COMP20008 Elements of Data Processing Assignment 1

Name: Max Yi-Hong Ruan

Student ID: 835040

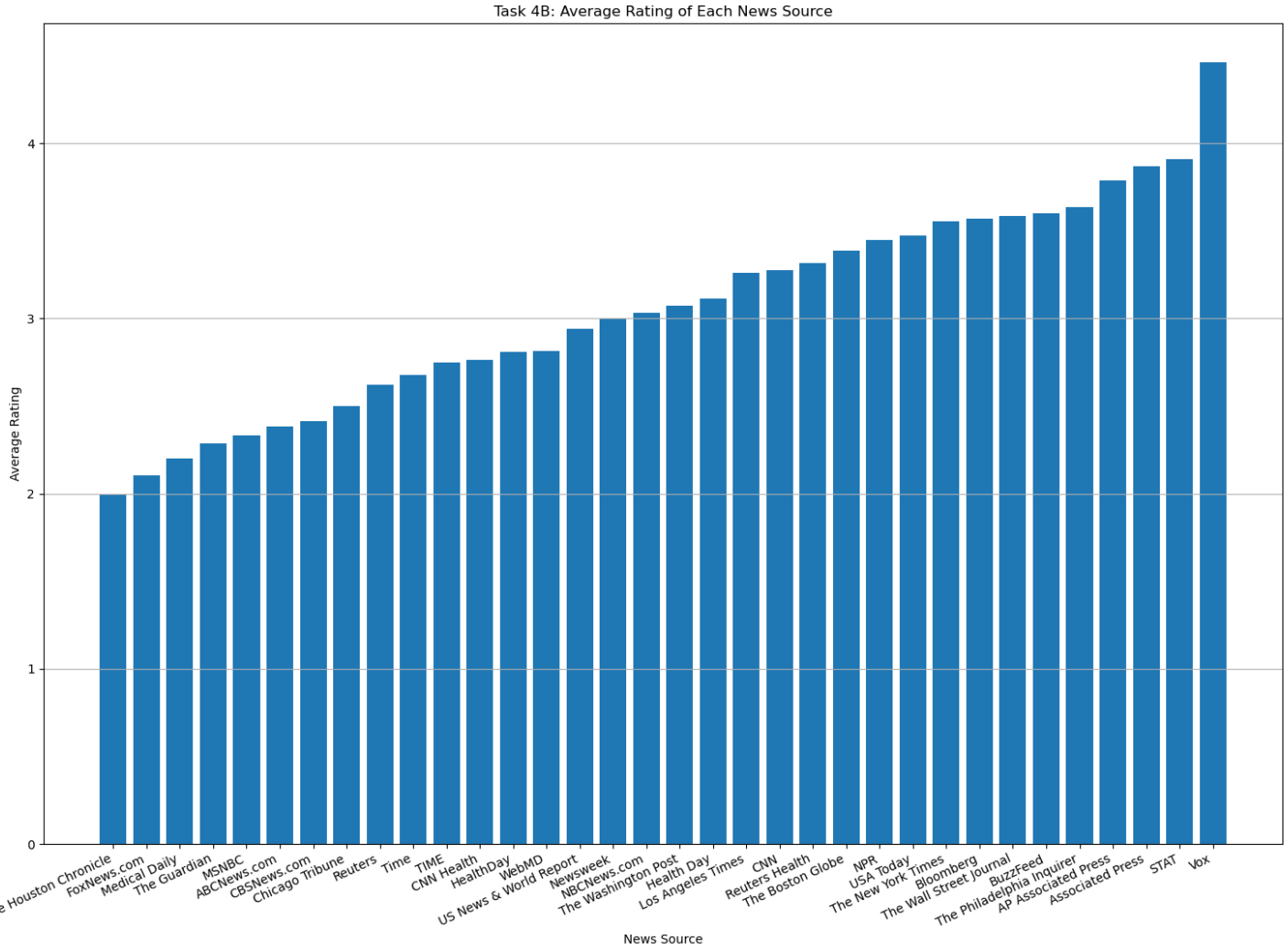
Word Count: 548 (Word Limit: 500)

Data

The dataset consists of news articles, reviews of these articles and tweets about the new articles.

