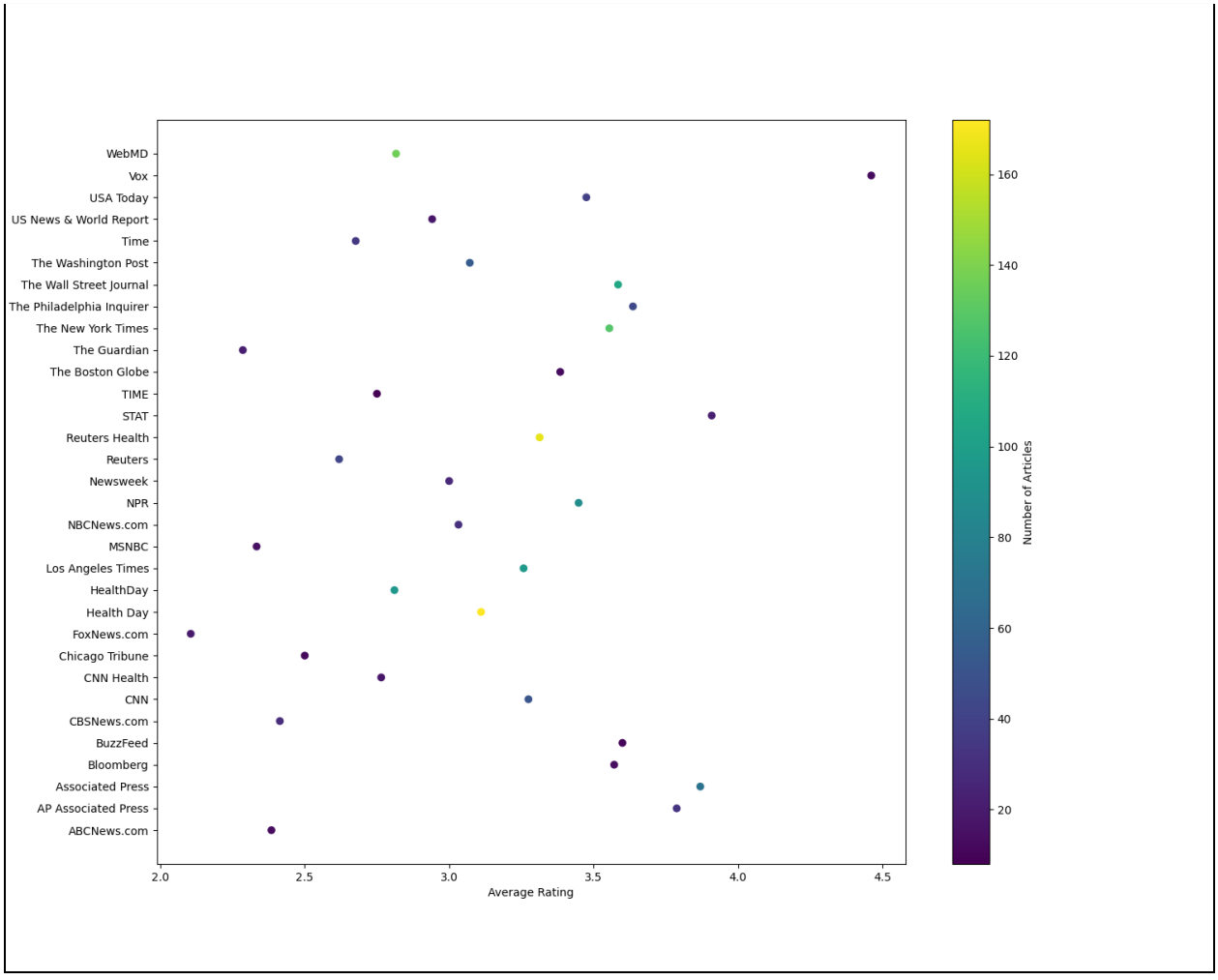
Assignment Report

Dataset

Most of the information we learn today comes from news sources, however there is plenty of news that contains untrustworthy information. This dataset contains more than 1600 health-related news articles with its review and twitter post. Based on that, we can determine the stereotype of fake news and real news.

