The data for task 4, give us the articles ids and the news source it came from. Moreover, it gives the rating of all articles. This allows us to get the average rating per news source. Plot 4 is a bar chart that represents the average rating of all the news sources that were available in the data given.

For task 5, allowed us to get the number of tweets per articles. Furthermore, we accessed the reviews folder to get more information on the articles and get their rating. Plot 5, shows the number of non-repeated tweets per rating.

Task6, we tokenized non repeated words in each article and removed all stop words. This was all taken from the text value of each article folder. The output represents all each unique word in all the articles and assigned an array of article ids that they can be found in.

Task7, we check the news reviews to find the rating of all the articles and classify them as real of fake based on their rating. The second output shows the frequency of log odd ratios for all words, this gives more insight in the distribution. The last output compares the words with the top 15 highest odds ratio and the bottom 15 words with the lowest odds ratio.

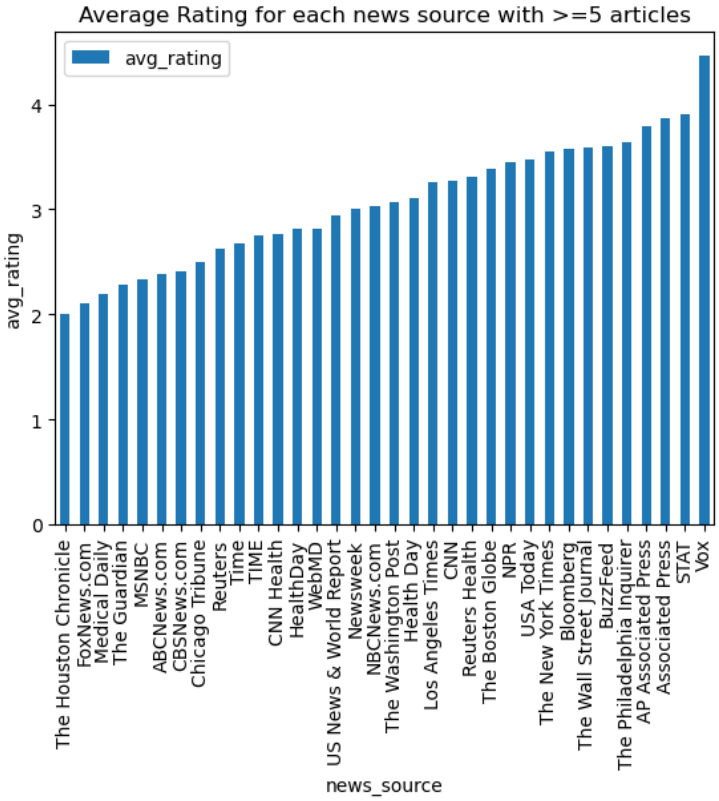
According to the graph represented in figure 4b, the news source with the highest average rating is “Vox”, followed by “STAT” then “AP Associated Press”. The lowest ones are “The Houston chronicles”, “FoxNews.com” and “Medical Daily”. A possible problem that could appear when evaluating the data is the number of articles per news source isn’t the same for each news source.

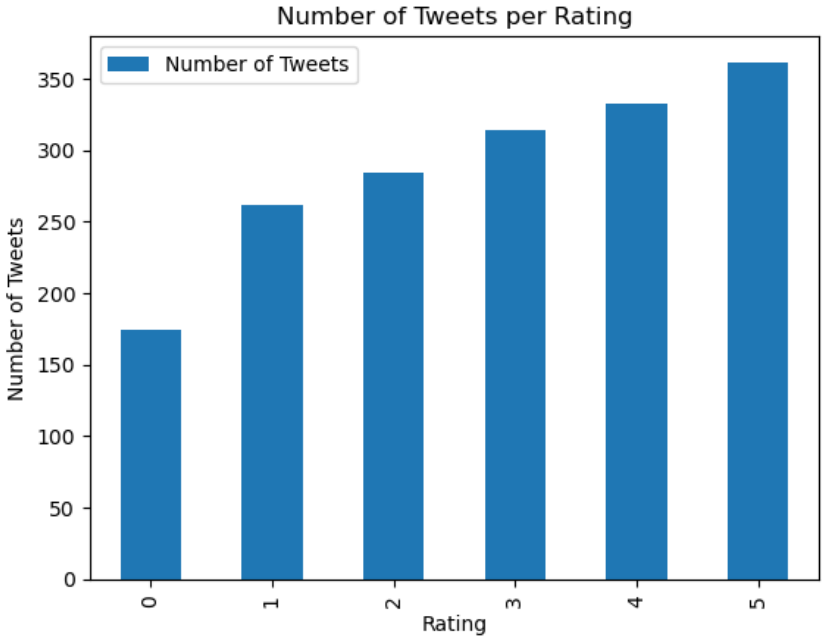
According to the output of task5, we can see that the higher the rating of an article the more Tweets it will receive. However, the change between number of tweets per rating isn’t high enough to make the number of tweets the only valid parameter to take into consideration.

