Truth Detectives

Maria Veronica Sayewich, Andre Lorenzana, Dalia Dahleh, Wasif Parvez

Final Report

'Fake News', as commonly referred to in recent years, has been an issue of rising importance. According to Golbeck et al., it may be defined as factually incorrect information that is presented as a particular medium (usually news stories) designed to mislead the consumer to believe that it is true. Specifically, in March 2018, Monmouth University reported that 52% of Americans are suspicious that what they are reading is considered "fake news", while only 9% do not consider these items to be suspicious [2].

With this issue in mind, the aim of this work is to provide the public with a database that has a variety of articles from various sources and mediums which have been labelled as either satirical, unreliable, or reliable news. We hope that this database will provide greater context for those who wish to decipher the nuances between these forms of media, and can therefore search for truth in this age of misinformation. To reach this end, we will employ the Extraction, Transformation, Load (ETL) process. ETL is a process in which data is extracted from multiple heterogeneous sources, transformed into common & aggregate variables, and then loaded into a final database [3].

Extraction: An outline of the sources of raw data

1. Fake News Kaggle Dataset: A dataset that was designed to support the creation of a system to identify unreliable news sources.

Raw Variables: title, author, text, reliability_label

2. New York Times: API calls for articles published between January 1, 2015 to January 7, 2019. All articles were labelled reliable.
3. Twitter: API calls for tweets from specific twitter accounts and labelled them accordingly.
4. Google News: API calls for articles published by: BBC, The Washington Post, Reuters, The Economist, and The Guardian AU. All articles were labelled reliable.

5. Snopes: Scraped Snopes for unreliable news headlines. All verified as false by Snopes.

Transformation: The mapping of the variables garnered by these heterogeneous sources onto a shared set of variables.

Final Database:

Schema Variables:

Title: The title of the news source

Source: The publishing body

Author: The publishing body or individual author

Text: Body of the news source

Label: [Reliable, Unreliable, Satirical]

Loading: Once the information from all sources were loaded into one data set, we parsed out the items into three CSVs by Label.

Reliable Data Set: reliable.csv

Unreliable Data Set: unreliable.csv

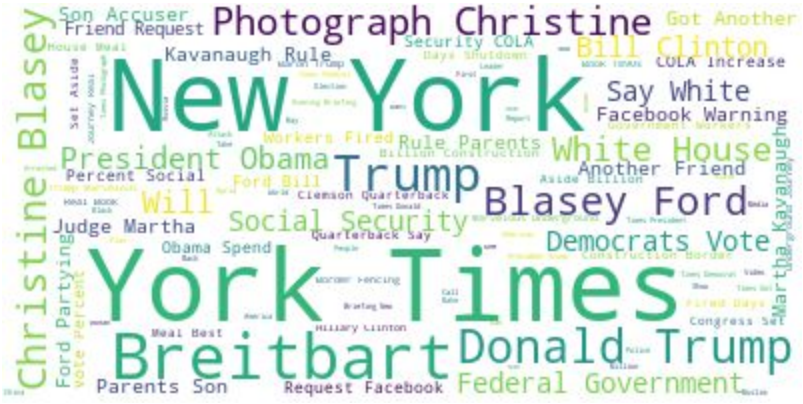
Satirical Data Set: satirical.csv

Using Python, Pandas and MatPlotLib, we transformed these data sets into word clouds and bar plots in an effort to gather a better understanding of the most common words, phrases, or topics in each Label. See the following results:

Word Clouds: Title Analysis

Unreliable Sources:

There are 19,124 unreliable articles in our combined dataset. There are 15,197 titles in the unreliable dataset which provide the following word cloud:



Reliable Sources:

There are 19,856 reliable articles in our data set. There are 17,728 titles in the reliable dataset, and the text from these titles provide the following word cloud:

Satirical Sources:

There are 3,281 satirical articles in our data set. Parsing through this data set proved challenging as we loaded content from the Twitter API, which does not have a governing data type and therefore showed images, hyperlinks, emoticons and other media.

