COMP90042 Assignment3 Report

Anonymous, Team of 3 students

1 Introduction

The information and statements on social medias can be a mixture of verified statements and rumours. Rumours are misleading and sometimes can propagate quickly and cause severe consequences. Thus rumour detection is useful for both governments and social media platforms. The rumour detection consists of experimental pipeline, features and models, compare the results of different models and end with critical analysis of these models. Then analysis were performed on COVID-19 tweets.

2 Experimental Pipeline

The entire experimental pipeline is shown in Fig. 7.In the first part of this project, we have been provided a set of tweet IDs and used Twitter API to crawl each tweet object. After that, we did some preprocessing on tweet objects in order to retrieve features and built some deep learning models such as Bidirectional Encoder Representations from Transformers(BERT), Neural Network(NN), Long Short-Term Memory(LSTM), and Attention to predict the labels. In the second part, we applied the model that we have trained to predict rumour/non-rumour on some COVID-19 tweets and performed analysis including topic extraction and sentiment analysis.

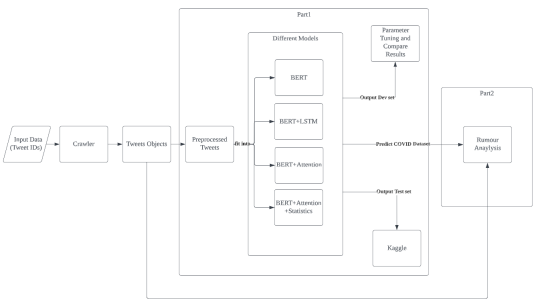

Figure 1: Pipeline

3 Features Engineering and Pre-processing

We were provided train and development dataset, the format of an instance of the dataset consists of a source tweet and multiple reply tweets, we ordered the reply tweets by time to represent the propagation path of the source tweet. Only texts from tweets are used for the feature because the content is needed to predict rumour. Each tweet sequence is padded to an equal length of 512 words. We preprocess the tweet sequence to exclude URLs and mentions(i.e @somebody) since they do not provide any meaningful context. Ideas are from the hugging face website [1]. After preprocessing, each tweet in a tweet sequence is joined using a [SEP] token and added [CLS] token at the front. This is preparing the dataset to be fed into the BERT model.

After careful analysis of the rumour and non-rumour COVID tweets(in Section 8), it appears that rumour have a significantly higher retweet count and rumour posters have more followers. So these two features are included with text or tweets to feed into the models.

4 Models

4.1 BERT + NN(Baseline)

In recent years, a pre-trained language model called BERT(Bidirectional Encoder Representations from Transformers) comes out and it is state-of-the-art on many NLP text classification tasks through the fine-tuning approach. It takes raw text as input and outputs [CLS] tokens as the embedding that captures the contextual representation of each sentence. (Devlin et al., 2018) Since we put a [SEP] token to separate individual tweets within tweet sequence and put a [CLS] before each source tweet. The output [CLS] token from BERT only captures each

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base

source tweet. Then the CLS tokens are fed to a NN with one layer for classification.

4.2 BERT + LSTM

LSTM is a variant of RNN which introduces gates to minimise the vanishing gradients problem of RNN. It has the ability to capture long range contexts. We stack LSTM on top of BERT to try to capture sentence level relationships. Normally, only the [CLS] tokens are used for further usage to build classification models. Here, we propose that since the dimensions and embedding format of [SEP] is the same as [CLS] tokens, we use [SEP] just like [CLS]. Where [CLS] captures the source tweet, and each [SEP] captures a reply tweet contextual representation. This model uses [CLS] and [SEP] tokens from input BERT and applies LSTM on top of that. The reason for using LSTM is to prevent gradient descent by using additional gates to control the flow of information and wish to catch longer dependencies than the vanilla RNN method.

4.3 BERT + Attention

This model uses outputs from BERT and applies Attention. Compared to the second model, Attention replaces LSTM in an effort to highlight and find relevant features from the BERT output and further boost precision compared to the LSTM model. In each path of tweets and at each time step of propagation, a weight is calculated using the softmax function with input from transformed embeddings using a linear layer and an activation layer. Then the new embedding is the weighted average of original embeddings. This transformation will give the tweet with more information a higher weight and thus can be beneficial for the classification.

4.4 BERT + Attention + NN

This model is similar to the third model and the only difference is that we added another one layer neural network after Attention. This is to include two more features: retweet counts and followers counts. Since after rumour analysis, these two features seem to be different for rumour and non-rumour tweets. This also circumvents the requirement of BERT which needs text input.

