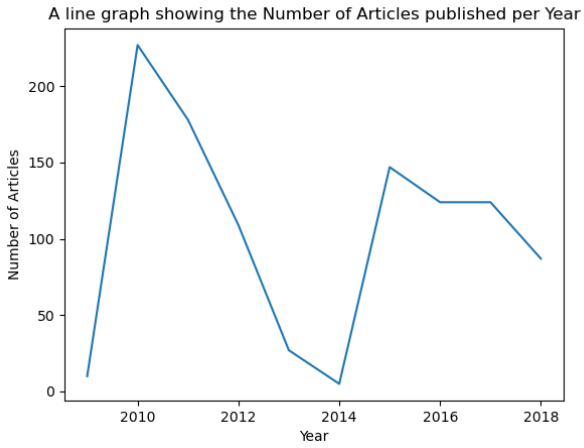
Analysis Report

Through various methods of data extraction and assessment, approximately 1600 articles and reviews and 500k tweets have been processed to develop an understanding of news credibility, popularity and likelihood of words appearing in real news and fake news. The 3 data sources given were:

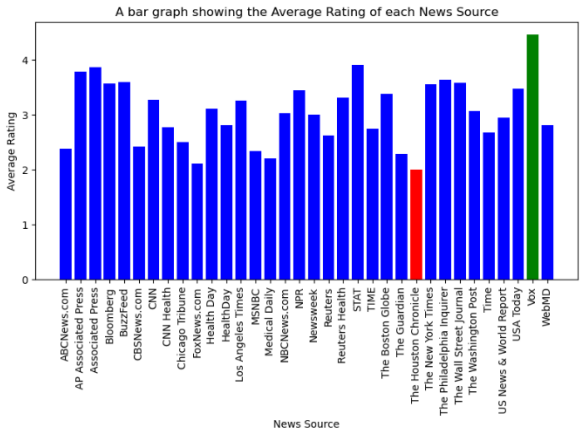
1. Articles: Filename is the news ID, give information on it's webpage, publish date and text.
2. Reviews: reviews of articles, gives us information on it's news source, a rating, and various criteria with decisions on if the article is satisfactory.
3. Tweets: each article has a set of values that represent a unique tweet, retweet, or reply.

Articles published per Year



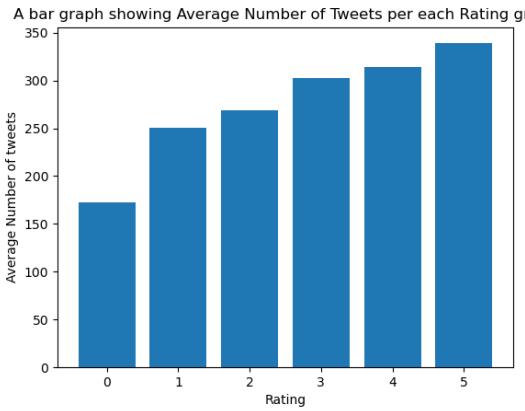
The graph shows no seasonality; the year with the largest number of articles produced is 2010 and the least is 2014. We can see a decline from 2010 to 2014 however is slowly increased from 2014 to 2015 slowly declining to the present.

Credibility of news sources



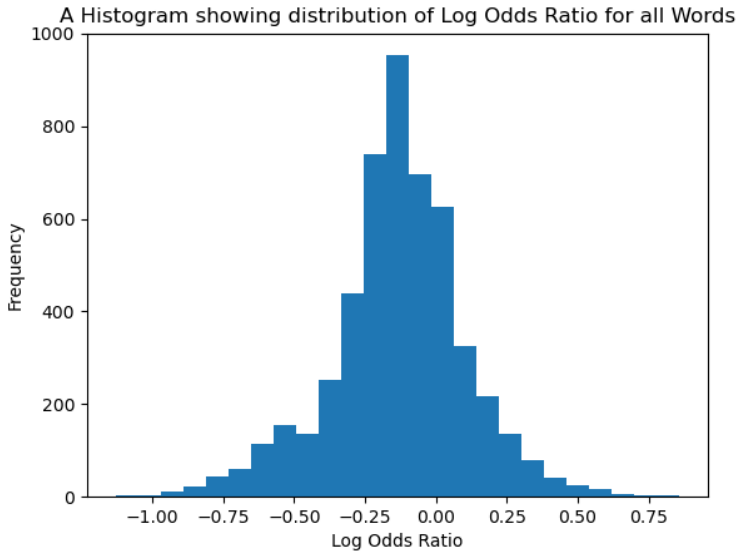
The Houston Chronicle seems to be the least credible source with an average of less than 2 whilst Vox is the most credible source with an average of 4.7. The majority of news sources average above 3 which shows that most sources have a satisfactory credit rating.

Popularity of Articles



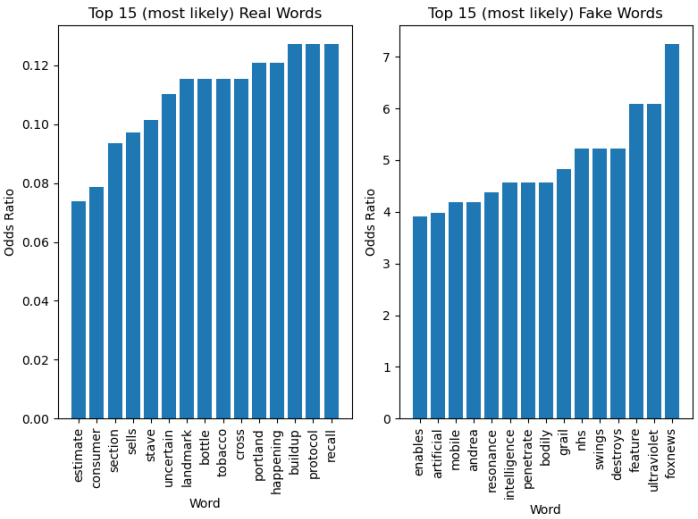
As we can that when the rating is higher, the number of tweets is also higher. This shows that articles with a higher rating are tweeted and retweeted more. This implies that a higher credibility article is more likely to be tweeted. This makes sense as better articles are more interesting and thus popular.

Log Odds Ratio for all interesting Words



The centre of distribution is at approximately -0.13 which shows that the mode of all words is more likely to be true. However is a large portion is centred around 0 showing the most words are not really indicative of being from real or fake news.

Top 15 real and fake words



Real words:
The top 15 most indicative words for real health articles shows that real articles often use more references such as ‘estimates’ to statistics, ‘happening’, ‘landmarks’ in current situations and rules such as ‘protocol’ and build up. One reason for this is that more credible news sources have citations, evidence and relate to the contemporary audience.

Fake words:
Most indicative fake words are about “AI”, “destroys” and UV. Linking this to my personal experiences, fake ads online tend to be clickbait on robots, UV rays, that are dangerous, and we should click on this ad immediately. Foxnews also seems to be the most popular word for fake news, which means the credibility of any article with Foxnews is extremely low.

Limitations of the dataset included missing values and vague notion of satisfactory levels. This could’ve been fixed by data imputations (finding the correct values for missing data) and more concise reviews. To improve our investigation, we can further analysis on the correlation between real and fake authors, words in criteria of reviews and expand our analysis on not just health articles.