Therefore we have decided to analyze the entire dataset counting the frequency of each words after cleaning the data of the unnecessary characters. The resulting word cloud therefore was created using a function which took in frequencies of words, rather than raw input text.

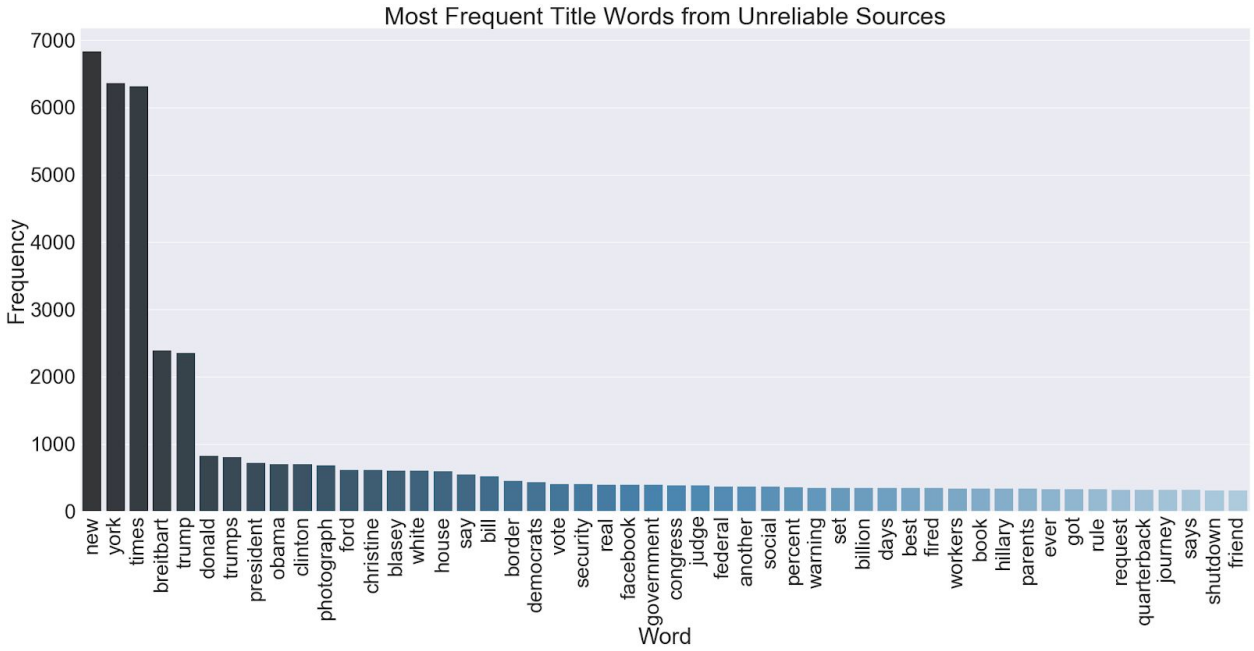
Bar Plots: Title Analysis

Method: The following bar plots were created by counting the occurence of certain strings in both the title and body text of all articles of each type. Due to the difficulty in working with raw heterogeneous text data from a multitude of different sources, certain measures needed to be taken in order to ensure that each word's frequency is counted as accurately as possible and to rid the dataset of unnecessary characters. The following algorithm may be found in Analysis_Tools/wordFrequencyCount.py file.

