Truth Detectives

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

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

Source:

Author:

Text:

Label: [Reliable, Unreliable, Satirical]

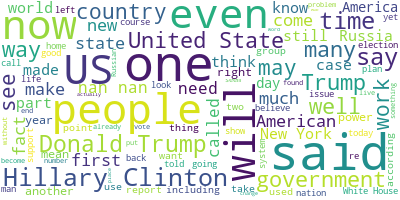
**Loading:** Once the information from all sources were loaded into one data set, we parsed out the items into three CSVs by Label.

**Reliable Data Set:**

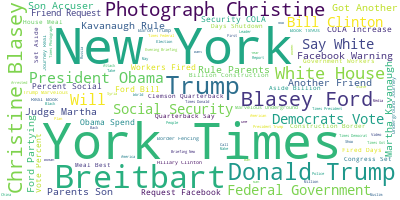
**Unreliable Data Set:**

**Satirical Data Set:**

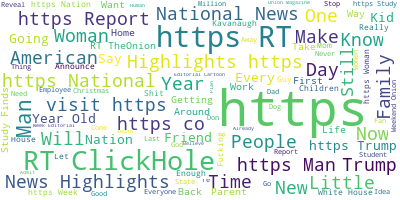
Using Python, Pandas and MatPlotLib, we transformed these data sets into word clouds and histograms in an effort to gather a better understanding of the most common words, phrases, or topics in each Label. See the following results:



*Reliable*

**

*Unreliable*

**

*Satirical – consider deleting!*

**Analysis**

* Common theme is politics
  + Specifically the political climate in the United States
* False news outlets *claim* to be affiliated with reliable sources (see “New York” and “York Times” in the unreliable word cloud
* Fake news items tend to be published in capital letters and have a lot of exclamation marks
* 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].

**Bibliography:**

[1] Stanford Fake News <http://web.stanford.edu/~mattm401/docs/2018-Golbeck-WebSci-FakeNewsVsSatire.pdf>

[2] Fake News, Monmouth University <https://www.statista.com/statistics/649234/fake-news-exposure-usa/>

[3] A Survey of ETL <https://www.researchgate.net/publication/220613761_A_Survey_of_Extract-Transform-Load_Technology>

[4] International Fact-Checking Network (IFCN) <https://www.poynter.org/fact-checking/2018/a-new-home-for-the-ifcn-code-of-principles/>