5 Evaluation Metrics

In this report, precision, recall and F1 score are used as evaluation metrics. Accuracy is not used because it can oftentimes be misleading for an imbalanced dataset which is our dataset as shown in Fig. 2. For precision, recall and F1, we used rumour to be our interested(positive) class. Precision indicates the proportion of actually positive is predicted as positive. Recall indicates the proportion of predicted positives is actually positive. F1 score considers both precision and recall. We will use the F1 score to compare the result since it is the sole determiner metric for the in-class Kaggle competition.

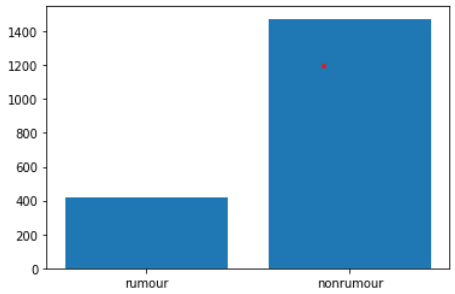

Figure 2: The distribution of Rumour and Non-Rumour for train dataset

6 Result

Table 1 shows the model performance results.

7 Critical Analysis

7.1 Result Analysis

Our BERT plus NN models achieved an F1 score of 0.86. This model doesn't use any assumption of [SEP] token containing the information of following tweets and only uses [CLS] for classification. The performance is not significantly worse than other models. This indicates that using only contextual representations of source tweets can construct a relatively well rumour classification model. Our BERT plus LSTM model performed the worst. This is vastly different from our expectations since LSTM often produces great results. This model wishes to use sequence modelling to capture the information flow during the propagation. However, the performance is not as good as the baseline model. This could be because LSTM failed to capture long range dependence, and with the very limited amount of training dataset(1,895 instances) causes the model to overfit. Another problem with LSTM is that it does not stack well. These could all be the reason for poor performance. Our BERT plus Attention model achieved second best performance with an F1 score of 0.884. When we further combine Attention with BERT, the performance

Model	Dev Accuracy	Dev Precision	Dev Recall	Dev F1	Kaggle Private F1
BERT	0.942	0.904	0.817	0.858	0.891
BERT+LSTM	0.937	0.886	0.809	0.846	0.853
BERT+Attention	0.962	0.852	0.916	0.883	0.873
BERT+Attention+Statistics	0.953	0.867	0.909	0.889	0.894

Table 1: Model Performance

is better than the previous two models. The attention seems to be able to capture the information related to the importance of each tweet and can make a better classification. Our BERT plus Attention and NN model performed the best with an F1 score of 0.889‘.This is a slight improvement to the third model because we included more features into this model. This indicates that followers count and retweet count are related to classification of rumour and should be used as features. Even with a small training set, NN does not overfit because we only have a single layer NN.

7.2 Parameter Tune

On the development set, we test whether different bert models can influence the result. From the Hugging Face website [2], there are several BERT-related models. We tried three BERT models namely “bert-base-uncased”, “bert-base-cased”, and “roberta-base”. Fig. 3 shows the performance on the 4th model(BERT + Attention + Statistics), we can see that “bert-base-uncased” gives us the best performance. As in the BERT-uncased model, the tweets are all converted to their lower case after the tokenization and this will reduce the total number of words significantly.

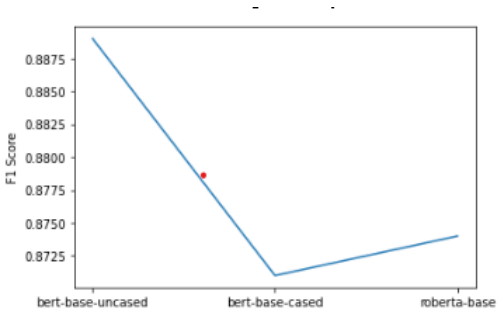

Figure 3: Different bert models

The reasons that BERT-cased performs the worst could be unlike the sentiment analysis task, the cased text and uncased text carries similar information in our task, but with cased words, the model might treat capitalised and uncapitalised words as totally different embeddings. “Roberta-base” also performs worse than bert-uncased, because the model is much larger and more complicated than bert-uncased and with such a small dataset, it is easy to be overfitting. In certain tasks like propaganda technique classification, using BERT-uncased can have significant improvement. (Altiti et al., 2020) .

