Fake News Analysis

Clayton Armbrust
Marquette University
Milwaukee, WI 53233, USA
clayton.armbrust@marquette.edu

Abstract
This Data Science Project analyzes Fake News published near the 2016 Presidential Election. The analysis formulates findings around how Fake News articles affected the election and how each political party benefitted from Fake News. This was achieved by various textual analysis techniques that identified common words in Fake News articles. This revealed

just how large the issue of Fake News really is. Following the finding of this report, it is recommended that the general public start to demand higher levels of reporting and journalism that we see on Google and Facebook.

Author Keywords
Fake News, Latent Dirichlet Allocation, Hierarchical Dirichlet Process

Introduction
Fake News has been a very popular topic since the 2016 Presidential Election. The effects of Fake News and how it has spread have been debated and analyzed over the last few months. Facebook and Google have received public backlash for their lack of integrity protection in the articles that spread through their services. Since this dilemma escalated quickly, there have not been large and successful research projects in this field. By attempting to begin this understanding of

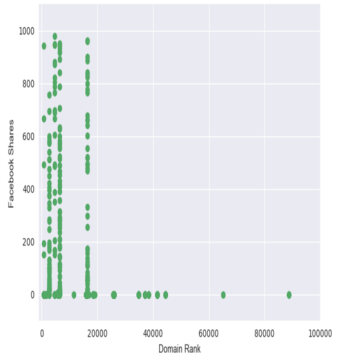

Figure 1: Comparing Facebook Popularity to Website Domain Ranking

Fake News, this project examines some of the viral Fake News articles that propagated through Facebook. The articles of interest in this research consisted of ones that mentioned Hillary Clinton or Donald Trump. Establishing this baseline allowed for direct analysis of the contents in those articles.

Exploring the Data

The data used for the project was collected from Kaggle project titled "Getting Real About Fake News" by Megan Risdal. She compiled her dataset by using a Google Chrome extension called "BS Detector" in unison with a web scrapping service that scanned hundreds of thousands of websites. From there, any articles that BS Detector labeled as Fake News were added to a csv file with various contents. So BS Detector, which is a public and free product, determined whether an article was considered fake. The tool also attempted to classify the articles into different categories such as Junksci, Satire, BS, and a few others. So the project heavily relies on BS Detector's accuracy of identifying fake news in order to effectively analyze the contents of Fake News.

The characteristics of the data ranged from domain ranking and Facebook sharing statistics to the text of the article. Other characteristics included information about the author, Twitter sharing statistics, and the country the article was published in. Additionally, it is important to note that the articles were gathered around the 2016 Presidential Election.

In their classification of Fake News, BS Detector included articles that they could not classify to a specific category. These articles were removed from the dataset in order to preserve the integrity of the dataset and to ensure that sound analysis can be done on all of the article's contents. Further exploration of the Social Media characteristics, the article's spam score, and the categories assigned to articles gave more insight into the data. The spam score attribute was computed by the BS Detector tool and it did not have any specific API's on the details of the calculation. Due to this lack of information, that attribute was ignored. The dataset also included Twitter sharing, but those numbers were minimal in comparison to those of Facebook. So after exploring more of these attributes, it was evident that the most important ones were Facebook data and the text of the articles. These two categories allow for proper analysis of the articles content while taking the article popularity into account.

Figure 1 illustrates how Facebook Sharing and Domain Ranking have a positive correlation, which shows how Facebook has a profound impact on Internet traffic. This allows us to assume that anything with moderate Facebook sharing should be taken as an influential article. For the existing dataset, the average Facebook Shares per article was found and anything below that threshold of 46 was thrown out of the dataset. This allows for investigating of only articles that had at least a moderate effect on Facebook users and potential voters. From this point forward, textual analysis of the articles was the only focus as it gave insight into whether Fake News impacted the 2016 Presidential Election.

Textual Analysis

This section of the project focused on analyzing the text inside the articles that were left in the dataset. Before examining the contents, it made sense to break up the articles into two groups. The first sub dataset consisted

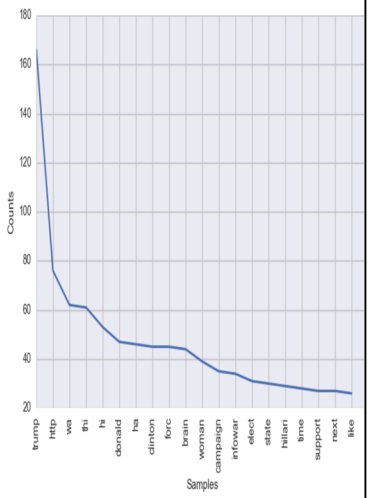

Figure 2: Trump's Frequency Chart

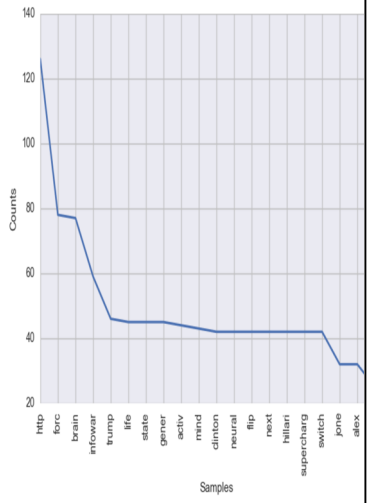

Figure 3: Hillary's Frequency Chart

of articles that mentioned Hillary Clinton and the second sub dataset consisted of articles that mentioned Donald Trump. This was done in order to separate the two political parties that could have benefitted from Fake News. Analyzing the contents of the articles in the subgroups and comparing the results will show who truly benefitted from Fake News. Additionally, having smaller subsets of articles will allow the LDA and HDP textual analysis models to run more efficiently. These small group sizes, of 19 articles for Trump and 14 for Hillary, could affect the relevance of using these algorithms.

