COMP90042 Project 2022: Rumour Detection and Analysis on Twitter

1068299, 1296306, 893164

1 Introduction

The importance of rumour detection is significant in the current age of digitization, especially in moderating in social networking sites, where majority of the population obtain their information. As rumours are easily created due to the individual's misinformation and interpretation of existing media, this emphasizes the need to remove them as the damage caused can cause a lasting impact.

This paper examines the various rumour prediction models that can be used to analyse tweets and label them with a binary classification, rumour and non-rumour. The prediction model would be further used to analyse and evaluate the nature of COVID-19 rumours and non-rumours, to provide deeper understanding and breakdown between the relation between tweet text and rumours.

2 Data Exploration

Using the data sets {train, dev, test}.txt files that include only the ids of the tweet and its replies, a script is created to extract ids of the source tweet, username that posted the tweet, whether the user is verified or not, source tweet replies, and the number of followers the user who posted the source tweet has. Since twitter has a limit of 900 queries per 15 minutes, a delay function is implemented as it will timeout multiple queries are requested. The extracted data is then saved as a comma-separated values file (csv) {train.data.txt, test.data.txt, dev.data.txt}.csv so that it could be easily loaded for later use.

Before creating and implementation of the rumour detection model, data exploration of the relation between the rumour label and the verified feature of the account, whether it is verified or non-verified, in both dev and train dataset to partially filter out some of the legitimate tweet sources.

Label	Rumour	Non-Rumour
Verified (Dev)	11.2%	28.0%
Non-Verified (Dev)	10.3%	50.5%
Verified (Train)	11.5%	27.4%
Non-Verified (Train)	8.8%	52.3%

Table 1 - Percentage of Verified accounts to Rumour

Label	Rumour	Non-Rumour
Verified (Dev)	60	150
Non-Verified (Dev)	55	270
Verified (Train)	180	429
Non-Verified (Train)	138	820

Table 2 - Count of Verified accounts to Rumour

From Table 1, there seems to be a form of relationship between rumour labelled tweets and verified feature of the Twitter account, with non-verified Twitter accounts spreading less rumours proportionally compared to the verified counterparts. However, this goes against the contrary sense that verified Twitter accounts would have less tendency to spread rumours as to get a twitter account verified, the user must submit identity verification to Twitter before getting verified status. Thus, verified users would not want to be flagged as a user spreading misinformation. Regardless, this could indicate that verification of accounts and rumour labels have a possible relationship with one another.

3 Task 1: Rumour Detection

3.1 Lemmatization

Lemmatization was used as the baseline as the benchmark for performance in comparison to other models such as stemmer. Stemming is avoided in this task as stemming causes a word to be reduced into its dictionary root, which the aggressive method would result in the word to lose its meaning and sentiment, which may affect the performance of sentiment analysis (Balakrishnan & Ethel, 2014).

The model was created by apply lemmatization on the text information of the source tweet, using the word tokenizer to split words into a list after pre-processing is done by removing other unnecessary characters such as links and punctuation. Once the text is normalised, Naïve Bayes and Logistic Regression classifiers were then implemented as seen Table 3.

3.2 TF-IDF

TF-IDF model was chosen to bring out the relevancy of a given word in a specific document in comparison to the inverse proportion of that word over the whole document corpus (Ramos, 2003). This would calculate how relevant is a given word in a specific document, thus providing the key topics of each document for analysis. The TF-IDF weights of the keywords in tweets would then provide information on whether the topics are related to rumours or non-rumours, providing the relevancy between words and the tweet rumour labels. This would allow the model to identify whether the tweet contains rumour related words or non-rumour related words.

As the implementation of the TF-IDF model focuses on keywords, as there are overlapping topics between rumour and non-rumour tweets, this would affect model evaluation. Depending on the classifier model, such as Logistic Regression used together with TF-IDF, resulted in varying results in comparison to Feed Forward Neural Network which provide a deeper relational analysis by extracting the feature vector at each interval from the tweet text and input into the neutral network (C P & Joseph, 2019), enabling the model to make a improved accuracy prediction.