Credibility based on task4b.png

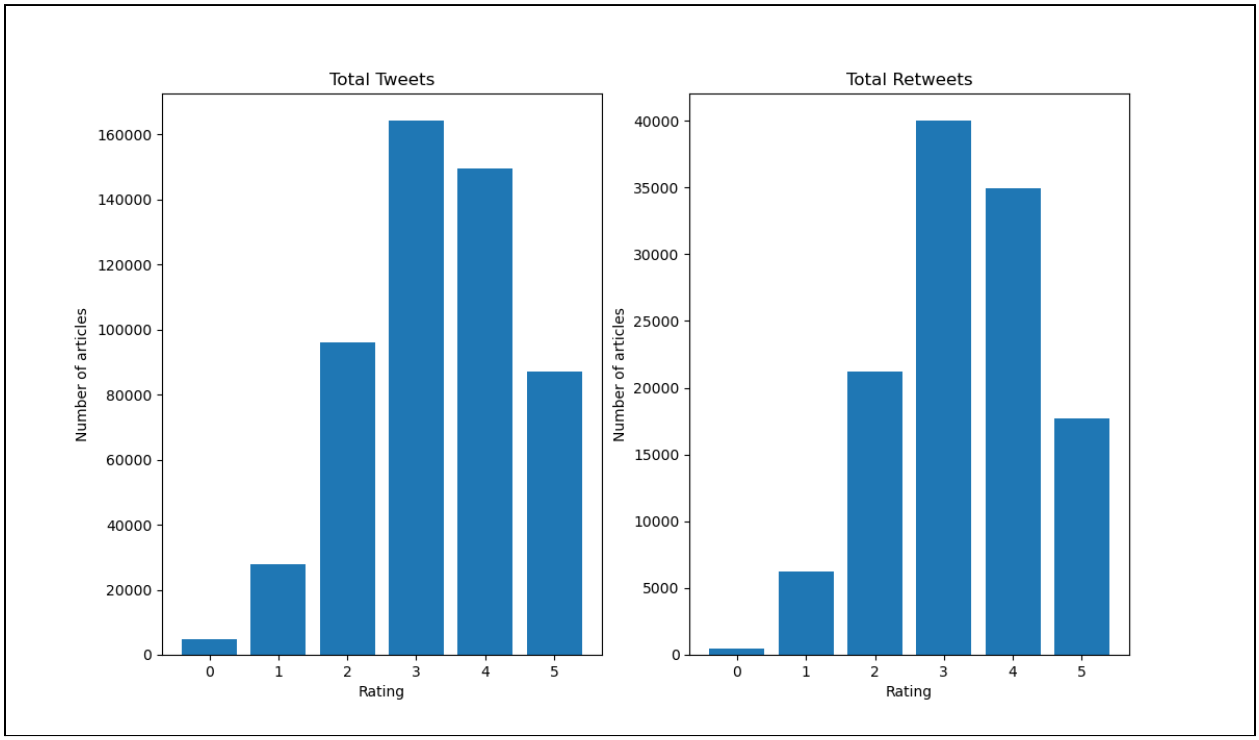


Task4b.png

This figure (task4b.png) shows the correlation between news sources and its average rating. Even though the average rating of 'The Wall Street Journal' and 'The New York Times" just exceed 3.5, they have over 120 articles. Moreover, 'Vox' is the most trustworthy news source with the largest average rating about 4.5. On the other hand, 'The Guardian', 'Fox News',

‘ABCNews.com’, and ‘MSNBC’ provide the worst news sources. Moreover, ‘WebMB” provides most of the fake news in the dataset.

