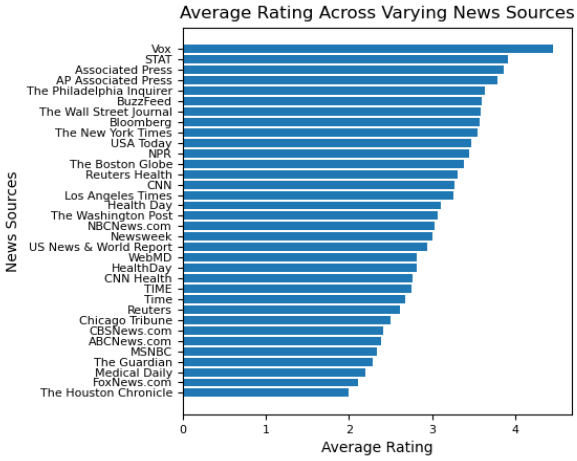
Liam Bong 1272915

Introduction:

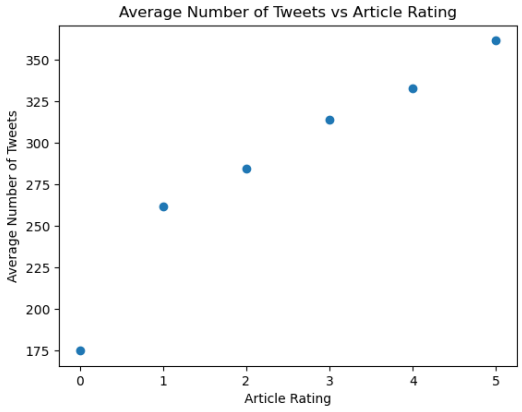
The data sources utilised in tasks 4-7 consists of a folder holding distinct files - each, for which there are over a thousand, containing information on a health-related news article – and two separate json files, which collate relevant reviews/tweets concerning the articles within the said folder. This data analysis centres around using the dataset to comment on the credibility across various news sources, identify any correlation between twitter popularity and article credibility. And finally, identify any difference in the distinct words that are contained in credible and untrustworthy articles.

Body:

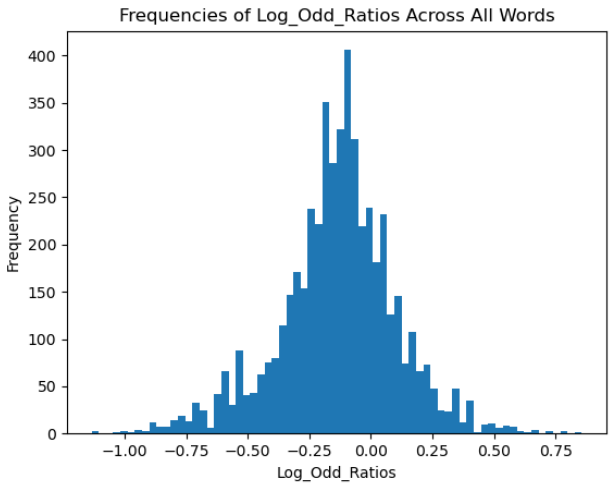


The graph above was constructed by utilising the data in the reviews json file to group each article and respective rating by their news source. Its construction has allowed us to identify “Vox” and “STAT” as the most credible sources and “The Houston Chronicle” and “FoxNews” as the least, with sources generally having ratings between 2 and 4.

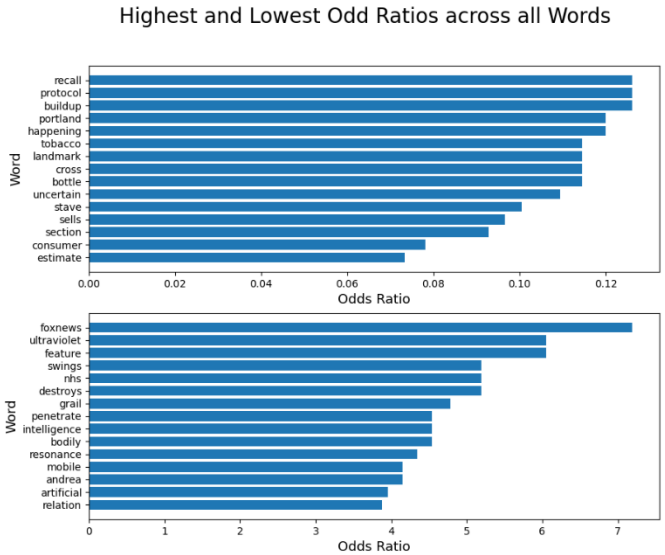
Similarly, task 5 utilises the tweet and review data sources to group articles by rating to assess correlations between twitter popularity and credibility. The scatter plot suggests a clear positive (perhaps linear) relationship between the average number of tweets and the rating of an article. We can consequently infer more articles receive higher number of retweets.



The task 7 process begun by flagging articles into “real” or “fake” news. Which facilitated the processing of the task 6 data, a vocabulary of distinct words and the articles they were present in, to ultimately assign “ratio(s) comparing the odds that w appears in a fake article against the odds that w appears in a real article” to each word.



The histogram closely resembles a bell curve, and the word frequencies are symmetrically distributed about its centre. The data could likely represent a normal distribution. Since the values of the log-odd-ratios are less spread out, we can infer that there is less standard deviation within the data. Furthermore, the negative mean suggests that the majority of words are slightly more likely to appear in real articles. However, generally only a minority of words are strongly indicative of the article credibility.



Thorough analysis has identified words like “foxnews” and “ultraviolet” to be most indicative of fake articles, while “recall” and “protocol” are identified to be the most indicative of real articles. I can only moderately agree to the reliability of these word pools for practical use. There seems to be a reasonable and logical commonality between indicative fake words, in that they are often confident and persuasive (eg; intelligence, destroys) comparative to less confident, real words (eg; uncertain, estimate). Despite this, my personal experience reading health articles leads me to believe that the identification of these “indicative” words in an article is not sufficient to effectively ascertain its legitimacy/credibility.

Conclusion:

The popularity and linguistic patterns of an article have been seen to vary as credibility varies. Although the dataset was concrete, the reliability of the reviews was limited by the number of and trustworthiness of the reviewers. Furthermore, our assessment of “popularity” was limited by the twitter demographic (25-34 years old). Besides this, our tokenisation process resulted in the generation of tokens which were not real words. Perhaps in future analysis, further processing of the tokens could be implemented and the dataset could expand past the twitter demographic.