3.3 BERT

BERT (Bidirectional Encoder Representations from Transformers) was chosen as it is a language model that can read text input bidirectionally, compared to other language model which could only read forward or backwards (González-Carvajal, 2020). This allows the BERT model to disregard the order of the process data. Tweets do not necessarily follow a specific sequenced language order (Kouloumpis, 2011), as everyone have their own personal lingo of expressing themselves which is a combination of both spoken and written language. Furthermore, users are required to express themselves in a tweet within the word limit of 140 characters. Thus, the BERT model would be an appropriate model, as having a bi-directional learning model would be optimized in for a masked language model (Devlin et al., 2019).

BERT pretrained model, “bert-base-uncased” was used in this report. Uncased is used as to negate the difference between uppercase letters with lowercase, as tweets are not necessarily case-sensitive.

Different Adam optimizer for learning rates were chosen in reference to other “bert-base-uncased” sentiment analysis used in other train cases base on their F1 and Recall score (Geetha & Karthika, 2021). Learning rate of 2e-5 and 3e-5 were tested to find a desirable learning rate with optimal weights for the BERT model. This would prevent the model from the process being stuck due to small learning rate and insufficient epochs, whereas a high learning rate would result in a suboptimal solution model (Huo & Iwaihara, 2020).

3.4 Rumour detection model analysis

The analysis of the Rumour detection models was tested against two data sets, dev data and 40% of the test set made available, with the classification report of the precision, recall and F1 score presented in Table 3.

3.4.1 Lemmatization: Logistic Regression & Naïve Bayes

The Lemmatization of Naïve Bayes have a overall low performance in the classification report for dev set, but however unexpectedly performed reasonably well for F1 test score at 0.8081, close to its precision score of 0.7920. Although with a high precision, the low recall of 0.7122 indicates a low quality of correct predictions. Lemmatization of Logistic Regression overall performance is higher than Naïve Bayes, but both similarly have a low F1 score performance.

3.4.2 TF-IDF with Logistic Regression

The performance of Logistic Regression worked reasonably well in with the dev set, with a F1 score of 0.8693, but however worked poorly on the test(40%) set, with a F1 score of 0.5915. This indicates that the classifier is not suitable for the regulation parameters we have set for TF-IDF.

Rumour Detection Model	Dev: Precision	Dev: Recall	Dev: F1	Test(40%): F1
Lemmatization: Naïve Bayes	0.7920	0.7122	0.7500	0.8081
Lemmatization: Logistic Regression	0.9474	0.6475	0.7692	0.8667
TF-IDF: Logistic Regression	0.9004	0.8860	0.8693	0.5915
TF-IDF: Feed Foward Neural Network	0.9474	0.9477	0.9475	0.8936
TF-IDF: Feed Foward Neural Network with Verified feature	0.9000	0.8609	0.8800	0.9072
Bert with Learning Rate 2e-5	0.9411	0.9402	0.9406	0.8571
Bert with Learning Rate 3e-5	0.9529	0.9533	0.9530	0.8889
Bert 3e-5 transfer learning to Feed Forward Neural Network	0.9385	0.9346	0.9358	0.8824
Bert 3e-5 transfer learning to LSTM combine model with Verified feature	0.9415	0.9421	0.9403	0.7907

Final Models used	Test(40%): F1	Test: F1
TF-IDF: Feed Foward Neural Network with Verified feature	0.9072	0.8481
Bert with Learning Rate 3e-5	0.8889	0.8846

Table 3 – Rumour Detection Model Classification Report

3.4.3 TF-IDF with Feed Forward Neural Network

In comparison with the TD-IDF Logistic Regression model, feed forward neural network performed well, with a dev set F1 score of 0.9475, and a test(40%) score of 0.8936. This is mainly due to the neural network being able to create connections between the different inputs produced by the TF-IDF vectorizer, providing adjustments to the weights that would best fit the model.

