ASSIGNMENT 1

REPORT

KUNAL PATEL (1291822)

1. Data Source description

The dataset given contains:

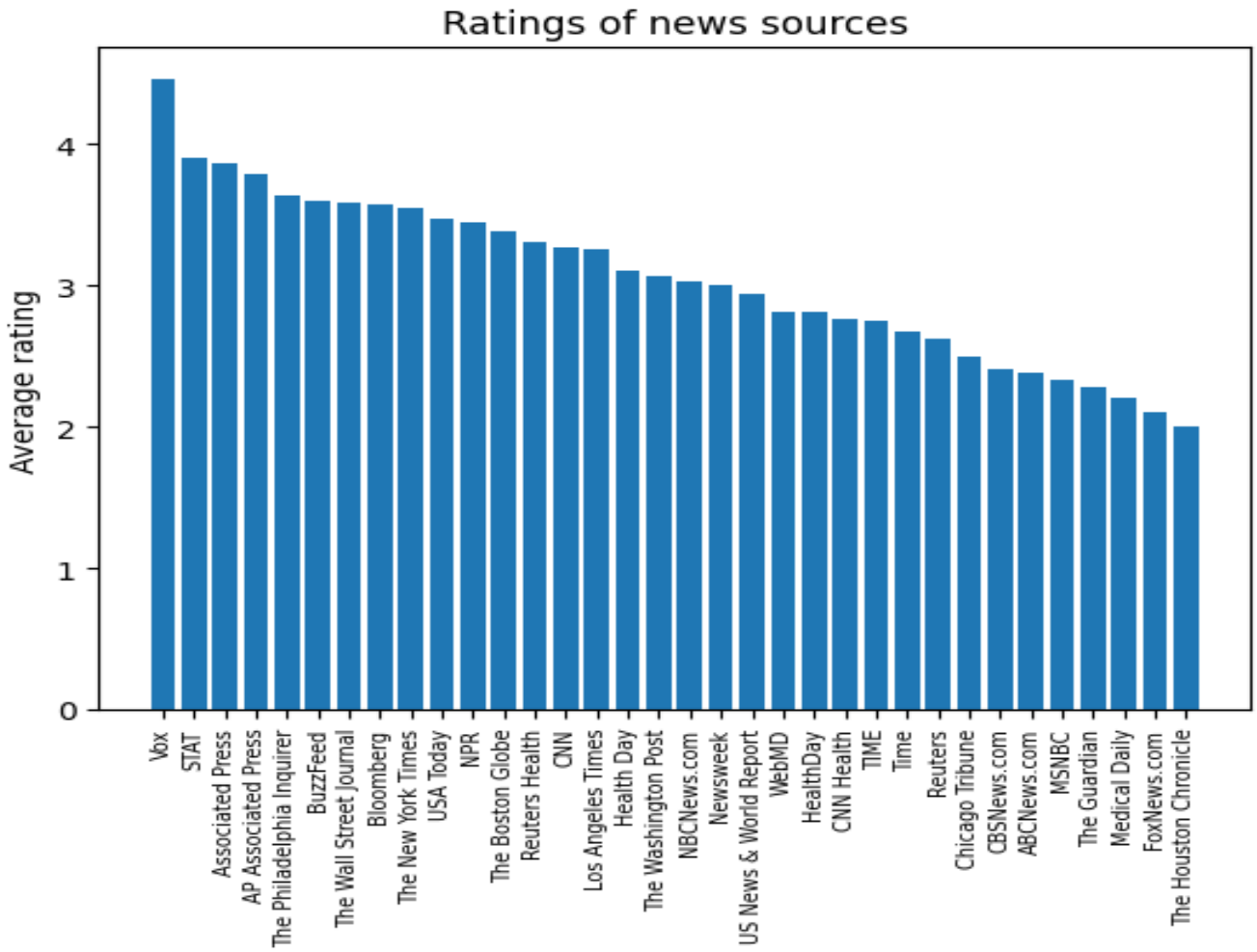
a. A folder **content** that has 1638 news articles from different news sources.
b. A folder **engagements** that contains a file 'HealthStory.json' which contains the info about tweets that the news articles received.
c. A folder **reviews** that contains a file 'HealthStory.json' which contains info about reviews of these articles done by experts.

2. Plot description

a. Task4 plot shows average ratings of the articles published by the news sources.
b. Task5 plot shows average popularity of the news articles based on their ratings.
c. Task7b plot shows the distribution of odds ratio or the likelihood of word being used in fake news compared to real news.
d. Task7c plot shows the 15 words most likely to be used in fake news and 15 words least likely to be used in fake news.

