CJ Armbrust

Shion Guha

COSC 4931

9 March 2017

Fake News Write Up

Fake News has become a problem within our country over the last five months. Then presidential candidate, Donald Trump, called major news outlets providers of "Fake News" which brought this whole topic into the spotlight. By labeling news sources like this, his aim was to discredit their publications and bring confusion into journalism. This began to blur the lines of credible news sources in a time when a contested election brought about many questionable scandals and stories. Around this time, Facebook began to receive part of the blame for this since their platform allows potential "Fake News" to be widely distributed at a rapid rate [2]. Many began to wonder whether Facebook should censor content or if the problem is best left to the journalists and the consumers who read news. Facebook has begun to attack this problem and identify Fake News, but an obvious question arises: how do we define fake news?

A Kaggle project used 13,000 articles and a tool called BS Detector to build a dataset to answer this question about fake news. The dataset contains columns that record domain ranking of the website, a generated 'type' of fake news, a generated spam score, Facebook statistics, author, title, and some other attributes. The various types assigned to the articles are Fake, BS, State, Bias, Conspiracy, Hate, Junksci, and Satire. To start exploratory analysis, I looked at the types and how many articles belonged to each one. The BS type had significantly more articles than all of the others because any article that could not be classified by BS Detector was labeled as BS. I took all of the BS articles out of the dataset due to this finding.

Once these articles were removed from the dataset, there was a good distribution between all of the different types leaving about 1,300 articles. In the next step I wanted to use a density graph or strip plot to compare some of the other column values to the various types of Fake News. Constructing a strip plot with the attributes type and domain rank illustrated the general popularity of each category of fake news. The most notable result was that Junksci and Bias articles had the highest domain rankings out of the group. With this is mind, it will be revealing to discover if the Junksci and Bias articles have any political leaning. Next I created a similar strip plot, but with type and spam score. The spam score could be between 0 and 1, which represents the spam percentage an article had. The results indicate that all of the categories had plenty of articles without any spam, but Bias and Conspiracy articles tend to have higher rates of "spammy" articles. It will be interesting to isolate the articles with high rates of spam and run various clustering algorithms on them to see whether other classifications could form. In this series of graphs, Facebook shares were the last attribute analyzed. The resulting graph showed that a lot of the article types had similar Facebook sharing trends, but Hate and State articles had very low sharing tendencies. After further analysis, Facebook's difficulty of moderating Fake News will be more distinctly identifiable based off of article content and classification from this dataset.

After comparing the different classifications of Fake News with each other, I wanted to see if there was any correlation between other attributes. To do this, I put two attributes on a standard plot. First, I compared domain ranking and spam score, which showed about three different groupings of articles. The websites that had higher ranking domains generally had lower spam scores. Websites with a domain ranking between 20,000 and 50,000 had variety in their spam scores, with most articles having lower spam scores. Websites with very low ranking domains had an even distribution between lower spam scores and higher spam scores. Second, I examined the relationship between Facebook shares and domain ranking. The results were intuitive, the websites with popular articles and higher domains had higher rates of Facebook sharing. All of the websites with higher traffic had varying Facebook sharing tendencies, which probably depended on the content of specific articles. This will be explored further in the semester once I examine the content of the articles in the dataset. Third, I plotted Facebook shares against the spam scores of each article. The results illustrated that articles with higher rates of sharing had very low rates of spam. This was a vital finding because it demonstrates that spam does not travel on Facebook very easily, but this does not ensure that the articles could not have misleading messages.

The last part of my exploratory analysis included boxplots comparing some of the attributes from above and a few other attributes. The first boxplot showed that Junksci articles had the widest range of Facebook sharing. The articles in this classification probably include exaggerated or invalid science claims that the average person without deep science knowledge finds engaging. Another noteworthy finding is that state and hate articles had no Facebook distribution, which could lead to taking out those classifications in further computations. Using a boxplot to analyze domain ranking illustrated that Junksci and Bias articles had more popularity than the other categories. Fake articles and some conspiracy articles had very low domain rankings, which will be important in the later steps to analyze if their specific contents affected the popularity of the website. Using a boxplot to analyze Facebook likes yielded similar results to the Facebook sharing data. In later steps, it will be useful to perform textual analysis on these articles to see what topics generated the most Facebook interaction.

Exploratory analysis of this dataset revealed some important trends in the data that will be used for future research. One of the most pivotal findings was that domain rank and spam score did not have a clear relationship between each other. This will be an important topic to continue to delve into because it can reveal if popular websites have tendencies to publish articles with spam. This has been a result problem for Google as their Snippet tool that uses data from website to answer questions quickly has been giving incorrect answers [1]. The other topic that will be interesting to do more analysis on is the relationship between Facebook engagement and article content. After understanding these relationships and other findings, a clearer and concise definition of Fake News could be constructed. Also, reaching conclusions about the sharing and publishing of Fake News can help companies like Facebook and Google censor their services and keep the public rightfully informed.

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/