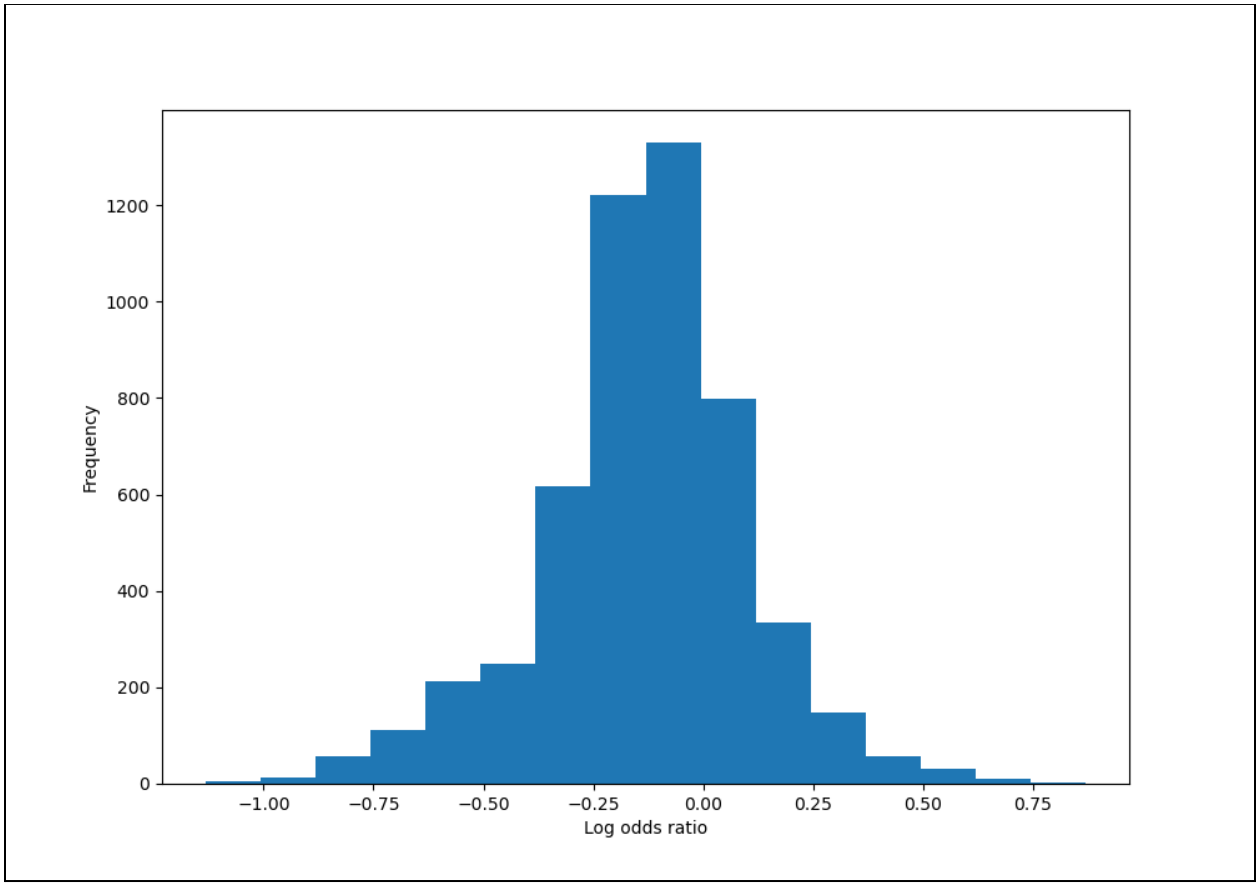
The correlation between number of tweets and credibility of articles



Task5b.png

This figure (task5b.png) shows the information about the number of tweets and retweets for each rating of news articles. In detail, the number of tweets and retweets lay around the rating of 3 and 4. On the other hand, articles that have ratings under 3, have less tweets about it than others. Then, it can answer that the number of tweets and retweets correlates with the high rating. Therefore, the more credibility of articles lead to the more it is retweeted.