3.4.4 TF-IDF with Feed Forward Neural Network and Verified feature

The model mentioned in 3.3.3 is further enhanced by incorporating Verified feature, which was shown to have a relationship with the rumour labels. This however had the adverse effect on the dev set, with a decrease of performance in the F1 score by 7.12% compared with the previous model. However, there is a slight increase performance in the test(40%) set, indicating that the feature does not have a strong relationship with the rumour label.

3.4.5 BERT: Learning Rate 2e-5

The F1 score for the dev set has a high performance at 0.9406, due to tweaking the hyperparameters with the dev set. The model has a lower F1 score with test(40%) set at 0.8571, as the BERT learning model is overfitting with the train and dev set, thus have a lower performance in test.

3.4.6 BERT: Learning Rate 3e-5

Due to the small size of the train and dev set provided, a higher learning rate is implemented to prevent the model process from being stuck. This resulted in an overall increase in the BERT model performance in the dev set, with dev set F1 score at 0.9530 and a test(40%) set F1 score at 0.8889.

3.4.7 BERT: Learning Rate 3e-5 with transfer learning to Feed Forward Neural Network

The combination of BERT model with feed forward neural network although have a small increase in overall performance in dev set, the test(40%) F1 score is marginally lower when compared with the BERT model itself, due to slight overfitting to the dev set.

3.4.8 Learning Rate 3e-5 with LSTM model with verified feature

Implementation of a LSTM neural network with the BERT model faired with an overall drop in performance for F1 score in both dev and test(40%) set in comparison to the feed forward neural network, with a test(40%) set F1 score of 0.7907, which shows that the model overfitting to the train and dev set.

3.4.9 Evaluation of Final Models

Two models were chosen for final evaluation, based on two of the best performing models, Bert and TF-IDF. Bert with a learning rate of 3e-5 got a F1 score of 0.8846 performed similarly to the test(40%) set at 0.8889. Although the TF-IDF model performed slightly lower in the test set, it could be likely due to the model having overfitting prediction model from the hyperparameters tweaked on the dev set.

4 Task 2: Data Analysis

To be able to analyse the COVID-19 tweets, collection of tweets was achieved by using 6

Twitter Developer API token interchangeably in turn to maximise efficiency before reaching the rate limit. For each COVID-19 tweet, information on its full text, verified feature, tweet metric, user metric, created date, and context annotation, were retrieved. Context annotation provided by Twitter that accurately extract topic from tweets was also retrieved and used for rumour topic trends analysis.

4.1 Characteristics and Topic Relation

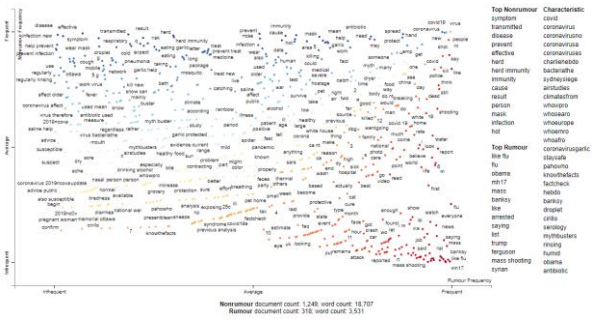

Figure 1 – Rumour and Non-Rumour Characteristic Distribution (Sukphasuth, 2022)

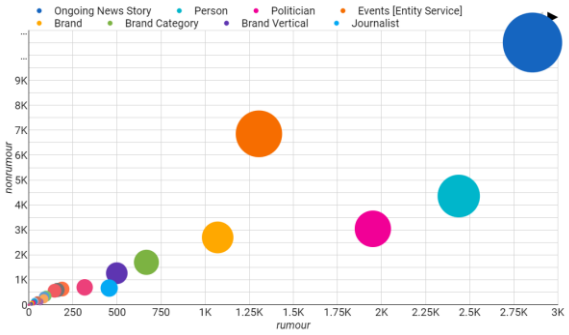

Figure 2 – Rumour and Non-Rumour Topic Distribution

From breaking down the difference in topics in relation to their rumour labels, as seen from Figure 2, there a positive correlation between the topics of tweets in both rumour and non-rumour tweets, with similar topics, as topics were popular during the period.