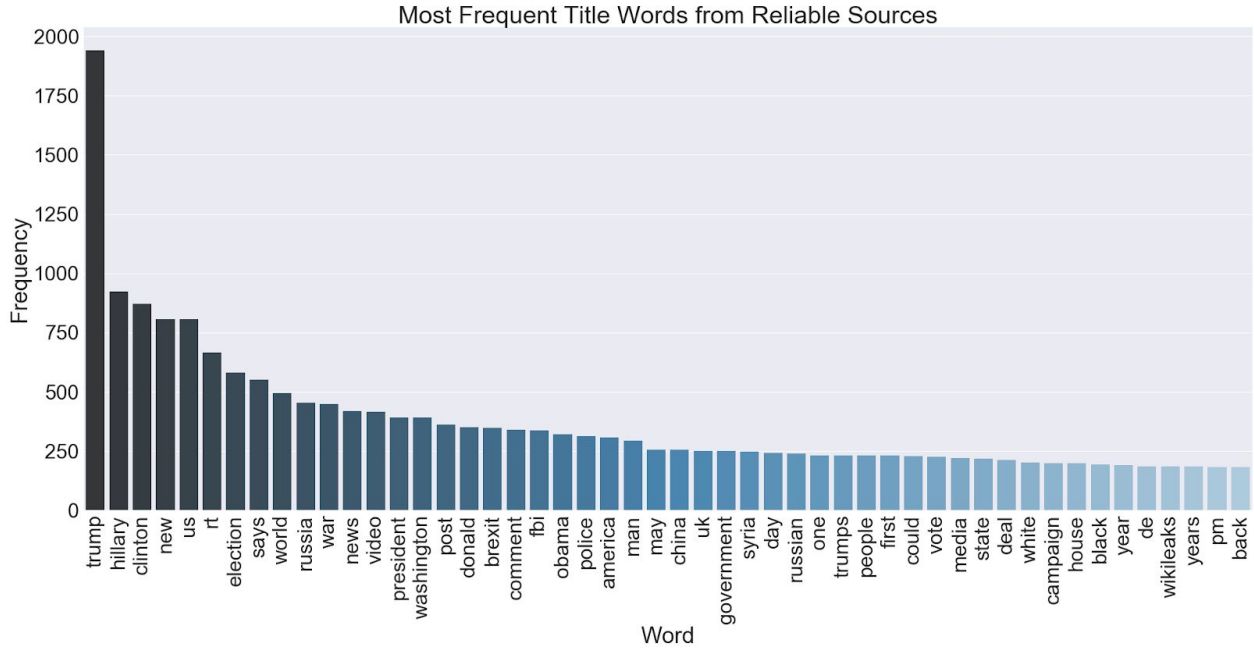
Required Transformations of Raw Input Text:

- Removal of apostrophe "'s" in order to attribute the count of that noun (e.g. Trump's) to the count of the apostrophe-less noun.
- Removal of all quotation marks (" ') which create noise within the data
- Removal of all URLs
- Removal of all non alphanumeric characters (not including underscores '_' and dashes '-').
- Removal of all digits
- Removal of all 'newline characters' (\n) which would result in a different word if it was found following or preceding a certain word.
- Removal of all capital letters, so that the count of each word, no matter how it was capitalized goes towards a single count.

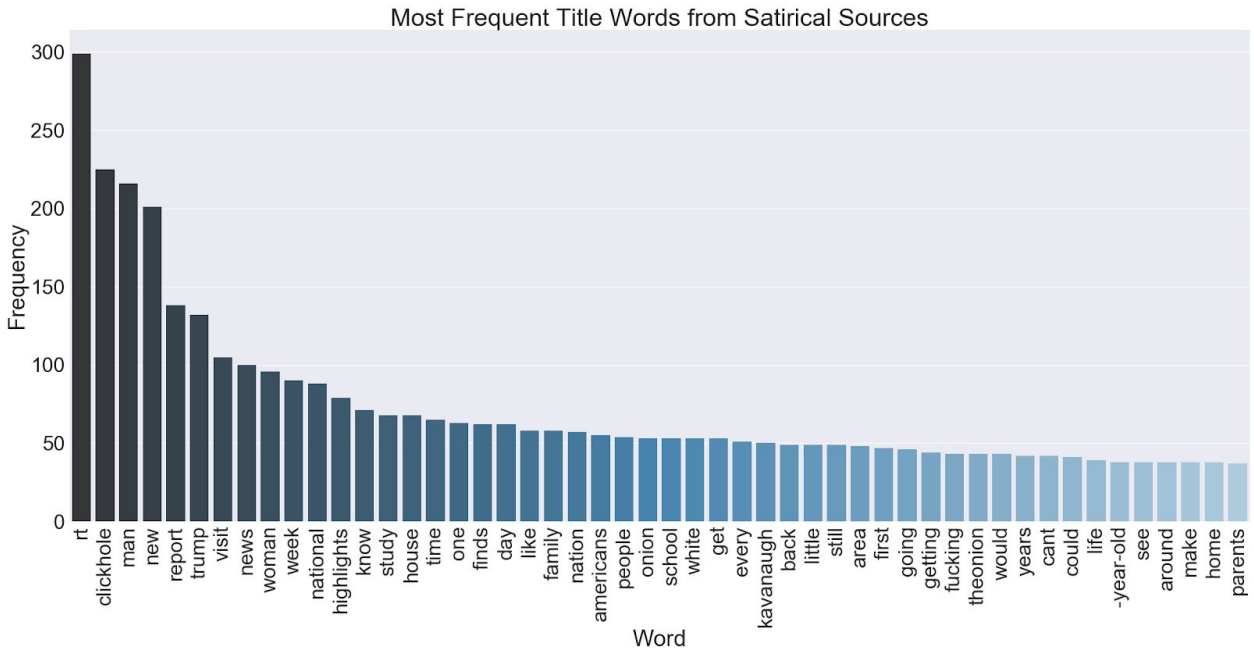
Unreliable Sources:



Reliable Sources:

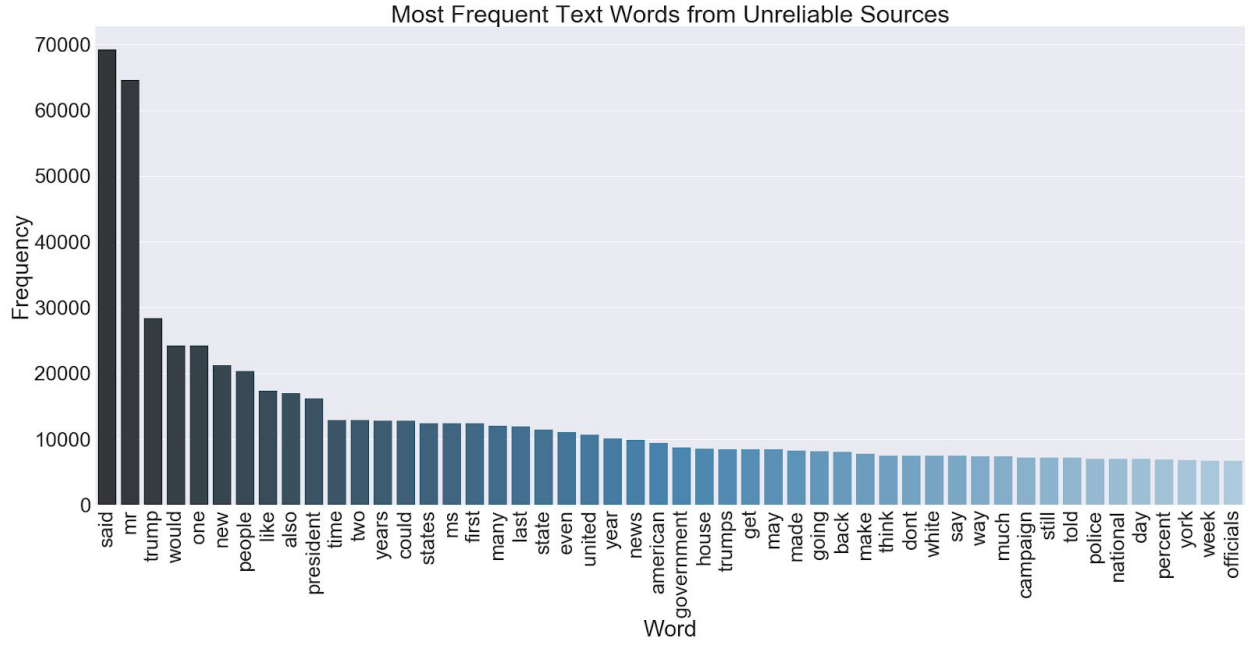


Satirical Sources:

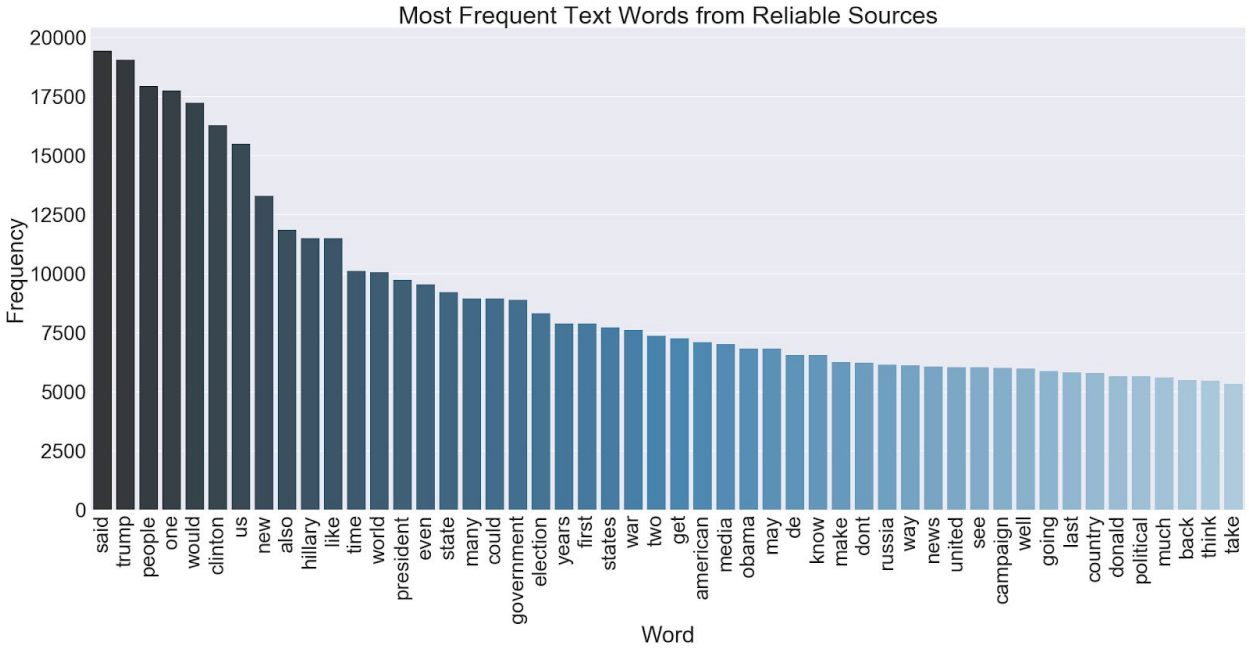


Bar Plots: Body Text Analysis

Unreliable Sources:



Reliable Sources:



Summary Statistics:

There are 19124 **unreliable** articles.

Total Unique Title Words: 18749

Total Counted Title Words: 153657

Total Unique Text Words: 114972

Total Counted Text Words: 5466930

<u>Sample of Unreliable Dataset Titles:</u>

N.F.L. to Spend $100 Million to Address Head Trauma - The New York Times

Chicago Is Planning to Add Hundreds of Police Officers - The New York Times

Ravens vs. Steelers - Friday/Game Status

Did Judge Martha Kavanaugh 'Rule Against' the Parents of Her Son's Accuser

A Quiet Giant of Investing Weighs In on Trump - The New York Times...

There are 19856 **reliable** articles.

Total Unique Title Words: 23359
Total Counted Title Words: 145639

Total Unique Text Words: 202235
Total Counted Text Words: 4419379

Sample of Reliable Dataset Titles:

A Candidate Spurned
Will China Trigger the Next Global Recession?
The Results Are In: This Is the Best Sex Toy for Women, Ever : Conscious Life News
The De Facto US/Al Qaeda Alliance : Information
Obama Furious After Fed-Up 'Deplorables' Drop 41,000-Piece Gift On Him...

