A Study of the Differences between Real and Fake News

Claudia Hidrogo
Madeline Odom
Jordan Sims
Xiomara Winkler

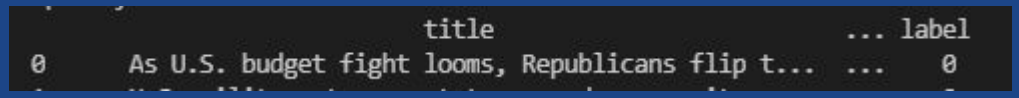
Questions Sought to Answer

1. Is there an association between length of article and the realness of news?
2. What words are most commonly used in real vs fake news?
3. Is there a time when more fake news started showing up (in the Harvard set)?
4. What categories (sports, politics, entertainment) are more prone to fake news?
5. What publications are more prone to publish fake news?

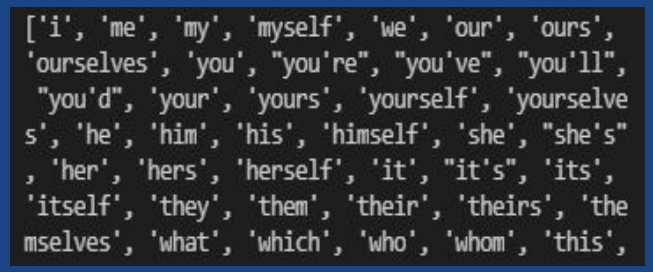
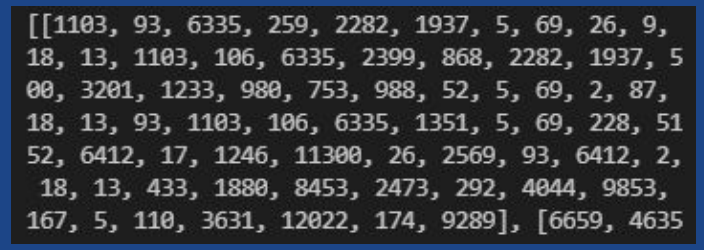
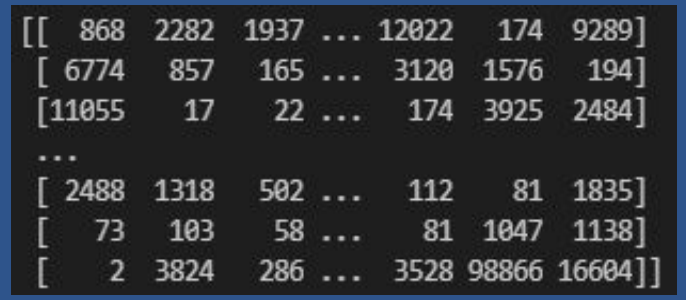
Datasets

Source: Harvard Dataverse	**URL:** https://dataverse.harvard.edu
A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles	
Source: Kaggle: Clément Bisaillon	**URL:** https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset?select=True.csv
Dataset consisting of fake and real news articles.	
Source: Kaggle: Samrat Sinha	**URL:** https://www.kaggle.com/samrat96/fake-news-detection?select=test.csv
Fake News Detection - Train dataset.	
Source: Kaggle: Paul Larmuseau and Deeachain	**URL:** https://www.kaggle.com/c/classifying-the-fake-news/data?select=training.csv
Classifying the Fake News – Training dataset.	
Source: Data Flair	**URL:** https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/
Detecting Fake News with Python and Machine Learning	
Source: Zenodo: Radu Prodan, Prateek Agrawal, Pawan Verma	**URL:** https://zenodo.org/record/4561253#.YWNCkNrMKUl
Generic word embedding dataset for fake news detection consisting of real and fake news from popular news sources.	

Data Pre-processing

Data Cleaning	Prepare and organize the raw datasets before doing the data integration. This step includes: Delete any empty rows Drop NaN values Correct inconsistent data Make sure that the label attribute only has a binary representation 
Data Integration	Concatenate the datasets containing the fake news articles with the real news articles to form a large dataset.