However, by looking into each characteristics, as seen from Figure 1, there are different characteristics of tweets between non-rumour and rumour, with top non-rumour characteristics such as "symptom", "transmitted", whereas top rumour characteristics are "like flu", "flu", "Obama", "mh17". This indicates that although there are similar characteristics, difference in characteristics frequency shows a relationship with the tweet's rumour labels.

4.2 Rumour Topic Trends

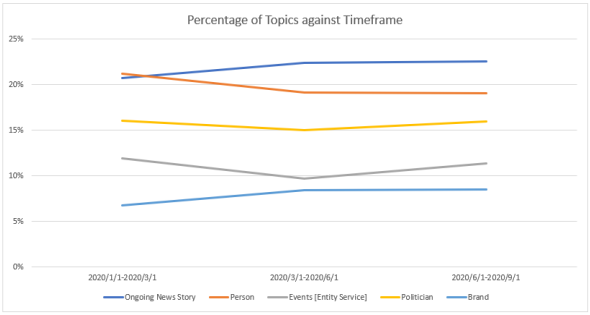

Figure 3 – Percentage of Rumour Topics against Timeframe

Topic trends of rumour labelled tweets were collected and categorised over 3 consecutive periods of timeframe. In Figure 3, the top 5 rumour topics with the highest count remained the same throughout the entire period of 9 months, with varying fluctuations. Fluctuations are caused due to one-time events, such as the elections, which can be seen from the initial trend at the start of 2020, when elections have not yet settled down, thus trends in "Person" and "Politician" during the timeframe was popular but it decreased overtime.

4.3 Popular Hashtags

Rumour Tweets	Count	Non-Rumour Tweets	Count
#covid19	305	#covid19	2062
#coronavirus	145	#coronavirus	951
#breaking	15	#breaking	82
#trump	11	#stayhome	39
#coronaviruspandemic	6	#china	38
#maga	6	#coronaviruspandemic	32
#cdnpoli	5	#covid	31
#blacklivesmatter	5	#cdnpoli	30
#stayathome	5	#lockdown	29
#foxnews	5	#stayhomesavelives	24

Figure 4 – Hashtags of Rumour and Non-Rumour Tweets

Hashtags were obtained by collecting tokens from the text of tweets that begin with '#' followed by the hashtag information. From Figure 4, both rumour and non-rumours tweets have common hashtags, with rumour tweets containing "#trump", "#maga", "foxnews", while non-rumour tweets hashtags focus on Covid-19 topic matters, which shows that the classifier correctly classifying rumour label in tweets.

4.4 Sentimental Analysis

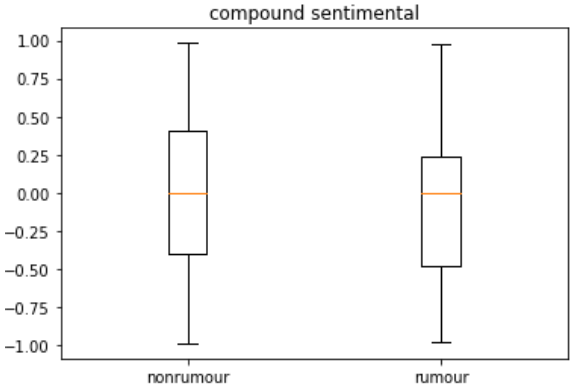

Figure 5 – Compound Sentimental Distribution

NLTK VADER Sentiment Analysis was implemented to obtain the compound sentimental values of the tweets' sentiments, into two categories depending on their rumour label, non-rumour and rumour.

The sentiment analysis also incorporated not only text of tweets, but also emojis. By adding sentimental evaluation on emojis, this would further enhance the sentiment of the text by adding a polarity(Guibon et al., 2016), which help classify neutral texts.

As seen from Figure 5, non-rumour tweets have a balanced sentiment distribution, with both positive and negative sections even in size. However, when comparing between non-rumour and rumour tweets, rumour tweets are slightly negative in sentiment values, indicating that rumours have a negative emotional connotation.

