Rumour Detection and Analysis

1183945 and **1227216** and **1212462**

Abstract

Information spreads quickly, specially through social networking platforms, including Twitter. Bidirectional Encoder Representation from Transformers(BERT) can be implemented to build a detection system to classify rumours or non-rumours. This report presents comparison of performance of different BERT-related models based on given tweets labeled with rumour or non-rumour, proposes our trained model to examine COVID-19 tweets, and analyzes the differences in these COVID-19 tweets. In our experiments, the BERT using TensorFlow framework with transformers TFBertModel, which is available on the HuggingFace Model Hub[1], outperforms other models, especially being fed with the transformed datasets, in the development datasets, the evaluation of F1 score can achieve to 94.22%.

1 Introduction

Misinformation increases rapidly during the COVID-19 pandemic (Shahi et al., 2021), detecting false information has drawn significant attention from researchers (Radu, 2020), mitigating consequences that rumours might have brought has been needed. This project aims to develop a rumour detection system using machine learning models to categorize rumour and non-rumour, and apply our fine-tuned best binary model to label tweets on COVID-19 tweets for further investigation.

2 Data Processing

2.1 The "text" extraction

Sets of tweet IDs were provided including source tweets and their replies as the training-dataset, dev-dataset and test-dataset, the Twitter API were used to crawl the given tweet IDs to retrieve their objects. A tweet object has items including "created_at", "id_str", "text", "user" and etc, each tweet was crawled and stored separately in json format at first. The source and the replies were concatenated together to form one event, among all the items, "text" was agreed as a primarily important indicator to detect rumours, as thus, we extracted source tweet ID for each event and all the "text"s event by event for further implementation.

Table 1: The Number of Tweets Overview

Dataset	Source Tweets	Retweets	Rumour	Non-Rumour	Total Tweets
Train	1895	28524	420	1475	30419
Dev	632	9922	139	493	10554
Actual-Train	1807	18507	411	1396	20314
Actual-Dev	595	6647	137	458	7242
Test	558	7438	N/A	N/A	7996

2.2 Initial Processing

After being concatenated, We first converted it to the tsv format for the baseline experiment, the new datasets in tsv format were created. At the initial stage of processing, in view of BERT's defaulted function of tokenisation using WordPiece, it can tokenize sentence. It did not need to do much preprocessing, even hyperlinks and "@user" were not removed from the input data, we just directly fed it to BERT to observe how it can perform.

Also, We converted these tsv format datasets into csv files just for the convenient of any further implementation.

2.3 Further Processing

We noted there were some http links and "@user" in some tweets, although the BertTokenizer had no problem to handle it, these elements were still part of the max length, it should affect the overall performance to a certain degree, in addition, they should not enhance much contextual interpretability, thus, in order to efficiently use the max length, these url links, websites, "@user", emojis and some irregular expressions were shortened or replaced

[1] HuggingFace Model Hub https://huggingface.co/models

with their normal forms. New datasets named denormalised_csv were created for fine-tuning.

2.4 Long Sentence Cutting/Data-Transform

The concatenated "text"s for a single event generally longer than 512 tokens, simply cutting the whole concatenated texts to the MAX-length that each model requires should affect the prediction performance. We were inspired by the long sentence cutting approach (Fiok et al., 2021). As such, we arranged the source tweet to concatenate with each retweet separately, and created new datasets named train_sr, dev_sr and testsr. As each source tweets have a couple of retweets, the matrices of the newly built datasets compared to their original form far exceeded, such as, the size of the training dataset increased to 22685 from 1895.

3 Task 1, Models Selection and Performance

3.1 Baseline-BERT, using Pytorch Framework

The BERT model (Devlin et al., 2018) was selected first to implement as the baseline model, As we practiced it in the subject COMP90042 Natural Language Processing. The pre-trained BERT model is designed to use self-attention networks to capture correlations between words, it can be fine-tuned with only one extra output layer to build a model to learn contextual representations, more importantly, it can precede to learn contexts without specifically task-related structure modifications. BERT is conceptually simple and empirically powerful. It has been broadly applied in natural language processing tasks.

