6 Task 2 – Rumour Analysis

To figure out the characteristics of rumours and understand their differences with non-rumour tweets, the first step is using the best-performing model to predict the dataset. Then, subsequent work on pointing out the factors of rumours will be undertaken by analyzing the topics, hashtags and sentiments of source tweets and their replies.

6.1 Popular topics in rumour and non-rumour tweets

To find the popular topics discussed in both rumours and non-rumour tweets, the frequency of unigrams in source tweets are counted, and word clouds are made based on the frequency in figure1. It is indicated that ‘Trump’ appears in rumours with a significant frequency. Similarly, words about political issues, more precisely, election of America in 2021, like ‘President’, ‘Biden’ and ‘America’ show high frequency in rumours.

Frequency is not the only factor that shows relevance between political issues and rumours. The possibility of tweets to be rumours with certain word are also counted. To prevent the case that a word appears in rumours coincidently, only words appear more than 100 times are considered. The most suspicious words are listed in figure2, the result supports the viewpoint that rumours mention political topics more frequently. An interesting fact is that tweets about China and Florida show high relevance with rumours, it will be further discussed in next part.

To explore more information, the frequency of bigrams in rumours and non-rumour tweets are also considered. The most discussed bigrams except ‘covid 19’ are shown in figure3. Bigrams about politicians still show high frequency like Donald Trump, Joe Biden, Mike Pence, and Nancy Pelosi. On the contrary, bigrams about epidemic prevention are ‘safer’ from rumours. The medical and scientific topics are more objective and trustworthy about Covid 19.

6.2 What rumours say about the popular topics

Next, what the rumours say about the topics is analyzed with TF-IDF and Non-Negative Matrix Factorization (NMF). The TF-IDF will be constructed based on the given texts and the NMF can extract key words from TF-IDF matrix. This method can point out what typical rumours look like. The top 10 topics are shown in table1.

The result shows that vaccine is the most mentioned topic in rumours (Topic 1). The rumours also like to mention the epidemic prevention policies (Topic 2, 5). It is supposed that the rumours like to impose on public by the credibility of governments.

Still, Trump makes a great ‘contribution’ to rumours, so as Mike Pence (Topic 3, 4, 7) The rumours about Trump like to mention ‘lie’, ‘call’ and rally in Tulsa. This fact supports the conclusion that the rumours are about the election of America in 2021.

The word ‘hoax’ appears in Topic 2, 6 and words like ‘die’, ‘death’ in most topic, especially in topics about Trump, which shows that the rumours are mainly against the epidemic prevention policies.

The result of NMF also shows what rumours say about China and Florida. The words appear in rumours about China show that the antagonism against communist and vaccination is the reason for rumours about China (Topic 1, 9). The rumours like to cheat the public on the number of deaths in Florida, this kind of rumours may be caused by regional discrimination to Florida in the U.S. (Topic 10)

6.3 Popular hashtags of COVID-19 rumours and non-rumour tweets

The hashtags of both source tweets and their replies are extracted for rumours and non-rumour tweets. It is found that there is not much difference between source and their replies. This is because that the relies are likely to discuss the same topic with the source. As for differences between rumours and non-rumour tweets, the most obvious one is topic about Trump. Hashtags like ‘#Trump’, ‘#MAGA’ show significantly larger frequency in rumours considering the number of rumours is only one third of the number of non-rumour tweets.

However, hashtags like ‘#TrumpVirus’, ‘#TrumpLiesAmericansDie’ appear frequently in rumours too. This means that not only Trump’s supporters, but also his opponents are making rumours about him. So, it is hard to say that he leads to lie on the Twitter, it is more proper to say that he is a controversial figure. Except the difference about Trump, there is no considerable difference on hashtags between rumours and non-rumour tweets

6.4 Difference on sentiment of rumours and non-rumour tweets

VADER is introduced for the sentiment analysis which is a lexicon and rule-based sentiment analysis tool in NLTK library. VADER will provide the ratio of negative, positive, and neutral words in a corpus and give a sentiment score based on the ratio. The sentiment score ranges from -100% to 100%, larger it is, more positive the texts are, and 0 means that the text is neutral in sentiment. The sentiment scores of the tweets are shown in table2.

The result indicates that source tweets of rumours are more negative than non-rumour tweets, with a compound sentiment score of 0.14% vs. -12.27%. So as the replies to the source tweets. The replies to rumour are more negative than replies to non-rumour tweets with a more severe decrease on sentiment score from -7.77% to -36.49%. The rumours are more negative than non-rumour tweet probably because that the negative information is more inflammatory. Another potential explanation is that the purposes of rumours tend to be negative.

Another interesting fact observed is that the reply tweets contain far more less positive words than source tweets and shows lower compound sentiment score (0.14% vs. -7.77% for non-rumour tweets and -12.27% vs. -36.49% for rumours). This indicates a worth thinking appearance that people like to vent their negative emotions online with less right sense.