Rumour Detection and Analysis on COVID-19 Twitter

Student ID: 1256470, 1097419

COMP90042, Semester 2, 2022

1 Introduction

As social media has been more and more widely used for news gathering, we are inevitably in an era of information explosion that leads to a higher probability of receiving fake news. This is the unmoderated nature that forces us to develop the capability for rumour identification, so that we could have an appropriate judgement on information correctness. To accelerate the validation process and reduce the time cost, we are motivated to devise a sensible way as the automatic solution. Many great machine learning techniques have provided us with strong support dealing with the challenging task. In this report, we are going to elaborate the methods to automatically differentiate rumour and non-rumour tweets, and conduct an analysis for COVID-19 tweets based on the results from our rumour classifier.

2 Dataset

The experiment has been provided with multiple datasets for different purposes, which includes the training, development, test, and COVID-19 tweet ID datasets. They are all in the format of one source tweet ID followed by its reply tweet IDs to be used for crawling raw tweet objects through the Twitter API. The training dataset and development dataset are also labelled with "rumour" and "non-rumour" for training model and validation purposes respectively. The rumour tweets in the training dataset are dominant with approximately 78% and the development dataset has the same rumour distribution.

3 Environment Setup

All of our detection systems were constructed on Jupyter Notebook with Python programming language. The libraries we used are NLTK mainly for natural language processing, and Pytorch, transformers for BERT-based model building. The experiment was finally performed on Google Colab with GPU.

4 Task 1 Rumour Detection

The objective of this task is to develop a binary classifier which can reliably predict whether a given tweet is rumour or not.

4.1 Model Selection

The primary classification model we have implemented was BERT-based model for contextual representation of each word in a tweet text, on top of which there is another sequential neural network layer taking the word embedding results to compute the final class (0 for "nonrumour", 1 for "rumour").

There are two versions of BERT-based model as our choices, the pure BERT and its variant RoBERTa which was initially published by (Liu et al., 2019). Both models are pre-trained and can capture the deep dependencies between words to learn contextual representation of a sequence, whereas RoBERTa uses dynamic masking, rather than static masking in BERT, to improve the downstream task performance (Durgia, 2021).Therefore, we selected this method as an alternative to observe and adjust our model performance.

4.2 Pre-trained BERT-based model options

We have used a variety of pre-trained models available on HuggingFace model library [1].

- *bert-base-uncased*
- *bert-large-uncased*
- *cardiffnlp/twitter-roberta-base*
- *cardiffnlp/twitter-roberta-base-2021-124m*

The difference between the first two pure BERT models is: *bert-base-uncased* has 12 layers, 768 hidden dimension, 12 attention heads, 110M parameters. *bert-large-uncased* has 24 layers,

[1] https://huggingface.co/models

1024 hidden dimension, 16 attention heads, 336M parameters.

The difference between the last two RoBERTa models is: *cardiffnlp/twitter-roberta-base* was trained on 58M tweets. *cardiffnlp/twitter-roberta-base-2021-124m* was trained on 124M tweets.

4.3 Text Preprocessing

The only feature we were using is the tweet text from the crawled data. To make the feature more fit to the model, we have implemented some text preprocessing.

For each of the tweet events in the dataset, we separated the source tweet from its replies so that we could have a source input and its replies as a pair input for the BERT-based word tokenization and embedding.

4.3.1 Rumour Tree

To better implement the separation, we built a propagation tree structure for each source tweet. The referenced tweets in each raw tweet object yielded from Twitter API are the source of our reply tree. The root node is the source tweet object. If a new tweet replies to a former tweet, we let the new tweet to be the child node of the former tweet. Thus, we have a rumour propagation representation from the source tweet to latest replies.

Then we extracted the propagation text sequence using the Depth First Search algorithm. A path can be found from the root node to each leaf node. Because people posted tweets sequentially and only replied to one tweet at a time, no loop is expected in this tree structure. We finally used each path as an input text sequence to feed into the word embedding layer in our model.

The benefits of this method involve:

• It intuitively reflects a better logic between sentences

• Data enrichment has been achieved by breaking one rumour to several rumour propagation sequence

4.3.2 Time Series

We also proposed another method to prepare the source-reply pair as backup, in which we sorted each tweet in a propagation event by the creation time in an ascending order so that the first one is the source tweet and the rest are its replies in a chronological manner.