We used the pytorch framework, and huggingface's transformers library, a suite of transformer models with a consistent interface. We first tried the max-length 300, any shorter length were padded to the MAX length, and longer length were cut down, attention mask vectors were established to direct BERT to attend real words tokens rather than the "PAD"s. Segments IDS were created to distinct sentences, then tokens were converted to vocab IDs, turned these IDs into torch tensors and fed these sequences to BERT. The hidden_reps was defaulted to conduct the contextualised embeddings for all tokens and passed it to a classifier, it had dimension[1, max-length, 768], the output layer produced probability. We used binary cross-entropy to define loss function and Adam to optimize our model, and saved the model which improved the performance.

The initial batch size was set to 64 , we achieved an F1 score of 81% in the test, then we fine-tuned the MAX-length to 512, wcich is the max-length of BERT model, batch size was down to 8 and learning rate Adam to $2e^{-5}$. the F1 score also improved to 82.978%.

3.2 RNN

An RNN is derived from feed-forward neural network that can be implemented to process variable length sequential sentence (Ma et al., 2016) by applying a recurrence formula. It uses a state vector to represent inputs, created an unrolled computation graph and apply back-propagation algorithm to calculate gradients. However, there are still some issues along with this method, it often mismatches training and decoding, and is not able to learn long-distance dependencies due to the vanishing gradients, early inputs will get much less updates. We tried this approach, the feedback as expected was even below our baseline model, the F1-score in the test on the Kaggle was 75.294%.The Long Short-Term Memory (LSTM) can solve the vanishing problem by creating memory units to store information over time. It could improve the RNN performance at certain degree.

3.3 BERTweet - using Pytorch Framework

BERTweet was published by (Nguyen et al., 2020). It was trained based on the RoBERTa pre-training procedure (Liu et al., 2019) using a large corpus of 850M English Tweets. The characteristics of Tweets are substantially different from traditional written English (Eisenstein, 2013), they have tendency toward a shortened format, an informal grammar structure, irregular vocabulary, some typo errors, emoji, and hashtags. Therefore, BERTweet model("vinai/bertweet-base") pre-trained on a larger corpus than its predecessor BERT was expected to have a better performance in our experiments. On the surface, it should suitably implement with tweets.

We adopted the pytorch framework same as the BERT base, and for the purpose of fine-tuning, we set up the batch size to 8, the model for 7 epochs using learning rate for AdamW optimizer $5e^{-5}$, an epsilon of $1e^{-8}$ and weight_decay of $1e^{-2}$.

The results increased about 6% compared to our baseline model on the development dataset. The feedback of BERTweet model based on the large corpus implemented to detect rumours on tweets apparently outperformed. however, its performance was at certain degree restrained to the max length of 128, the length of an event was much longer than it. It prevented our model mounting to a much higher accuracy rate.

To solve this limitation on max-length, we used datasets we processed according to the long sentence cutting approach, and adopted the max-voting method to decide the final classification. The performance turned out to be better than directly deploying the Max-Length. The transformed datasets encouraged us to explore further.

3.4 BERT - using TensorFlow Framework, with a dropout rate

To improve our rumour detection performance, we considered using TensorFlow framework to add a fully connected hidden layer after BERT layer with a dropout rate=0.3 to mitigate the potential overfitting problem. In the TFBertModel, a fully connected layer was added to the intermediate layer and the output layer, so as to map the trained distributed feature representation to the sample label space. The function relu and sigmoid were used as activation functions for the fully connected layers. A higher accuracy than all the models using pytorch framework was returned.

As inspired by the improvement using the cutting long sentence approach in the prior BERTweet model, we employed these transformed datasets, the result was superior to all other models we previously experimented.

