Analysis of Real and Fake News Articles

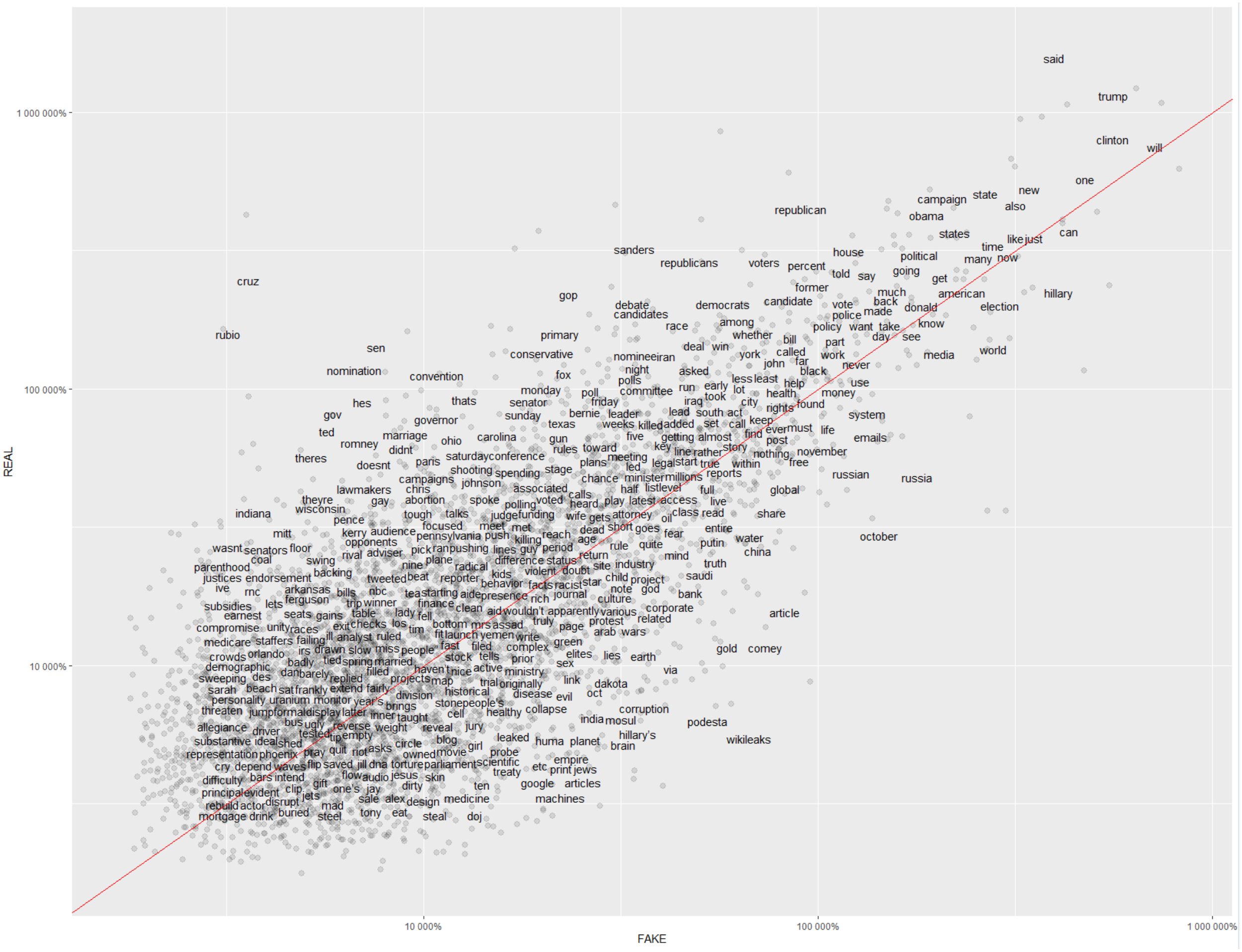
November 29th, 2023

Report Presented by:

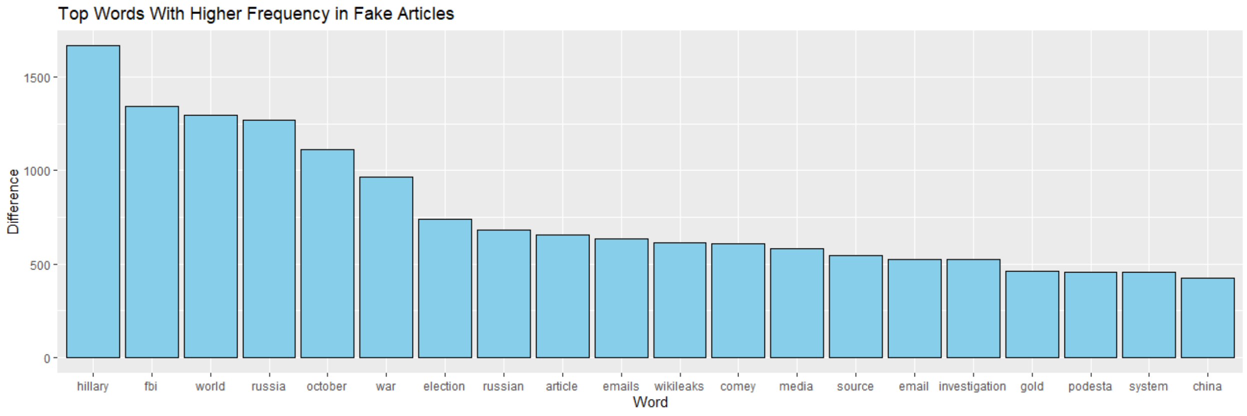
Carson Edwards

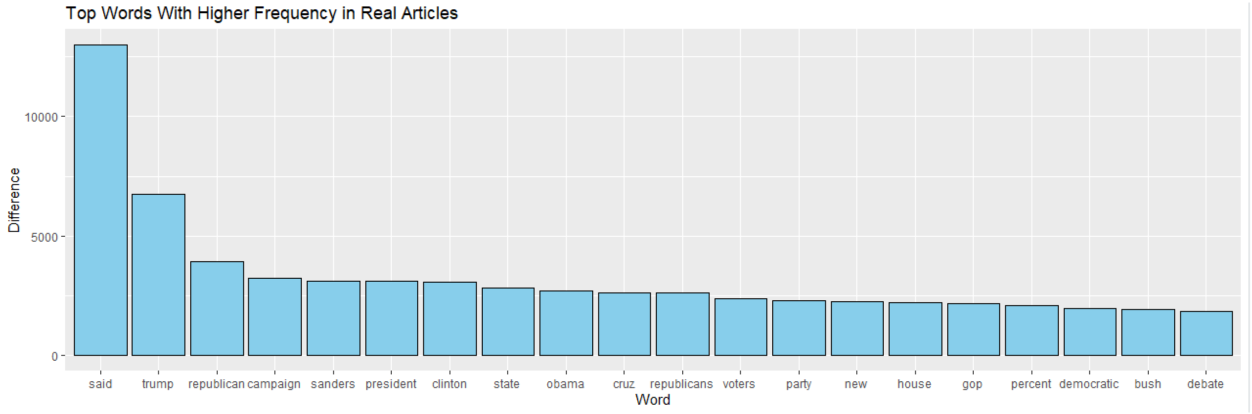
In today's day and age, people can access any kind of information more easily than ever before. An important skill that an aspiring informed citizen must have is the ability to verify something they see on the internet. Following recent elections, the term "fake news" has been used often to describe articles containing false information. Unfortunately, false political information spreads significantly faster online than true political information (*Wayback Machine*). This report will analyze a Kaggle dataset which has article titles and text already labeled as fake or real (*Fake News Prediction Dataset*). The data contains 3164 fake articles and 3171 real articles. Through this analysis, we will get an understanding of what we can use to tell apart fake and real news articles by looking at their differences in language and content.

Below is a chart that can give us an idea of the word frequencies in real versus fake article text. Words closer to the center represent equal frequency in both categories, words below the line have a higher frequency in fake articles, and words above the line have a higher frequency in real articles. The words included in this chart all appear in the text over 30 times. This decision was made to try to isolate common themes across many documents. Some interesting terms that show up towards the fake side include China, truth, lies, Russia, elites, evil, sex, Arab, fear and Hillary. There seems to be a recurring theme here focusing on political scandals and speculation, with several terms appearing to reference the Hillary Clinton email controversy. For real articles, high frequency terms include Sanders, said, Trump, Obama, republicans, democrats, and gun. It is interesting that "Hillary" has a higher frequency in fake articles compared to "Trump" or "Obama".



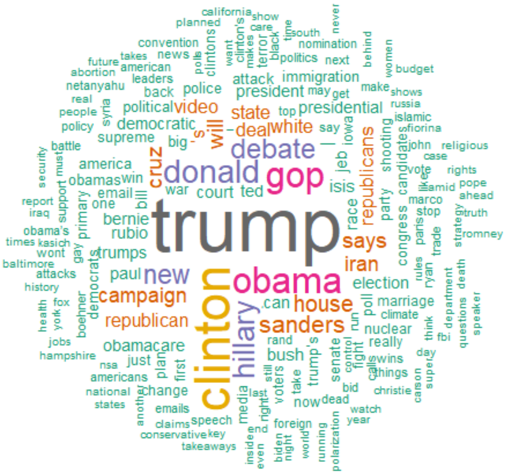
While this chart above gives a general idea of some different words in both categories, the charts below show the top words with the highest difference in frequency for real versus fake articles. This will clearly show the most identifying terms for real and fake article text.



