Analysis Report

Student ID: 1193813

This assignment required us to analyze a number of articles written by various news sources. With our analyses, we were able to categorize the various articles,their news sources and their words into groups and decide how reliable they were.

In task 4, we are able to visualize how reliable each news source is by taking the average rating of all articles from a specific news source and looking at it in comparison to other news sources. The average rating of each of these news sources is based on data collected from more then 5 articles, making the analyses quite credible.

In task 5, we looked at a number of tweets and retweets on each article and examined if the rating of the article affects the number of tweets that it receives. The correlation between the rating and the number of tweets of an article is linear .We observe that as the rating of the article increases, the number of tweets also increases. That does not make it necessarily true that the more the credibility of new_source, the more the number of tweets, but there is a possibility.

In task 6, we analyzed the text of each article, and we tokenized it, bringing it to a form that is easier to analyze. We then stored the articles where each word is used for further analyses in task 7.

In task 7, we find out the log odds ratio of each word that has been used throughout the article reviews, and use the article ratings to analyze if each word is more likely to be used in a real story or in a fake story. A key analysis from figure 7b tells us that the 3rd Quartile is very high for the log odds ratio compared to the 1st and 2nd Quartile. The words above the 3rd Quartile also have a relatively high log odds ratio. This suggests that words that have high odds log ratios are a lot less and alot more likely to be repeated than those the words with low log odds ratios.

In Figure 7c, we analyze the words with the top 15 log odds ratios and alongside the words with the lowest 15 log odds ratios. We see that the top 15 log odd ratios (most likely indicated fake news) are all significantly higher than the bottom 15. They increase steeply as we go more towards the highest value, as compared to the bottom 15, which can be observed to have almost constant log odds ratios.

This data set would’ve been better if it did not contain a number of fields that were empty or had missing data. The result could also have been improved if we were given further detail on the news_sources, such as the authors of these articles. In the future, it can also be further analyzed to compare if the ratings vary for different years to give an analysis on the growth of fake news over time.