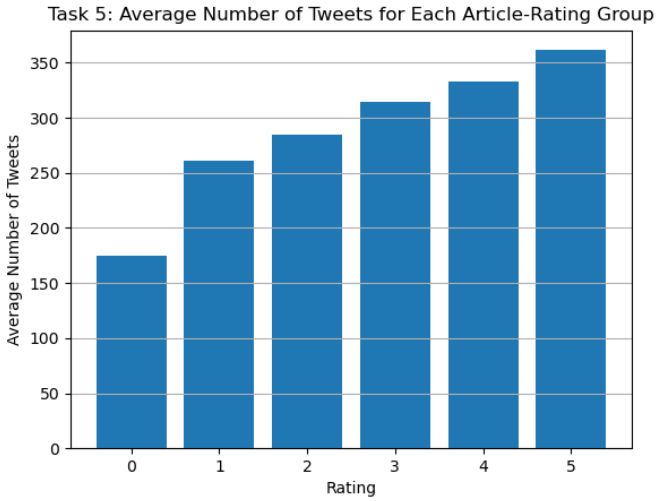
- Article data captures attributes including the article title, author/s, full body of text and news source. Article id is found in the JSON file name.
- Review data captures attributes including the review title, description, article information, overall review rating and a list of criteria with an answer and explanation for each.
- Tweet data contains lists for article tweets, replies and retweets.

Methods/Output



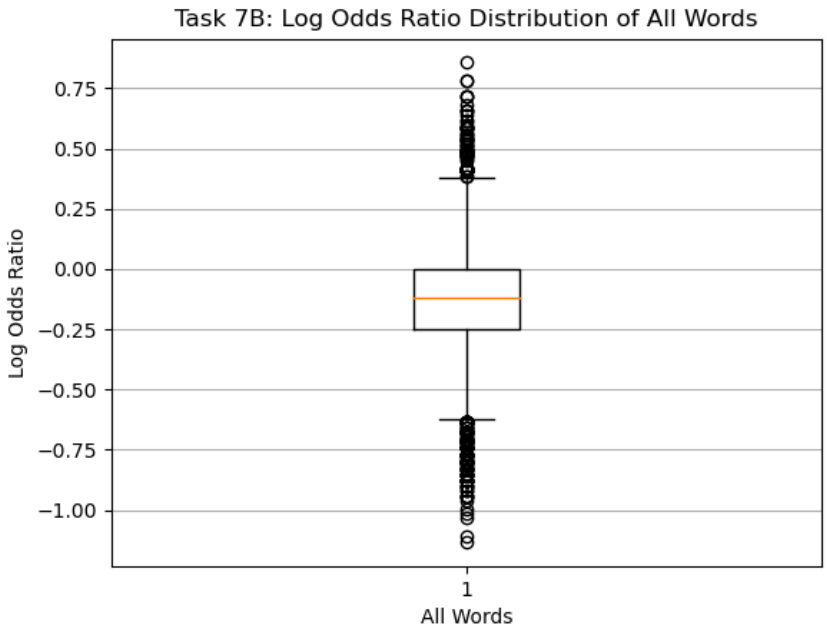
The bar graph above shows the average rating for each news source. The news sources are sorted by average rating from lowest to highest. Only news sources with at least 5 articles are present and articles with a blank field for news source were excluded.

It can be inferred that Vox, STAT and Associated Press are some of the most credible news sources whereas The Houston Chronicle, FoxNews.com, Medical Daily are some of the least credible news sources.