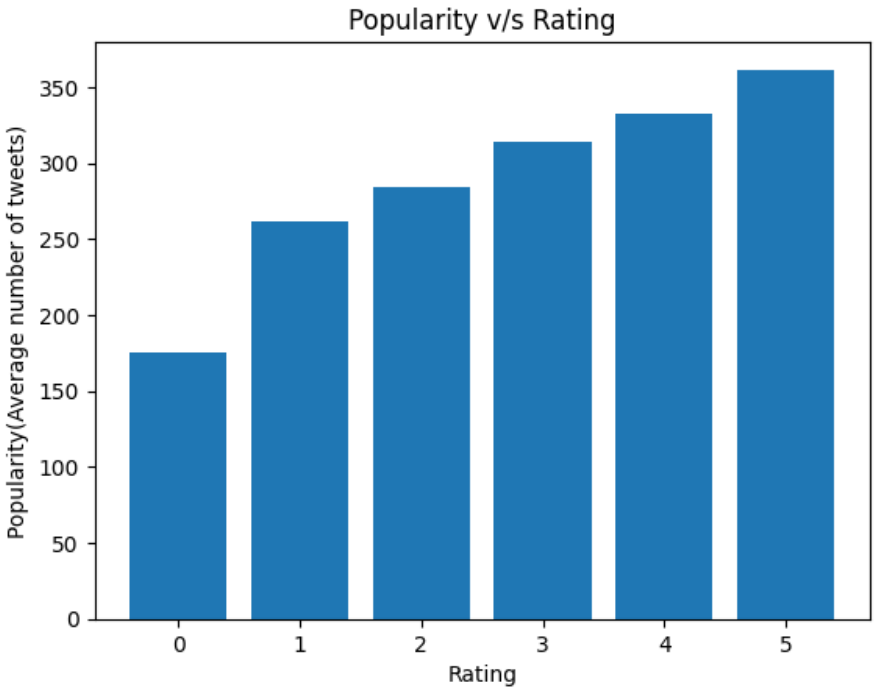
3. Credibility of news sources

Based on the graph of task4 below we see that **Vox news** is the most credible among the given sources with average rating of 4.46/5 while **The Houston Chronicle** is the least credible with average rating of mere 2.0/5.



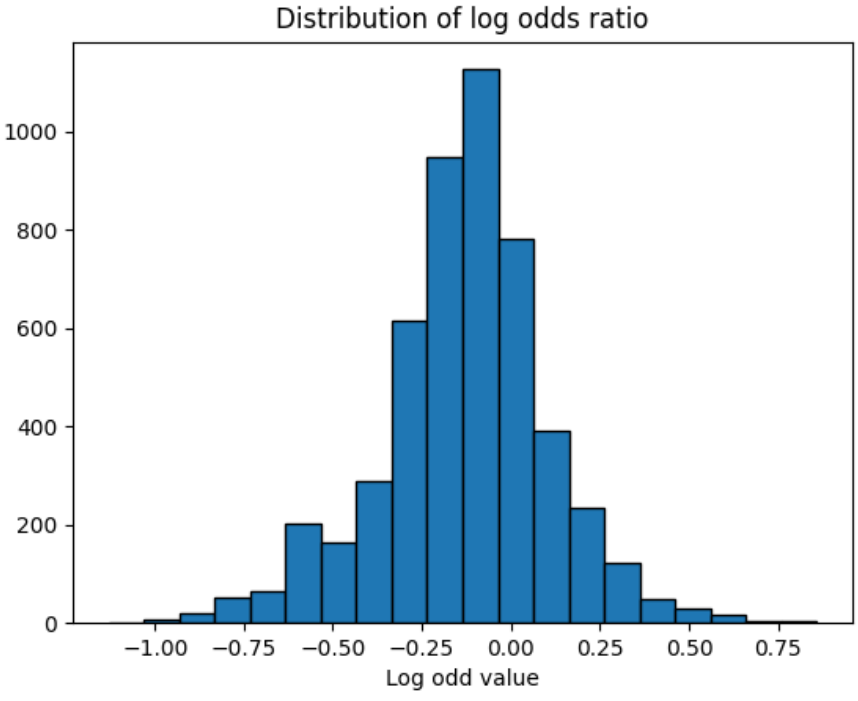
4. **Credibility and tweets**

   From the graph of task 5 given below we can see that the average number of tweets increase steadily with ratings, which means the more credible the news article the more it is tweeted.