Data Pre-processing

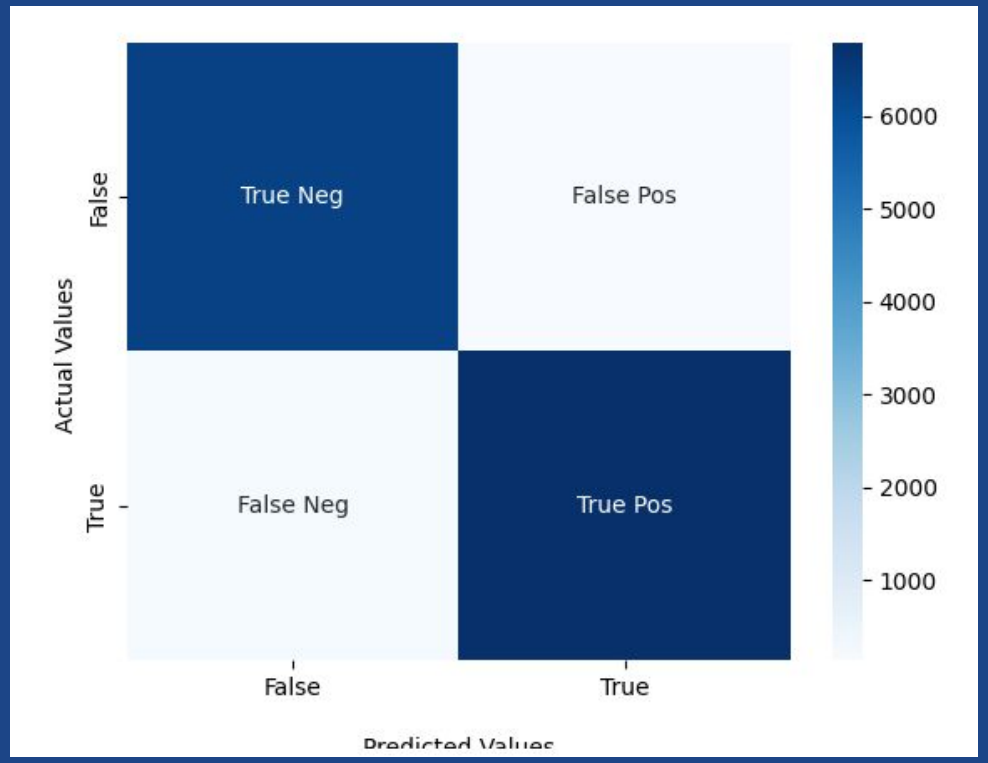
Data Transformation	Stop Words: Identify and remove/drop all the words in each news article that do not have any significant meaning Tokenization: Convert the list of words into tokens (numbers) for easy manipulation of the data.	 
Data Reduction	Use pad sequences to reduce the news articles to an equal length.	