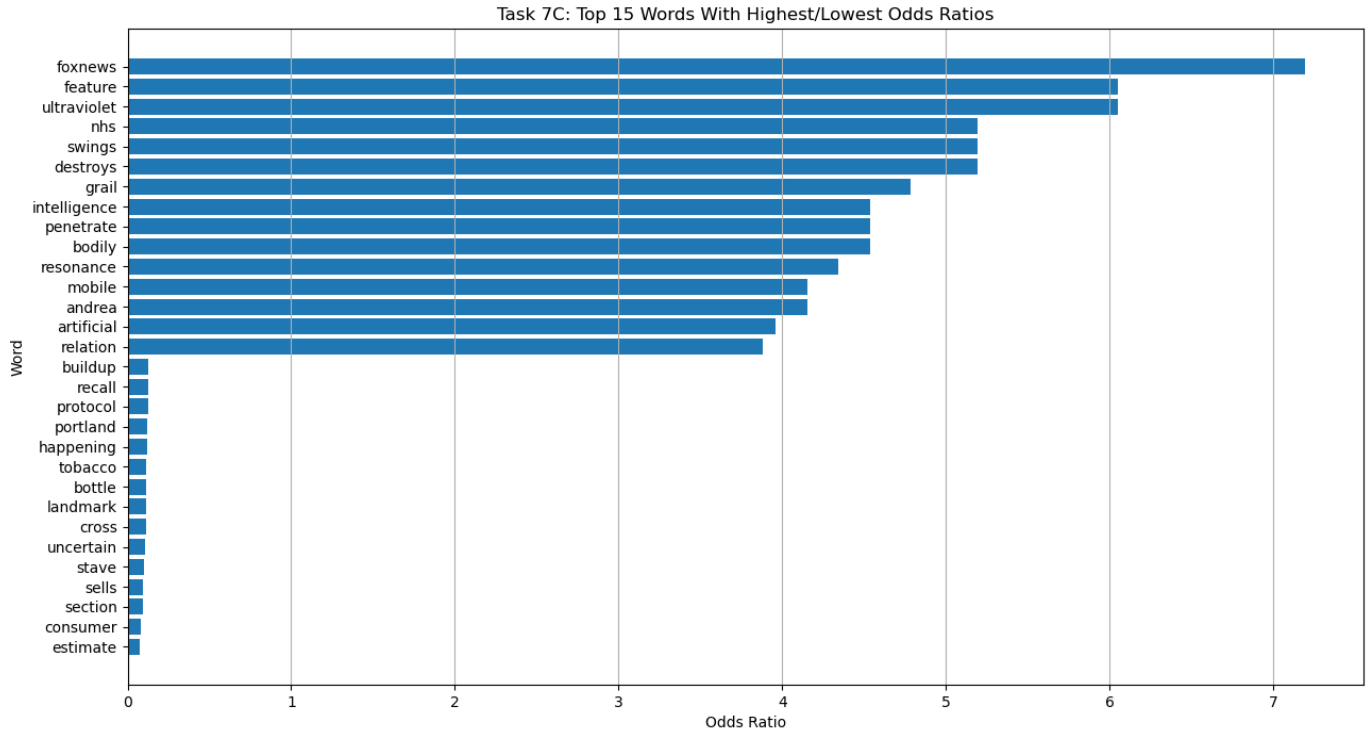
The bar graph above shows the average number of tweets (non-duplicate sum of tweets, retweets and replies) for each article-rating group.

It can be inferred that there is a positive relationship between average number of tweets and rating group. This suggests that more credible news is shared more than less credible news. However, the graph doesn't necessarily tell us whether more credible news articles are retweeted more as tweets also includes original tweets and replies.



The box plot shows the distribution of the log odds ratio of all words. However, it does exclude words that appear exclusively in real or fake news and those that appear in either fewer than 10 articles or in all articles.

The median/orange line (50% words are above/below this point) hovers between 0 and -0.25. A positive logs odds ratio indicates that a word is more likely to appear in a fake article. Thus, it seems to suggest that there are more words that have a greater likelihood to appear in a real article. The upper and lower ends of the box represent quartiles 3 and 1 respectively. Data points beyond the 1.5x IQR whiskers represent outliers.



The horizontal bar graph above shows the top 15 words with the lowest odds ratios and highest odds ratios for fake news, sorted by odds ratio value. This excludes words that appear exclusively in real news or fake news and excludes those that appear in fewer than 10 articles or appear in all articles. An odds ratio greater than 1 indicates a word is more likely to appear in a fake article.

I am indifferent to some of these words but do agree with words such as ‘artificial’, ‘destroys’ and ‘penetrate’. I associate the latter two with damage and fearmongering and often said in articles related to viruses and COVID. ‘artificial’ is one I can associate with unsupported alternative health fads such as naturopathy.

Conclusion

There are several improvements that could be implemented. For example, we see that the current text pre-processing in Task 6 is improperly strips letters/words with accents. In addition, analysis in Task 5 could aim to breakdown tweets into original and retweets or outright remove original tweets from consideration to more accurate analyse the relationship between credibility and social media sharing. Empty string columns can be standardised to be missing values e.g. blank news source in Task 4.