There are 3281 **satirical** articles.

Total Unique Title Words: 8302
Total Counted Title Words: 26517

Sample of Satirical Dataset Titles:

Men Are Not Oppressed,' Says Woman Who Has No Idea What It Like To Take Two Whole Escalators To Get To Your Clothi…
LeBron James To Star In 'Space Jam' Sequel
Single Woman Getting All Dolled Up To Watch Room Full Of People Make Out This New Year's Eve
Trump Slams Worldwide Jewish Conspiracy For Not Doing More To Prevent Synagogue Shooting
Earth Passes Through Temporal Vortex Hurling Planet Into Year 2019 …

Analysis:

- Common theme is politics
    - Specifically the political climate in the United States
- Unreliable news outlets *claim* to be affiliated with reliable sources (see “New York” and “York Times” in the unreliable word cloud.
- Unreliable news items tend to be published in capital letters and have a lot of exclamation marks, which may be attributed to their intent to be sensationalist or expressing extreme bias [1].
- The truth of the matter is that there are few organizations dedicated to fact-checking news items as the spread of misinformation surpasses our ability to debunk false claims. Although some organizations exist, there is no overarching governing authority on fact-checking [4].
- With regards to the body text in the analyzed news sources; reliable articles seem to have a more ‘balanced’ spread in their use of words as compared to unreliable ones. This may be evidenced by the final two bar plots, where the top 50 words used in reliable sources have a smaller spread in frequency than those in unreliable sources.
- Unreliable sources make use of the word ‘said’ over 3 times as much as that in reliable articles, given that we analyzed roughly the same amount of each.

Evaluation:

We had originally set out to create a database that can help in public understanding of the nuances between reliable, unreliable, and satirical news sources through providing a variety of different mediums. However, while our efforts have proven capable to provide for the desired end, as evidenced by our brief analysis, much more must be provided in order to achieve a suitable database in which anyone, anywhere, may benefit from in their battle against misinformation.

Jennifer Golbeck et al. outlines 9 necessary criteria for a successful “corpus” to better understand the issue of fake news & misinformation [1]:

1. Availability of both truthful and deceptive instances
2. Digital textual format accessibility
3. Verifiability of ground truth

4. Homogeneity in lengths
5. Homogeneity in writing matter
6. Predefined timeframe
7. The manner of news delivery
8. Language and culture
9. Pragmatic concerns

Using these 9 criteria, Golbeck and 27 other Stanford researchers were able to construct such a corpus of: **204 satirical articles and 284 unreliable articles** [1]. This highlights the difficulty of the undertaking that the Truth Detectives have chosen.

With our ambitious aim, we have accumulated 42,261 news articles of different types, and performed aggregate analysis on the texts. However, we hope that the work we have provided here may be used as a stepping stone to the difficulties of compiling a useful dataset for the purpose of combating misinformation in our world today.

References

[1] Golbeck, Jennifer, et al., "Fake News vs Satire: A Dataset and Analysis". WebSci '18, (May 2018), *Stanford University,* http://web.stanford.edu/~mattm401/docs/2018-Golbeck-WebSci-FakeNewsVsSatire.pdf

[2] Perceived frequency of online news websites reporting fake news stories in the United States as of March 2018, Statistica.com, (2018), *Monmouth University,* https://www.statista.com/statistics/649234/fake-news-exposure-usa/

[3] Vassiliadis, Panos. "A Survey of Extract-Transform-Load Technology", Researchgate.net, (July 2009), *International Journal of Data Warehousing and Mining*, https://www.researchgate.net/publication/220613761_A_Survey_of_Extract-Transform-Load_Technology

[4] Ramos, Dulce. "A new home for the IFCN Code of Principles", Poynter.org, (July 2018), *International Fact-Checking Network (IFCN)* https://www.poynter.org/fact-checking/2018/a-new-home-for-the-ifcn-code-of-principles/