Constructing a frequency analysis chart of the most used words in each cohort does not provide a lot of useful information. From the constructed charts, the only useful information gathered was that the article text in the dataset includes erroneous values such as website URLs and other nonsensical terms. These words on a frequency chart can be shown in Figure 2 and Figure, which display Donald Trump's and Hillary Clinton's frequency charts respectively. This severely hurt the results from the Frequency Distribution graphs that utilized String Tokenization methods. The first Tokenization method deleted unnecessary words that the NLTK library denoted as "stop words." These words refer to the most common words in the English language that do not offer any value in terms of textual analysis. The second method deleted any punctuation marks from the article content. The third method utilized stemming techniques that shortened words to their base.

Even though the frequency analysis charts did not provide any useful information, the methods that led to them became useful when conducting LDA and HDP methods on the articles in each subset. These methods such as tokenizing, lemmatizing, and stemming the data allowed the LDA and HDP algorithms to more accurately analyze the data.

LDA and HDP Findings

The resulting dataset for Donald Trump had 19 articles and that of Hillary Clinton had 14. The original dataset consisted of nearly 13,000 articles, so the previously mentioned methods of parsing the data significantly decreased the amount of text that could be analyzed. For the LDA and HDP algorithms, we predefined the amount of topics to return as ten with five words in each topic. Both of these methods looked to identify topics that were most frequently discussed in the various datasets and to return those topics with words that best summarize them. If you are not familiar with these methods, please conduct some external research in order to fully understand how they work. The NLTK library was used in order to implement these algorithms, which has a very extensive API online.

The algorithms produced similar findings in that the articles of Fake News that mentioned Hillary Clinton had frequent mention of the words email and FBI. Other words that pertained to her investigation appeared frequently. Another interesting keyword that was returned by LDP was a link to various articles. It was considered to further clean up this data to get rid of links and any other non-dictionary terms, but the significance of the links show how these Fake News articles attempt to make money. The more links that readers click on, the more advertising revenue the Fake News authors will receive. This shows that these authors and websites have a blatant disregard for integrity and only care about money. The lack of

detailed results from Hillary's dataset of articles provided an interesting comparison with those that mentioned Donald Trump.

The keywords and topics identified in Donald Trump's dataset included more variety than that of Hillary. Words ranged from "campaign" to "hat" to "woman" to "khan." These words reference a lot of different segments of Donald Trump's Presidential Campaign. The words in Donald Trump's topic list provided more insight into his campaign when compared to the words from Hillary's dataset. The only consistently significant words from her dataset referenced to her tribulations with the FBI probe. It is to be noted that this could have occurred because Hillary's dataset of articles had five less than Trump's. Yet, the aforementioned observations are still relevant when comparing the effects of Fake News in each political party.

The overall findings from these articles reveal two main takeaways. The first being that the authors of Fake News articles include excessive links in order to further profit off of their malicious content. The links that are published often point to an article on the same website in which that article was published. This phenomenon illustrates that the authors manipulate readers for the sole purpose of making money. The second main takeaway refers to the comparison between the Donald Trump and Hillary Clinton Fake News articles. The Donald Trump articles contained wide-ranging topics that made detailed references to his campaign. It is hard to tell whether these references are positive or negative, but comparing that with Hillary Clinton's dataset illustrates a strong contrast. Hillary Clinton's articles largely focused on her FBI hearings where her personal data servers received much public scrutiny. These articles likely made false and negative claims about her servers, that could have potentially harmed her campaign. Yet, it is hard to tell whether the articles benefitted or harmed one candidate based on the algorithms. It is certain that Hillary's article content solely focused on her servers, implying her untrustworthiness, while Trump's article content covered a wider array of his campaign.

Conclusion

In summation, this research highlighted the dangers of Fake News and how it could severely affect the future of news. While no conclusive evidence of benefitting one political party was found, it is evident that Fake News authors profited off of the 2016 Election. The articles in this dataset show that Hillary Clinton received more targeted attacks on her private server than that of her opponent Donald Trump, yet Trump's articles covered a lot more of his campaign. The effects of this cannot be quantified, but it should be known that Fake News authors will continue to exploit their techniques as their audience continues to fuel their material.

In order to continue with research in this field, one must find an accurate way to classify Fake News. Whether this is through fact checking or revealing author bias, the general public and the technology company's responsible for content distribution need find clear ways to identify Fake News. It is evident that even if it is hard to quantify the impact Fake News had on the 2016 Presidential Election, the issue will only become more troublesome in the future if immediate steps are not taken.

Acknowledgments

Thank you to Shion Guha for teaching the Data Science class at Marquette University that gave me the opportunity to pursue this research.

Github

https://github.com/cjarmbrust1/cosc4931-introdatascience

References

1. Condliffe, J. Google's Algorithms May Feed You Fake News and Opinion as Fact. Retrieved March 8, 2017, from MIT Technology Review: https://www.technologyreview.com/s/603796/googles-algorithms-may-feed-you-fake-news-and-opinion-as-fact
2. How Big Is The Fake News Problem For Facebook?, 2016. Retrieved March 8, 2017, from Forbes: https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/greatspeculations/2016/12/21/how-big-is-the-fake-news-problem-for-facebook/&refURL=https://www.google.com/&referrer=https://www.google.com/
3. NLTK Documentation 3.0. Retrieved April 11, 2017, from Natural Language Toolkit: http://www.nltk.org/