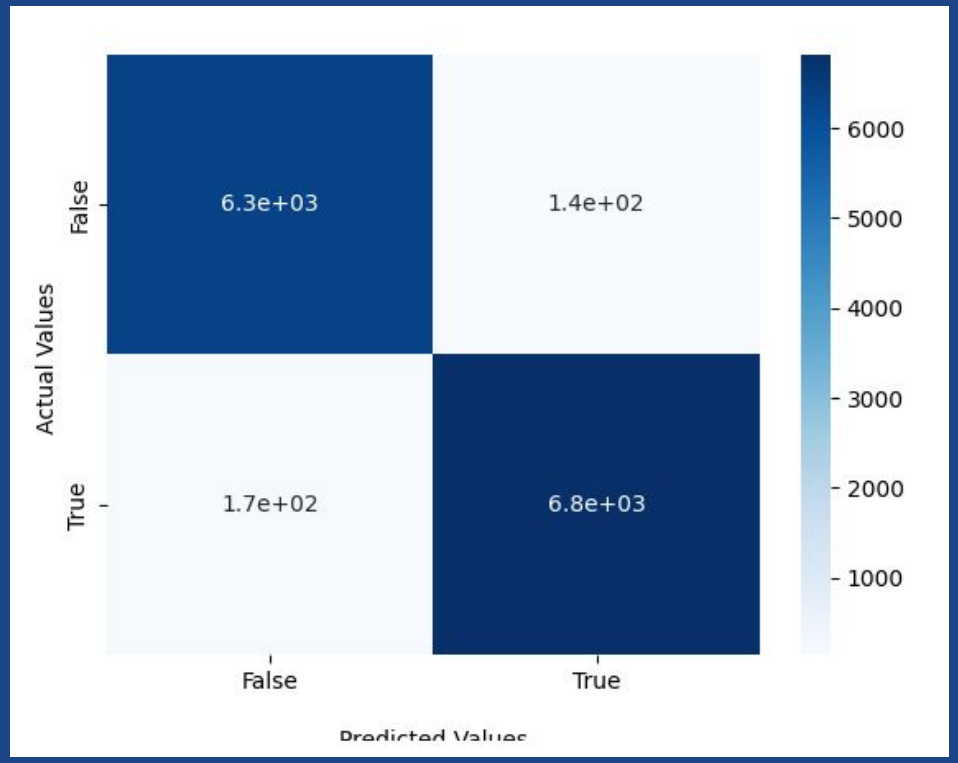
Tools

Pandas	For data frame manipulation
Numpy	For numeric analysis
Matplotlib *Seaborn*	For data visualization
Plotly	To create interactive plots – dynamic visualization
WordCloud	To represent frequency of words
Keras *Tensorflow*	For deep learning
NLTK	For language processing