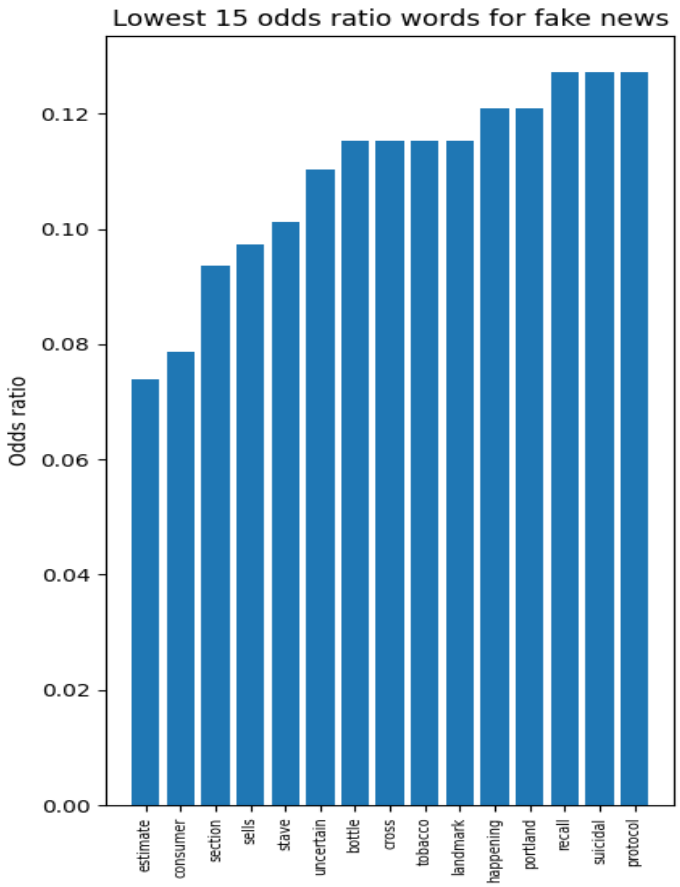
5. Distribution of log odds

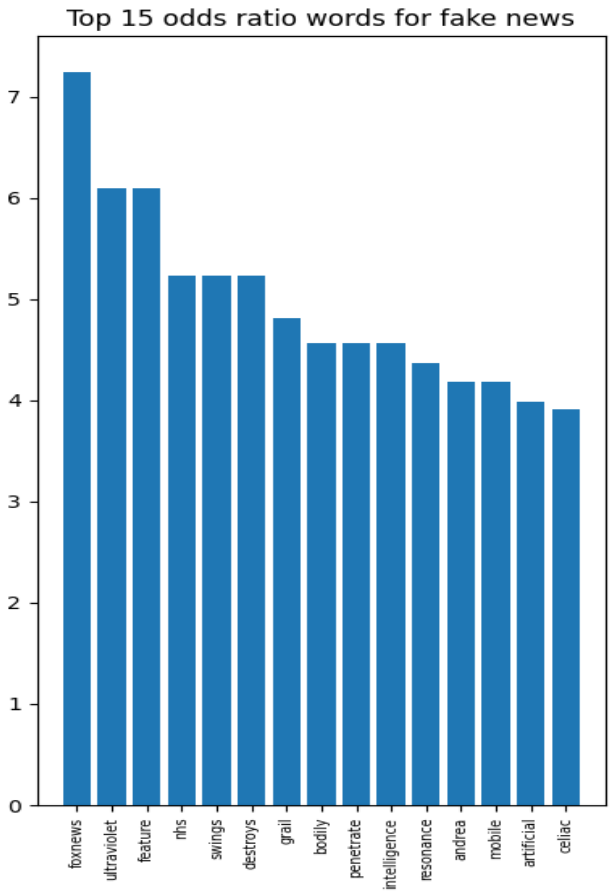
From the distribution of log odds in task7b shown below, we see that most of the words have log odd near zero which means that most of the words are used equally in both fake and real news, while only a very few words are more significantly used in fake news and vice-versa. This means we can look at those few words at the extremes of the distribution rather than all the words and tell whether the news is fake or real with high confidence.



6. Agreement with the ‘most indicative’ words

I agree with the most indicative words of fake news and real news shown in graph of task7c, given below. Words like estimate, protocol, recall are used in research journals and are mostly used by people in research and thus the articles having them have a high chance of being real. In my experience, FoxNews has not been a credible news source and I have seen words like destroys, intelligence, ultraviolet, artificial to be mostly used like ‘This method increases your intelligence’ or ‘This contains artificial this and is harmful’ etc. Such articles are either written by non-experts or are funded by companies who want to market products like natural oils. Thus, these words are more likely to be found in fake news.





7. Limitations

In the dataset we have:

a. 27 articles with 0 rating
b. 107 articles with 1 rating
c. 338 articles with 2 rating
d. 525 articles with 3 rating
e. 452 articles with 4 rating
f. 241 articles with 5 rating

The difference in number of articles for each rating is huge and can create biases. The number of articles must be uniformly distributed wrt ratings or at least the number of fake news articles must be almost equal to the number of real news articles for a better analysis.