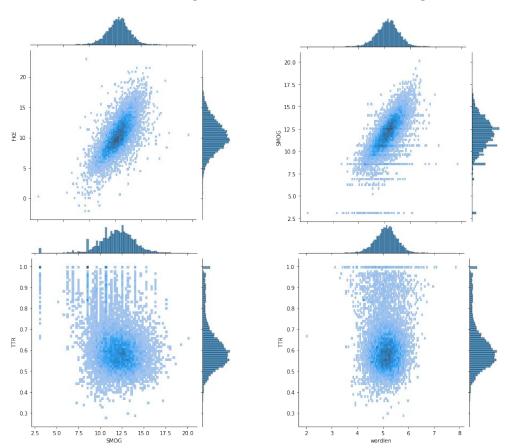
$$0.39*(\frac{Totalwords}{Totalsentences}) + 11.8*(\frac{Totalsyllables}{Totalwords}) - 15.59$$

Avg. Word Len per Article Distribution and Mean Plot





Readability Features Analysis

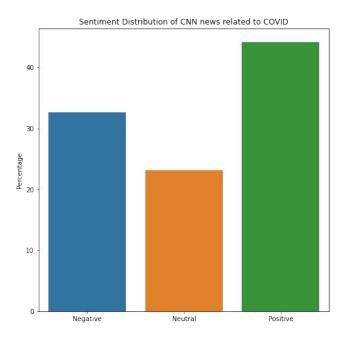


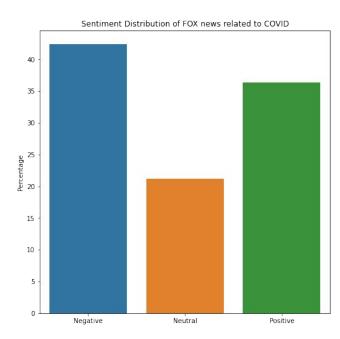
Bonus!

Covid + Biden Articles

- **newsapi**: To obtain urls of Covid and Biden Articles
 - News obtained from October 2021
- BeautifulSoup: Parses HTML urls
- NLTK: To obtain the sentimental polarity score of the text
- **Seaborn**: To plot the figures
- Rake: To extract keyword from raw text

Covid Articles



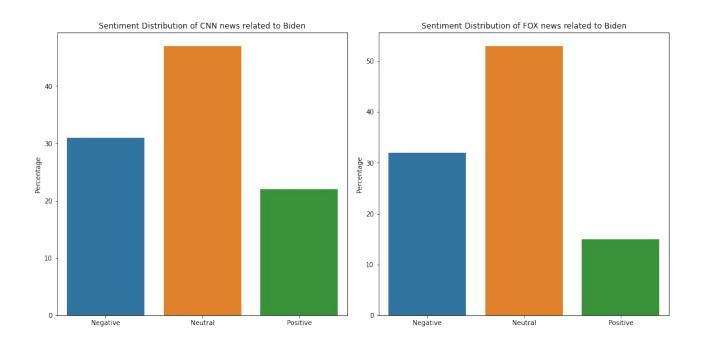


CNN Fox News





Biden Articles



CNN Fox News





THANK YOU