	Rumour Tree	Time Series
PureBERT(base)	92.17	94.57
PureBERT(large)	92.24	93.88
RoBERTa(58M)	93.69	94.17
RoBERTa(124M)	94.62	95.04

Table 1: Development Performance

	Rumour Tree (Public)	Rumour Tree (Private)	Time Series (Public)	Time Series (Private)
PureBERT(base)	76.00	77.46	76.19	82.99
PureBERT(large)	82.47	84.21	83.17	82.93
RoBERTa(58M)	85.42	83.23	89.80	85.00
RoBERTa(124M)	88.24	84.06	**90.00**	**83.44**

Table 2: F1-score on test dataset in %

4.3.3 Other preprocessing steps

For each tweet source-reply pair, we have further applied the following preprocessing.

1. Convert each word to lower case
2. Replace user mentions and URLs by "@user" and "@http" respectively
3. Remove the hashtag symbol ("") while using MaxMatch algorithm to tokenize it

4.4 Results

We subsequently fine-tuned our model by using different text preprocessing with AdamW optimizer using learning rate of $1e^{-5}$ for batch size of 16. We also set 2 warm up epochs at the beginning of training and weight decay to $1e^{-2}$. The results are shown in Table 1 and Table 2.

From both development and final evaluation scores, it is arguable that an increasing number of layers indeed contributes to the model performance improvement (PureBERT(base) vs. PureBERT(large)). The dynamic masking mechanism in the RoBERTa model helps to represesnt the context more comprehensively by learning different parts of the sentence and truely enhances the downstream classifier performance. In addition, with learning on larger tweet size, RoBERTa(124M) captures more semantic meaning so achieved a better result compared to RoBERTa(58M), and it has the best result

among all experiments with Time Series text preprocessing.

It is not difficult to find out that models with Time Series preprocessing outperformed those with Rumour Tree. The reason behind may be:

• There are missing tweets that break down trees in many cases. If one of the key replies is missing, all of its children can not have a path to the source tweet.

• Model may overfit to the source tweet and early replies, because their representation can be found more frequently than other tweets, as each input text sequence of a rumour at least contains the source tweet.

• Time Series maintained the integrity of all tweets in an event. Even though there are missing tweets, they are skipped to the next closest reply so the path is still complete. This is due to the missing tweets are not coming from the crawled data instead crawling stage.

• Time Series better reflects the sequential relationship between source tweet and replies, which may be more suitable for the mechanism of downstream neutral network.

We also have observed that our model entirely under performed on test dataset. The possible reason may be there are numerous tweets not written in English, from which the sentence information could not be captured by the model. The best submission in the private leaderboard resulted in a lower performance. This is firstly because this particular model may be overfitting to noise in the public dataset where we simply discard development performances. Besides, we only focused on the development F1 score to evaluate our model, the improvement could be plotting the learning curve and check for development loss. Lastly, we performed ensemble learning on our output label using a majority voting mechanism, this prediction has the highest private score at 87.50%, which proves that our model overfits to the public leaderboard to some extent so that single model output was unstable.

5 Task 2 Rumour Analysis

The objective of this task is to use the previously built rumour classifier to predict whether the given COVID-19 tweets are rumours or not. We have selected our best ensemble learning classifier to implement the task even though it was not selected for the submission. Within over 200K COVID-19 tweets in our dataset, we have investigated that approximately 94% of the source tweets are non-rumours. Subsequently we have explored the characteristics of those tweets with respect to their topics, hashtags, trend, and sentiment.

5.1 Topics discussed in rumours and non-rumours

Table 3 has given us the most common words and bigrams discussed in rumours and non-rumours. We can observe that rumours are predominantly talking about the consequence and precaution of COVID-19, such as the death, cases, and wearing mask. Politics has been heatedly discussed as well involving President Trump and White House etc. However, the discussion focus in non-rumours is surprisingly similar to rumours with the only difference that political topic is shifted to United States and China.

Another common topic mentioned by Twitter netizens is name. In both rumours and non-rumours, Donald Trump has been discussed at most, followed by other US presidents, politicians, and those of many other countries, such as Joe Biden, Obama, Pence, and Putin. As we all know President Trump is a Twitter fanatic and prone to publish opinions involving global issues so the topics are easily related to him.

5.2 Hashtags in rumours and non-rumours

Plenty of overlapped hashtags has been discovered in rumours and non-rumour tweets. For example, covid19, coronavirus, and pandemic appear in both cases. These words co-share the characteristic of a COVID-19 topic, which is ambiguous to identify the rumour from non-rumours. However, non-overlapped hashtags in each classified side show a stronger directivity. What we can see in the rumour hashtags are america, dumptrump, democrats etc.,, whereas the exclusive hashtag in non-rumours are chinavirus, healthcareheroes, lockdown, contest etc.