4.5 Characteristics of users

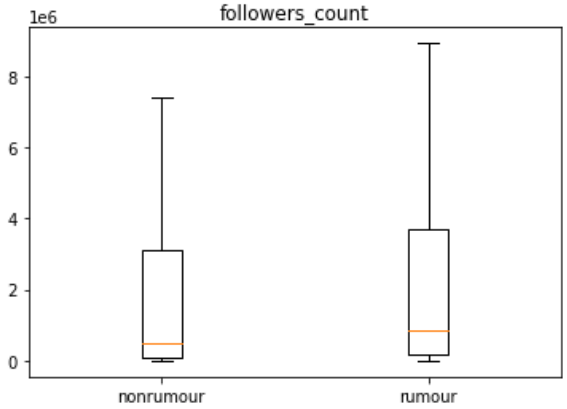

Figure 6 – Followers Count Distribution

Follower counts of users of the main tweets were obtain and plotted against whether the tweet was labelled as rumour or non-rumour. As seen from Figure 6, users who posted rumour tweets have an overall higher follower count in each quartile, however, no clear relationship can be distinguished between users who tweet rumours and non-rumours.

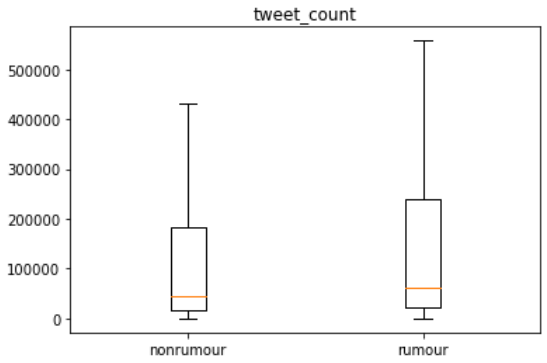

Figure 7 – Tweet Count Distribution

From Figure 7, users of main tweets that were labelled as rumour have an overall slightly higher tweet count in compared to their counterpart, showing that they tend to have a higher tweet activity interaction. This is clear above the upper quartile, with a higher tweet count ceiling in users of tweets labelled as rumour than non-rumour.

5 Conclusion and Future Work

Currently, due to the limitation of Twitter API tweet extraction rate and time restriction, not all replies of tweet and was unable to fully utilized and extract all the information provided from Twitter.

From our part 2 analysis, we found that the rumour a mostly political related, tend to stick out during the political event time frame such as election, and relatively gives a more negative sentimental score. We also found that account with high following count that have been verified are more likely to produce rumour. This led to a conclusion that the rumour on twitter is largely served the purpose to spread fear to aid the election campaigns for a specific political body by trusted account.

Furthermore, the model is trained on the real tweet send by human on twitter.com and it have been polluted with racist and hate speeches. In future work, additional improvements on our part 1 classifier model to be able to detect bias from sentiment values can be implemented to improve our model to an even higher ethical standard.

References

Balakrishnan, V., & Ethel, L. Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, *2*(3), 262–267. https://doi.org/10.7763/lnse.2014.v2.134

C P, P., & Joseph, S. (2019). Deep Learning Approach For Rumour Detection In Twitter: A Comparative Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3437620

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*. https://doi.org/10.18653/v1/n19-1423

Geetha, M., & Karthika R.D. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *International Journal of Intelligent Networks*, *2*, 64–69. https://doi.org/10.1016/j.ijin.2021.06.005

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

Guibon, G., Ochs, M., & Bellot, P. (2016). From emojis to sentiment analysis. In WACAI 2016.

Huo, H., & Iwaihara, M. (2020). Utilizing BERT Pretrained Models with Various Fine-Tune Methods for Subjectivity Detection. *Web and Big Data*, 270–284. https://doi.org/10.1007/978-3-030-60290-1_21

Kouloumpis, E. (2011). *Twitter Sentiment Analysis: The Good the Bad and the OMG! | Proceedings of the International AAAI Conference on Web and Social Media*. Aa. https://ojs.aaai.org/index.php/ICWSM/article/view/14185

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *In Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29-48)

Sukphasuth, L., (2022). Rumour and Non-Rumour Characteristic Distribution. Kan. https://rdat.herokuapp.com/