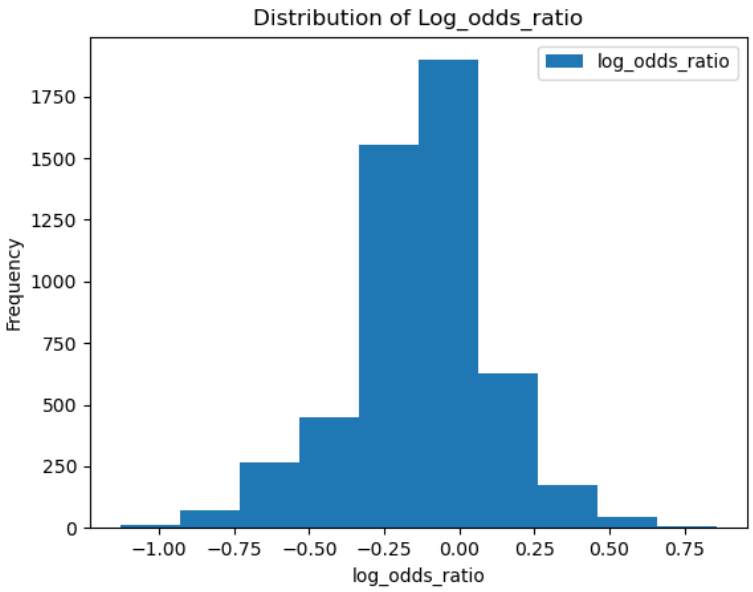
Plot 7b, shows a high number of words appear will appear in both Real and Fake news articles (log_oddratio =0). Moreover, the distribution of negative log odd ratio appears to be more prominent meaning that more words appear in Real news rather than Fake news.

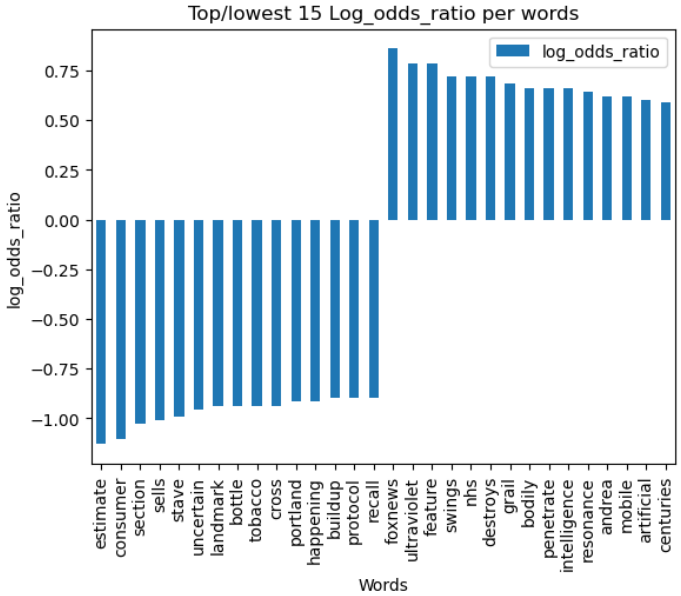
The graph shown in task7c, includes common words which are associated with real articles such as estimate, sells and consumer. This was unexpected since in my experience more sophisticated words that are associated with the theme of an article would represent whether and article is real or fake. However, after some reconsideration sophisticated words would be more common in both articles since they all cover Health. Of that reason, it would appear that checking the ratio of words would not prove to be helpful in sorting real and fake news. Rather a combination of words(phrases) would prove more effective in my opinion.

Overall, words aren’t the only parameter we should consider when judging an article. The number of tweets and the rating are related since they show that if the rating is higher the more tweets it will get however, it is important to consider that the number of tweets isn’t a good indicator of rating as Viral fake news seem get more attention in social media nowadays in my experience (example: negative/positive effects of Vaccines news). Furthermore, I agree with selecting the source of the news and comparing it’s rating to get an overall idea of real/fake of article.


Average Rating for each news source with >=5 articles
avg_rating
avg_rating
news_source
The Houston Chronicle
FoxNews.com
Medical Daily
The Guardian
MSNBC
ABCNews.com
CBSNews.com
Chicago Tribune
Reuters
Time
TIME
CNN Health
HealthDay
WebMD
US News & World Report
Newsweek
NBCNews.com
The Washington Post
Health Day
Los Angeles Times
CNN
Reuters Health
The Boston Globe
NPR
USA Today
The New York Times
Bloomberg
The Wall Street Journal
BuzzFeed
The Philadelphia Inquirer
AP Associated Press
Associated Press
STAT
Vox


Number of Tweets per Rating
Number of Tweets
Number of Tweets
Rating


Distribution of Log_odds_ratio
log_odds_ratio
Frequency
log_odds_ratio


Top/lowest 15 Log_odds_ratio per words
log_odds_ratio
log_odds_ratio
Words
estimate
consumer
section
sells
stave
uncertain
landmark
bottle
tobacco
cross
portland
happening
buildup
protocol
recall
foxnews
ultraviolet
feature
swings
nhs
destroys
grail
bodily
penetrate
intelligence
resonance
andrea
mobile
artificial
centuries