4 Results and Discussion

The results of our models were evaluated on the development set shown in Table 2. Since the performance of the RNN model did not even remarkably exceed the performance of the baseline (BERT base), we decided not to improve the RNN model. The performance of the BERTweets model and BERT_TL_Dropout was significantly better than the baseline model's performance in the development dataset and the final results of the Kaggle private board. What's more, using the long sentence cutting approach, the performance of the BERTweets_transform model and BERTweets_TL_transform model outperformed than the result of models without using the long sentence cutting approach.

Comparing the performance in the development dataset to the test on the Kaggle private board, the F1-score in the test was slightly lower. The two datasets were probably not split evenly, it caused the model overfitted to the develop data set. Since Kaggle selected the best two public submissions from the public board for the private, it selected the results of the BERTweets_TL model for the private board instead of the results BERTweets_transform model and BERTweets_TL_transform model, which performed much better than the ERTweets_TL model in private board.

Table 2: Evaluation Score in %

	Performance					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC-Score**	**Test-F1**
BERT	84.87	80.52	45.26	74.36	70.99	84.34
BERTweet	93.28	83.92	87.59	90.66	91.29	86.25
BERTweet_transform	95.53	95.05	83.48	88.89	92.64	89.94
BERT_TL_Dropout	93.78	86.23	86.86	91.25	91.36	85.23
BERT_TL_Dropout_transform	97.58	96.36	92.17	94.22	99.31	90.36
RNN	76.97	75.34	79.61	77.42	82.73	N/A

5 Future Work

There are some space to optimize models, such as using translation technique to handle some non-English tweets. There are some superior models performing better than our models, take BERTweet for example (Nguyen et al., 2020) released their advanced version "vinai/bertweet-large" with a large set of parameters 355M using 873M English [2] in August 2021, and researchers (Kim et al., 2022) published the improvement in late of April 2022. It can be used in the future work.

6 Task 2, Analysis on COVID-19 Tweets

In this section, a prediction was made using the best model we previously built to label rumours or non-rumours. There were 35895 predicted to rumour, and 165595 labelled to non-rumour using long sentence cutting. We examined the prediction, investigated the difference of features between the two labels with respect to their hashtags, trends evolving over time, and topics.

6.1 Hashtags

Hashtag is an important feature of the tweet and facilitates a search for a specific topic of interest (Antoine, 2016). Users make use of hashtags to categorize their posts (Godin et al., 2013). To examine what hashtags were prevalent on both sides, we

[2] VinAIResearch/BERTweet `https://github.com/VinAIResearch/BERTweet`

extracted them and summarized their frequencies for each label. We observed that a large number of hashtags overlapped between two groups, such as #covid19, #coronavirus, #coronaviruspandemic, #covid, #pandemic, #corona, #covid19pandemic, #trumpliespeopledie. These tags demonstrated the common interests shared in the two groups.

Apart from these dominant hashtags, there were some distinct hashtags as shown in the Figure 1.

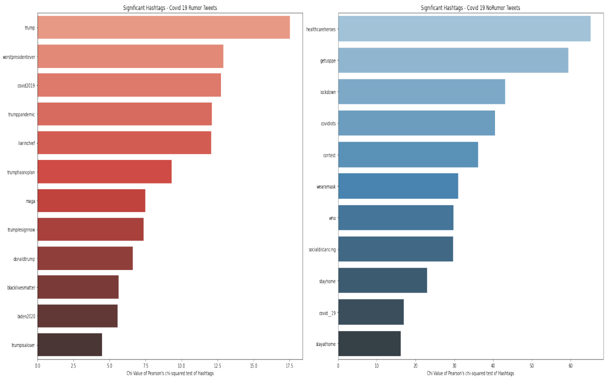

Figure 1: Typical Hashtags distribution for labels

From the above graph, we observed that the typical hashtags in the rumour generally focused on #liarinchief, #trumpisloser, #fakenews, #trumppandemic, #chinesevirus, #worstpresidentever, #chinavirus, as we know these words were widely regarded as rumours. Whereas, hashtags in the non-rumours mainly concentrated on #stayhome, #socialdistancing, #waremask, #getupppe, #healthcareheroes, #contest, #who, and #lockdown, these words reflected that their users laid emphasis on facts rather than spreading virtual things, and had more positive impact. In a summary, these hashtags particularly showed us each side's popular interests and their views on COVID-19 during a specific time period.