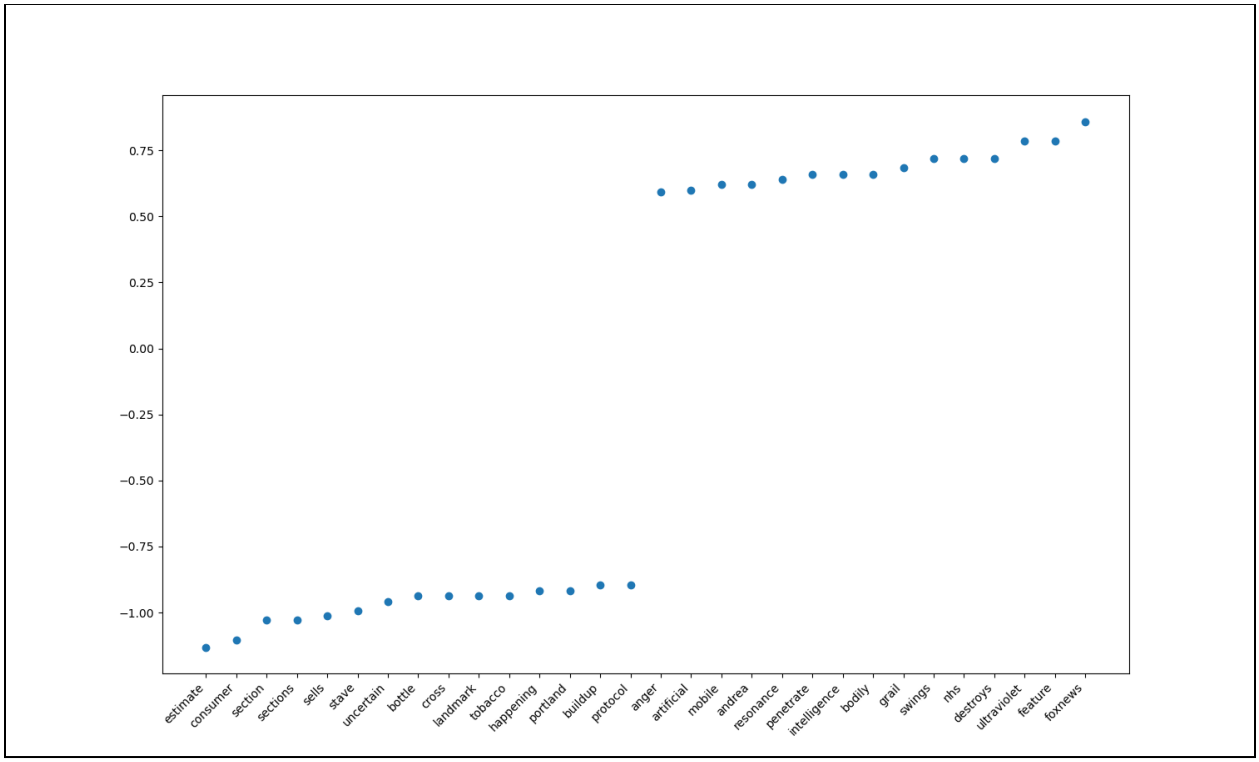
Distribution of log odds ratio



task7b.png

This figure (task7b.png) shows the distribution of log odds ratios for all words that appear in all articles. From this figure, we can see that there are groups of words that are more likely to appear in fake news and trustworthy ones, and we can use this group of words to determine which news is likely to be fake and real.

The most indicative words of fake news and real news



task7c.png

This figure (task7c.png) shows the 15 largest and 15 smallest log odd ratios of words used in articles. The largest log odd ratios represent the words that are frequently used in fake news, vise versa. Most of the words used in fake news describe certainty and vagueness, or new technology that people are not well informed about. In real life, scientists are continuously researching about the human body, and there are so many things that are uncertain. Therefore, the more certainty the words in news are, the higher probability that it is fake news.

The limitation of this dataset

Even though this dataset contains a large amount of articles, it is not objective because the number of fake news and real news are not the same. In detail, the result of all the plots using this dataset may change because most of the articles have ratings bigger than 3. We cannot use this dataset as a sample for machine learning to detect fake news. Moreover, the only information about a tweet is the number, we need to determine which tweet has a positive or negative approach.