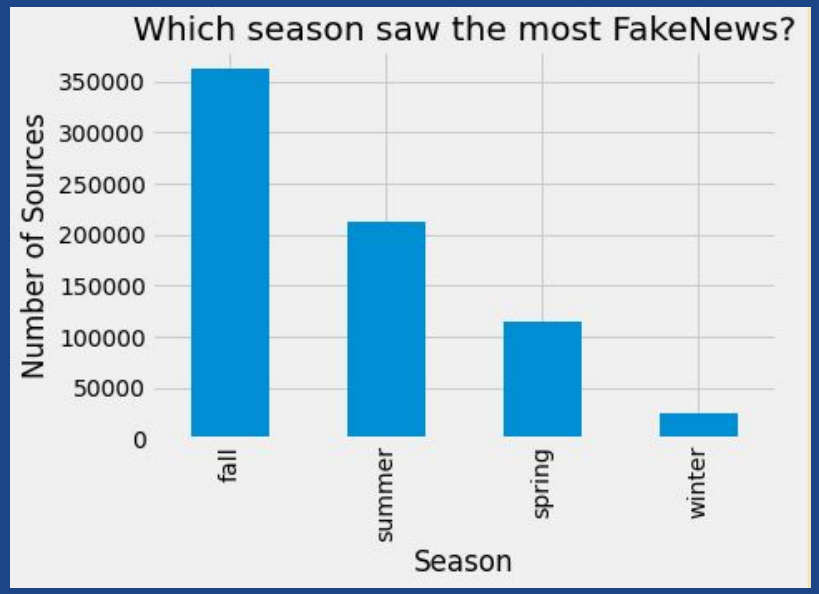
Evaluation

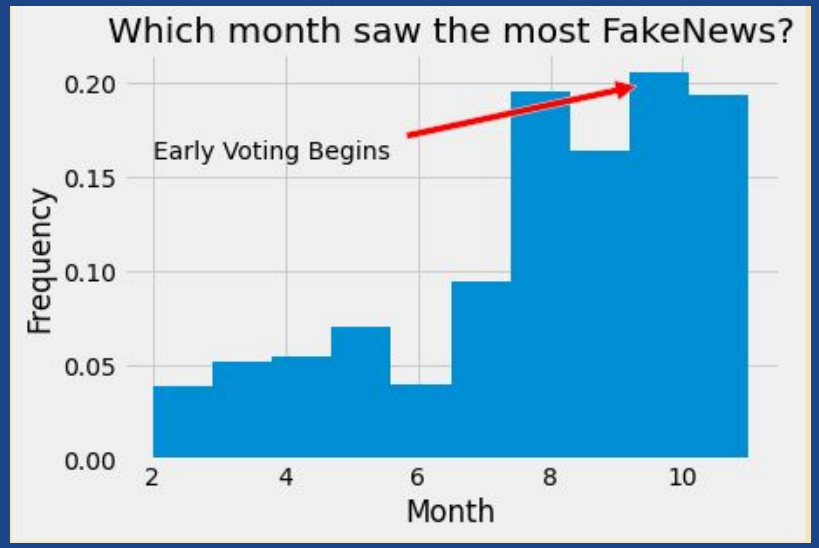
Confusion Matrix





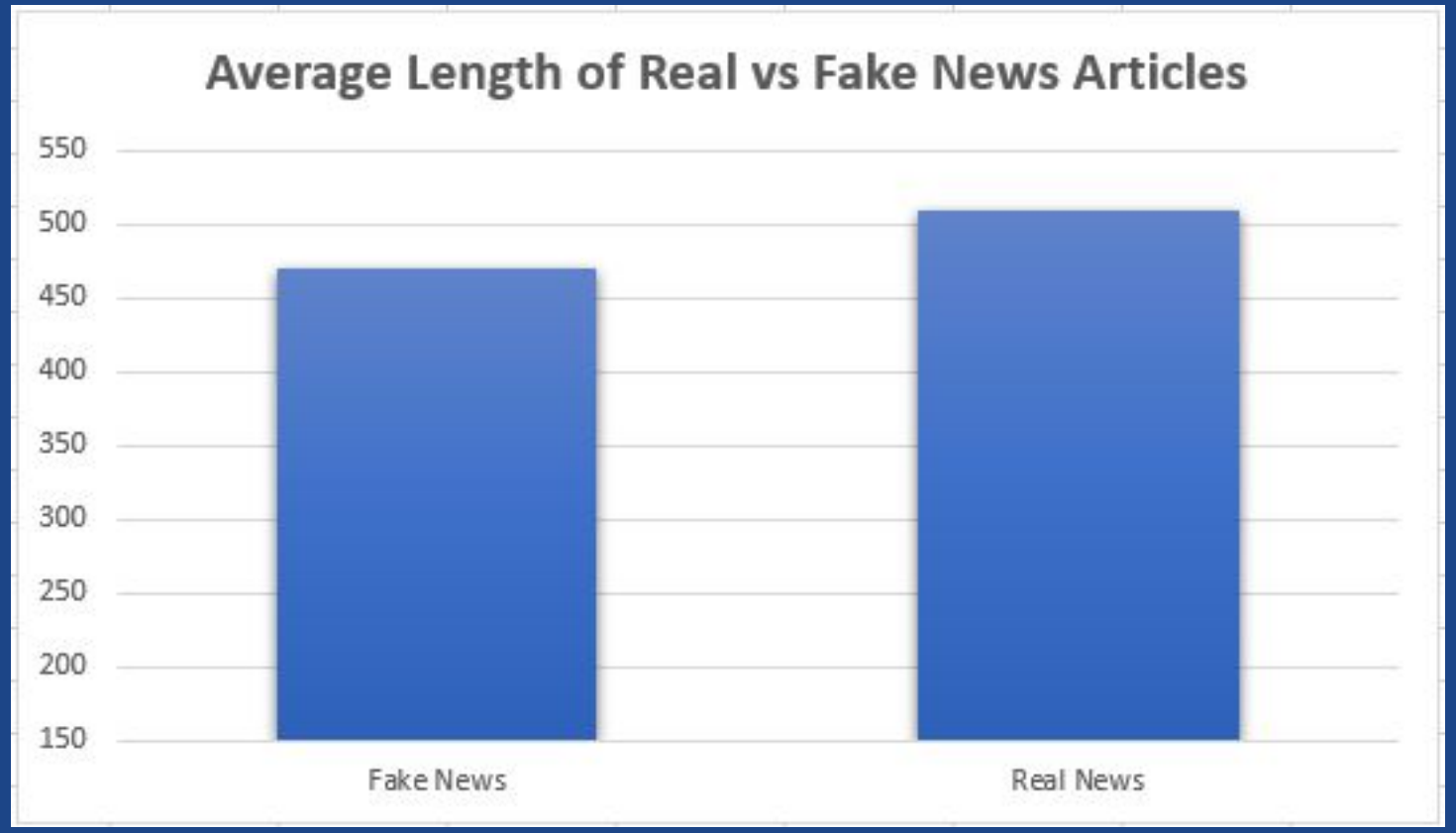
Knowledge Gained





Knowledge Gained

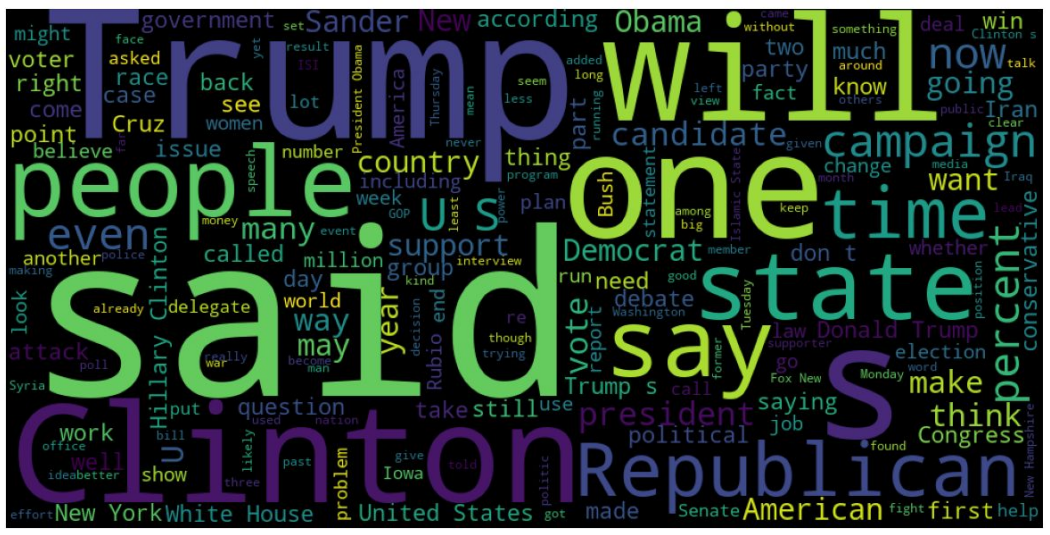
Average Article Length



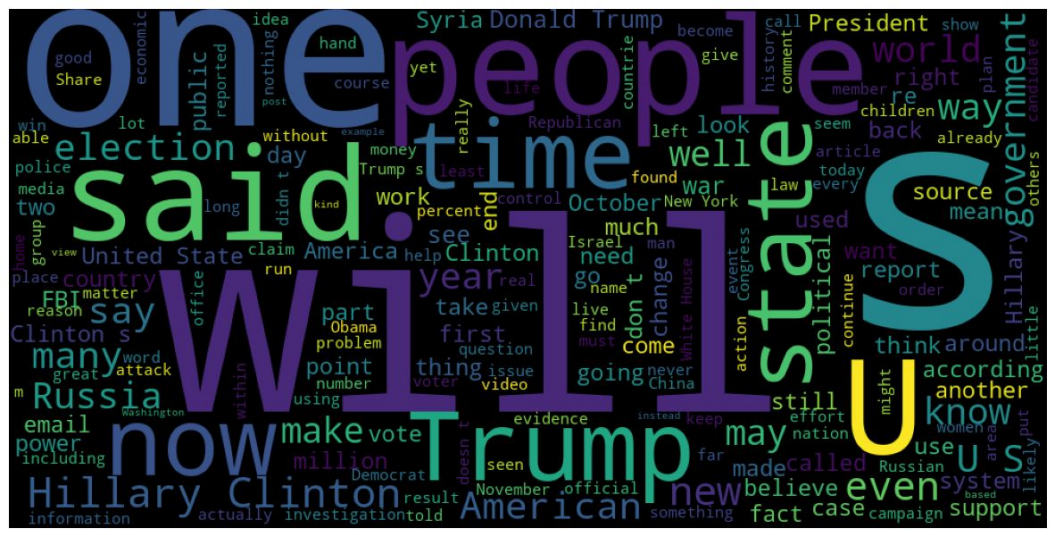
Knowledge Gained

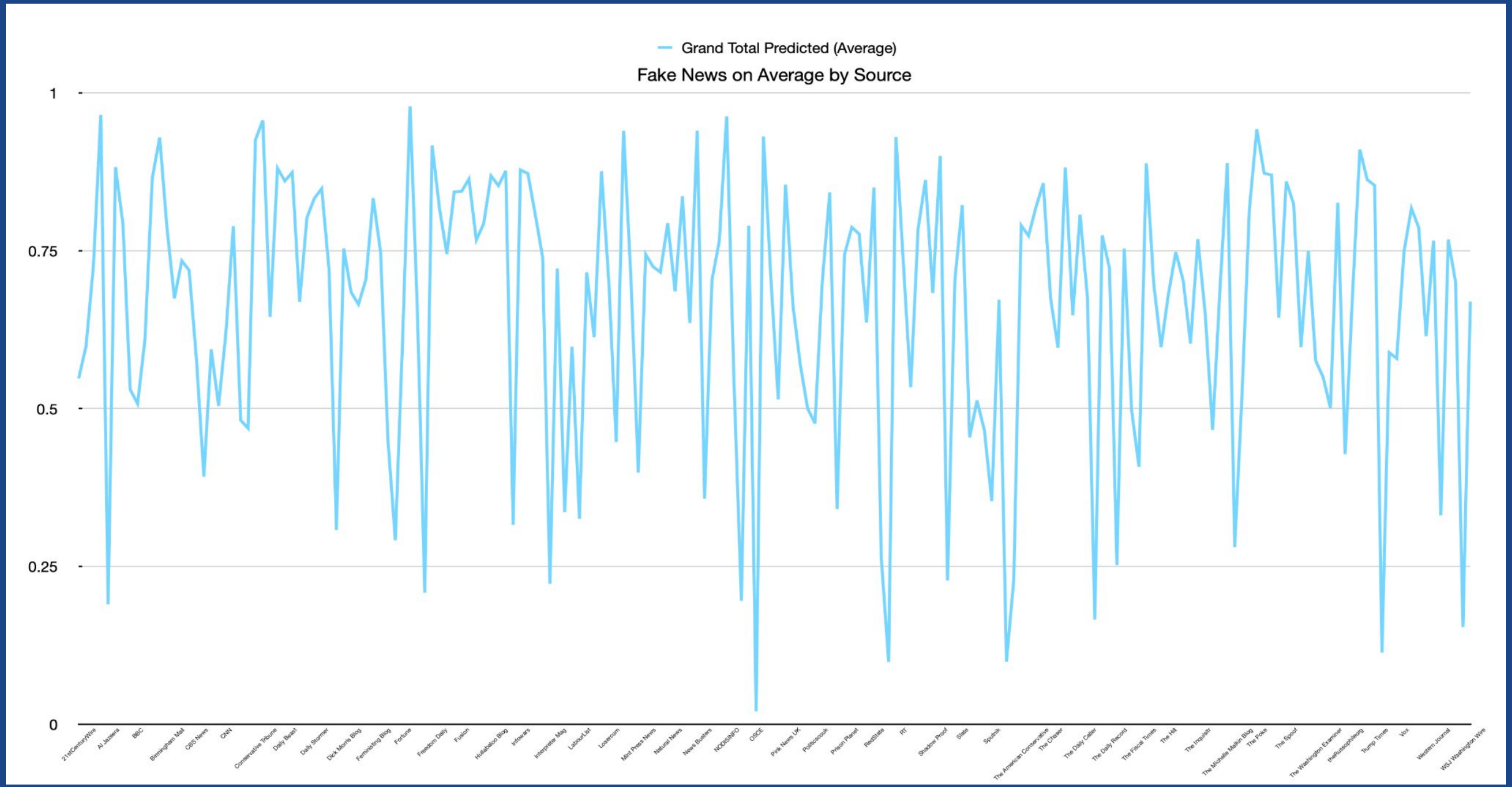
Word Usage Per Article Type

Real



Fake




Grand Total Predicted (Average)
Fake News on Average by Source
1
0.75
0.5
0.25
0
21stCenturyWire
Al Jazeera
BBC
Birmingham Mail
CBS News
CNN
Conservative Tribune
Daily Beast
Daily Stormer
Dick Morris Blog
Feministing Blog
Fortune
Freedom Daily
Fusion
Huffington Blog
Infowars
Interpreter Mag
LabourList
Lowercom
Mint Press News
Natural News
News Busters
NODISINFO
OSCE
Pink News UK
Politicscouk
Prison Planet
RedState
RT
Shadow Proof
Slate
Sputnik
The American Conservative
The Chaser
The Daily Caller
The Daily Record
The Fiscal Times
The Hill
The Inquisitr
The Michelle Malkin Blog
The Poke
The Spoof
The Washington Examiner
thefluxsophilesorg
Trump Times
Vox
Western Journal
WSJ Washington Wire

Applications of Knowledge

Findings

- The average reader will not be able to deduce if an article is fake based off article length alone
- The average reader will not be able to deduce if an article is fake based off word choice alone
- Fake news articles tend to be posted more frequently during election cycles in the United States
- Political news is by far the category most prone to fake news
- It is clear that certain publishers of news generate much more fake news than others, so it is important when getting news to verify the integrity of the publisher before trusting what you read