χ^2 test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies of a hashtag in rumor and norumor tweets. For a hashtag h, the χ^2 value of hashtag $\#h$ can be calculated as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, i \in \{rumor, non{-}rumor\}$$

When degrees of freedom $df = 1$ and $p = 0.05$, in one-way test scenario, the χ^2 value is 3.84. When χ^2 is greater than 3.84, #h is a distinct hashtag.

6.2 Trends Evolving Over Time

Tweets of the COVID-19 dataset were posted or retweeted over the period between 09/Jan/2020 and 01/Aug/2020, it was a special historical period of the pandemic outbreak, individuals were restricted on movement or were ordered to stay at home (Jacobsen and Jacobsen, 2020). This subsection will explore trends changing over the rough eight months.

We divided the dataset into four sub-datasets roughly two months for each. A handful dominated hashtags being both sides were removed so that other major hashtags were clearly demonstrated in our graphs. The summary was presented in the Table 3.

Table 3: COVID-19 Tweets Summary

Period	Total	Rumour	Non-Rumour
09/Jan/20-10/Mar/20	11184	4986	6198
10/Mar/20-10/May/20	84285	13604	70681
10/May/20-25/Jun/20	66502	11123	55396
25/Jun/20-02/Aug/20	39518	6182	33336

In the first stage, users were more concerned about #maga, #hoax, #fakenews, #trump, #trumpliesaboutcoronavirus and #trumpvirus in the rumour side, while the counterparts were more interested in #wuhan, #askwho, #coronavirusoutbreak, #china and #cononaoutbreak. Over this period, the COVID-19 just began in the world, people had not much self-perception about how the virus would affect their living.

In the second period, the tweets increased significantly, thus, the hashtags presented more variously, people had known the seriousness of the coronavirus and the importance of self-protection. Also, in this period, the most populous country India was becoming out of control the the spread of the virus. These changes were apparently demonstrated in the hashtags in the non-rumour group, like #itelsaysstaysafe, #healthcareheroes, #getusppe, #rentrelief, #precaution, #india, #who, #socialdistancing, #lockdown, #ppe, #stayhome and #quarantine. On the contrary, their opposite side did not care much about these topics, they extended the last period's topics to blame Trump and mislabelled the virus to China, these hashtags #wuhanvirus, #chinavirus, #usa, #trumppandemic, #worstpresidenever, #trumpliedpeopledied were popped up to spread rumours.

In the third time period, rumour spreaders remained blaming Trump and also added topics about

	Rumour Topics
1	control dr february fauci disease coronavirus went u march trump
2	covid 19 family one like many know life lost would
3	coronavirus trump :folded_hands: say :police_car_light: fine
4	said remember coronavirus would doctor big trump worry say deal
5	usa blame coronavirus trump dems rest president handle pelosi nancy

Table 4: Rumour Topics

	Non-rumour Topics
1	2 week month 1 3 two data ago 4 day
2	way 'm coronavirus 'll sure would long think 'd best
3	health public house coronavirus white cdc pandemic official expert :rolling_on_the_floor_laughing:
4	control political party coming left biden racist joe allowed coronavirus
5	lockdown india govt pm taken country time measure #covid19

Table 5: Non-rumour Topics

the black-man Floyd, new hashtags #blacklivesmatter, #liarinchief and #georgefloyd occurred with high frequency. Whereas in the non-rumour cohort, just a few new hashtags, #hypochlorite, #disinfectant, #bleach and #hydroxychloroquine were added, non-rumour holders kept their interests in health and protection.

In the final part, rumour holders strengthened their anger by adding new hashtags #trumpresignnow, #votetrumpout, #tre45on, #trumptraitor, and #russianbounty. While the non-rumours holders, five times as many as rumour holders, just merely focused on a couple of topics, high frequent tags like #contest and #tmobiletuesdays [3] appeared among them. Promotions on internet connection and telecommunication had highly drawn their attention, these themes became much more essential for them being restricted to staying at home.