5.3 Trend Over Time

During the 2020, the COVID-19 rumours were exploded at the beginning of the year when the pandemic was initially outbroke. It shrank down in the following months with more and more countries

	rumour topics	non-rumour topics
1	people die, covid health, coronavirus case	covid death, people die, die covid
2	donald trump, american die	pandemic, test positive
3	wear mask, get covid	covid pandemic, test
4	white house, president trump	donald trump, China
5	stay home, nurse home	social distance

Table 3: Topics in rumour and non-rumour tweets

over the world successively entered lockdown with increasing confirmed cases and death rate. Many countries such as Japan and Germany were impacted on their economic recessions as the evidence to refute the rumours. The non-rumours were dominantly spread over the Twitter until a new infection caused a sharp rise in America in the late June, whereas afterwards rumours revived to a higher proportion and reached the highest in July. The trend reasonably coincides with the COVID-19 timeline published by The New York Times (Taylor, 2021).

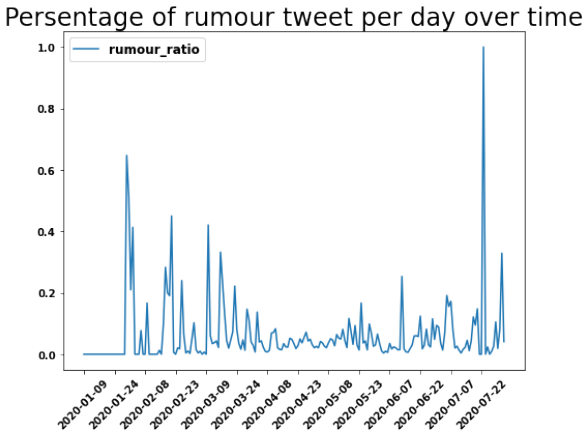

Figure 1: percentage of rumours over time

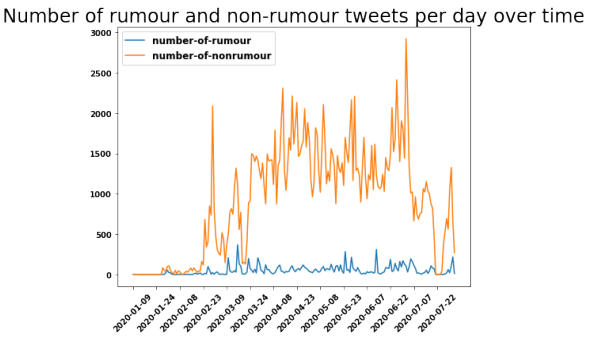

Figure 2: rumours and non-rumours change over time

5.4 Sentimental Analysis

The predicted results present the negative sentiment in both rumours and non-rumour events, which indicates the pandemic is primarily related to negative emotional expression. In terms of the proportion of each sentiment in rumours and non-rumours, negative tweets exist more in rumours compared to non-rumours (55.42% vs. 49.99%), and positive tweets are taking less part (27.29% vs. 32.36%). The ratio of neutral tweets in both cases are similar (around 17%). This is rational since negative sentiments are usually displayed more in rumours and non-rumour tweets are generally more positive.

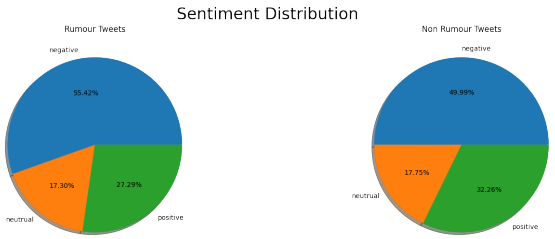

Figure 3: sentiment distribution

6 Conclusion

Overall, in this report we have described two rumour classification models making use of BERT architecture, as well as two major text preprocessing methods we used to adjust our model performance. Our best performed ensemble learning classifier has achieved 87.50% F1-score in the final evaluation, which has been applied to the COVID-19 tweets for the prediction and helps us to understand the nature of rumours and non-rumours propagated on Twitter.

References

Chandan Durgia. 2021. Chandan Durgia exploring bert variants (part 1): Albert, roberta, electra.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Derrick Bryson Taylor. 2021. Derrick Bryson Taylor a timeline of the coronavirus pandemic.

A Team contributions

A.1 Student 1256470

Brainstorming the methods and models in Task 1
Brainstorming the analysis aspects in Task 2
Building the code architecture and model details, and implementing experiments
Providing charts and tables support in report

A.2 Student 1097419

Brainstorming the methods and models in Task 1
Brainstorming the analysis aspects in Task 2
Providing code support and methodology research
Writing report