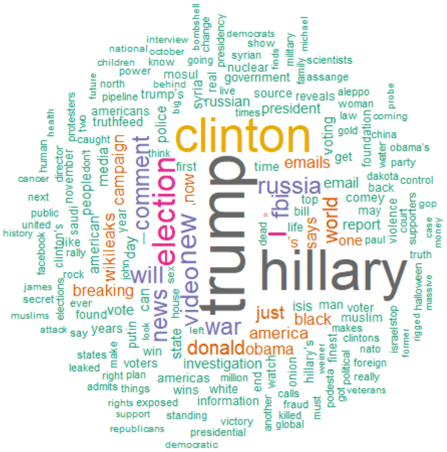


One thing to note when reading the charts above is that the scale is not adjusted for either, the count of these words is much higher for the real articles. The highest occurring word for real articles, “said”, occurs over 12,500 times more often than in fake articles. However, for fake articles, “Hillary” occurs over 1,500 times more often than in real articles. One takeaway from this is that real articles are much more likely to quote someone directly, using the word “said” in the text. This can be a key identifying feature as it is likely that fake articles will not lie about a quote someone did or did not say. Real articles are also more likely to reference real people like Donald Trump, Bernie Sanders, Barack Obama, or Ted Cruz. Another takeaway is that if an article references “Hillary” without “Clinton” then it is more likely to be labeled as fake news in this dataset.

So far in this analysis, the charts have been focused on the content inside each article. However, articles online are often shared before someone even reads the entire article. People who don’t share the article may only read the title and get their news from a general set of headlines. So we will now take a look at the titles of the articles. Below are word clouds made of the most common words in real and fake article titles from our data set.



Real Article Titles



Fake Article Titles

One new thing that this shows us is that most articles from our dataset are about Donald Trump whether they are real or fake. It also tells us that Hillary Clinton is a topic found often in both real and fake article titles, and confirms that her first name is mentioned more often in fake articles. For fake article titles, terms like Russia, FBI, war, wikileaks, and election are some of the most common terms found that are not as common in real articles. These terms can be categorized as key identifiers for judging an article on its title alone. Real article titles are more likely to have terms such as Obama, GOP, debate, campaign, house, and republican. One interesting thing to note is that the terms "white" and "house" are more common for real articles, most likely paired together. When looking at the data, 119 titles contain "white" and 74 of those contain "white house" which leaves 45 instances that are not about the White House. In addition to that, the term "black" is more often found in fake article titles, with 70 instances of the term in fake titles and 28 instances in real titles. Many of these fake article titles reference Black Lives Matter and black voters. Another interesting language difference is the presence of the word “now” in article titles. Fake news article titles contain “now” more often in the data set, with 110 instances compared to the 59 instances in real article titles. The term is most likely used as a clickbait tactic signaling urgency to the potential reader.

Becoming familiar with these patterns would help improve one’s internet literacy. Fake articles are easily identifiable by their speculative title content such as Russia, FBI, war, wikileaks, and Hillary Clinton emails. A sense of urgency can be found in these titles as well with the term “now” more frequently appearing in fake article titles. Their text content is typically even more telling with terms such as China, truth, lies, elites, evil, sex, Arab, and fear more frequently appearing. Real articles are more likely to directly quote people with the term “said” showing up much more often than in fake article text. The real article titles typically contain more common political terms such as republican, debate, campaign, presidential, and democratic. In conclusion, there is no one rule that can be used to identify whether an article is real or fake. However, having an understanding of the common language and context of each type will allow readers to have a better initial screening of articles they come across.

References

Fake News Prediction Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/rajatkumar30/fake-news/data

Robinson, J. S. and D. (n.d.). 7 Case study: comparing Twitter archives | Text Mining with R. In www.tidytextmining.com. Retrieved November 29, 2023, from https://www.tidytextmining.com/twitter.html

Rul, C. V. den. (2019, October 20). *How to Generate Word Clouds in R*. Medium. https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a

Wayback Machine. (n.d.). Web.archive.org. Retrieved November 29, 2023, from https://web.archive.org/web/20190429073158/http://vermontcomplexsystems.org/share/papershredder/vosoughi2018a.pdf