6.2.1 Topics

With regard to the topics, we aimed to extract commonly discussed topics, investigated the prevalence, and analysed the difference between them. We visualized high frequency words clouds and correlations between words first, and constructed a topic model using Gensim Linear Discriminant Analysis-Latent Dirichlet Allocation(LDA) algorithm [4], researchers applied this algorithms to extract topics (Chakkarwar and Tamane, 2020) (Bastani et al., 2019). After removing emojis, urls and stopwords, we tokenized and lemmatized the tweets to produce a base model, then tracked the progress for hyperparameter tuning. The LDA model produced the topics for each side. Table 3 and Table 4 illustrated the top five topics for two sides.

Topics were sorted from the most relevant to the least for each topic, "coronavirus" occurred in both groups, it reflected the mutual interest of two groups. However, comparing with the position where it stationed, the frequency in the rumour side was apparently much higher than the counterpart. "trump" was the second frequent one in the rumours, whereas it did not appear in top topics in the opposite group, non-rumours users showed less interests in it. And "covid 19" and "control" existed in both teams, they showed more frequently in the rumours than non-rumours.

Apart from those common and most frequent words, others in the topics for each category were evidently distinct. Rumours showed interests in "dr","remember","said", "disease", "blame" and so on. While non-rumour users were generally concerned with "week", "day", "way", "health", "public", "lockdown" vice versa.

The topics that the LDA model generated enhanced our understanding of import themes over that specific history period.

6.3 Conclusion

The analysis on the hashtags, trends and topics from the COVID-19 tweets can be useful to help us understand the public's opinion and gave us insights into the impact of COVID-19 on society.

[3] T-Mobile Tuesday Promotion https://www.t-mobile.com/offers/t-mobile-tuesdays

[4] Latent Dirichlet Allocation (LDA) https://radimrehurek.com/gensim/models/ldamodel.html

References

HA Antoine. 2016. Hashtagging: What you need to know about hashtags as trademarks, hashtag litigation, the ftc, viral campaigns, and more. *The Computer & Internet Lawyer*, 11(6).

Kaveh Bastani, Hamed Namavari, and Jeffrey Shaffer. 2019. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271.

Vrishali Chakkarwar and Sharvari Tamane. 2020. Social media analytics during pandemic for covid19 using topic modeling. In *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, pages 279–282. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Krzysztof Fiok, Waldemar Karwowski, Edgar Gutierrez-Franco, Mohammad Reza Davahli, Maciej Wilamowski, Tareq Ahram, Awad Al-Juaid, and Jozef Zurada. 2021. Text guide: Improving the quality of long text classification by a text selection method based on feature importance. *IEEE Access*, 9:105439–105450.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd international conference on world wide web*, pages 593–596.

Grant D Jacobsen and Kathryn H Jacobsen. 2020. Statewide covid-19 stay-at-home orders and population mobility in the united states. *World medical & health policy*, 12(4):347–356.

Myeong Gyu Kim, Minjung Kim, Jae Hyun Kim, and Kyungim Kim. 2022. Fine-tuning bert models to classify misinformation on garlic and covid-19 on twitter. *International Journal of Environmental Research and Public Health*, 19(9):5126.

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. *arXiv preprint arXiv:1907.11692*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Roxana Radu. 2020. Fighting «infodemic»: Legal responses to covid–19 disinformation. social media & society. 1-4. *Doi*, 10(1177):2056305120948190.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22:100104.

A Team Contribution

Table 6: Team Contribution

1183945
Tweet Objects crawl, All the data processing, Baseline model, BERT-tweet Model, Task 2 COVID-19 Model
1227216
BERT-using TensorFlow, RNN, BERT-using TensorFlow applying data-transformed
1212462
Part of Baseline Model, tried to BERT-Tweet Large, Report