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

8 COVID Rumour Analysis

8.1 COVID Exploratory Data Analysis

The COVID dataset, after we used Twitter API to crawl the specified tweets, contains 15,955 tweets in total. All tweets are from January to August 2020. Then the best performance rumour classifier is applied to the unlabelled COVID-19 tweets dataset to distinguish rumour tweets. We perform exploratory data analysis on both rumour and non rumour tweet sets to examine the potential characteristics of the dataset. We performed an extensive analysis of the COVID dataset. The COVID tweets were preprocessed with four steps first, then basic analysis like length distribution, popular hashtags, and rumour creating user’s characteristics was conducted to examine the difference between rumour and non-rumour sets. Moreover, we examine the differences between the content of rumour and non-rumour set, so topic extraction with Latent Dirichlet Allocation(LDA) and sentiment analysis with Valence Aware Dictionary and sEntiment Reasoner(VADER) were performed to study the content of the COVID dataset. The preprocess of COVID tweet text includes 1. Remove URLs and user mentions, 2. Tokenize and lowercase the texts, 3. Remove tokens that do not contain the English alphabet, and 4. Remove stopwords. After the classification of the COVID dataset, there are 4,267 rumour tweets which consist of 26.7 percent of the entire dataset, and 11,688 non-rumour tweets which consist of 73.3 percent of the dataset. The dataset is dominated by non-rumour tweets and the

number of non-rumour tweets is roughly 2.75 times higher than rumour tweets. The retweet counts of rumour are 1.4 times higher than non-rumour tweets.

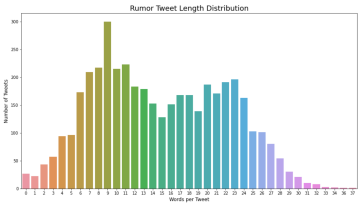

Figure 4: Rumour Tweet Length Distribution

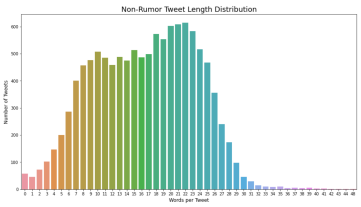

Figure 5: Non-Rumor Tweet Length Distribution

From Figure.4 and Figure.5, rumour tweets length distribution is similar to a bimodal distribution with a peak at lengths 9 and 23, meaning that most rumour tweets length are centred around these two peaks. Non-rumor tweets have a right-skewed distribution with most tweets having a length between 11 and 22. Non-rumour tweets are on average longer than rumour tweets with an average length of 16.45 compared to rumour's 14.84.

8.2 Popular Hashtags of Rumour and Non-rumour

We analyse popular hashtags to better understand the topic or trends of the COVID dataset, which sets up for later topic modelling. We extract only the hashtags from tweets' text and create a frequency distribution of these hashtags for both rumour and non-rumour tweets. For both rumour and non-rumour the top two hashtags are 'covid19' and 'coronavirus', and they appear 5 to 8 times more frequent than the third most popular hashtags. This is apparent considering the dataset consists of COVID tweets. Out of the top fifteen most popular hashtags, four hashtags appear in both rumour and non-rumour tweets: 'covid19', 'coronavirus', 'breaking', and 'coronaviruspandemic'. And rumour popular hashtags tend to involve U.S. President Trump and China whilst non-rumour popular hashtags are more about how to deal with COVID-19 such as social distancing, lockdowns, and staying safe.

8.3 User Characteristics

Aside from COVID tweets' text, analysis of rumour and non-rumour generating users was performed. We examine the follower counts of users to see if a discrepancy exists. The mean followers count of rumour user is 4,608,595 and the mean followers count of non-rumour user is 5,004,710. It is apparent that rumour generating user have less followers.

8.4 Topic Modelling With LDA

LDA is an unsupervised ML model that takes documents as input and produces the distribution of topics for each document and the distribution of words for each topic. It tries to find some natural group of topics from the documents. LDA uses Gibbs sampling which first randomly assigns words to topics, then for each iteration calculates the probability of the topic given a document and the probability of a word given a topic and assigns words to the most likely topic. After a large number of iterations, reassignment of words to topic stops and now we have the topic models created by LDA. (Jelodar et al., 2019). LDA is easy to implement and works well with shorter text. The COVID dataset also satisfies LDA's assumption that the input words are related since all tweets are about COVID. LDA has proven successful in extracting topics from COVID data in South Africa. (Mutanga and Abayomi, 2022)

We had preprocessed our tweet text further to achieve better results with topic modelling. Words like 'covid' and 'coronavirus' were removed since all COVID tweets are about COVID and coronavirus. Based on the length analysis from above, tweet-length less than four are excluded since tweets with minimal text hardly convey a topic discussion. We chose the number of topics for our LDA model to be fifteen and used Gensim Library to create a bag-of-words representation of tweets to feed into the LDA model.

The topics of rumour and non-rumour sets are very different. All of the topic models are similar to Figure.6 Rumour topics are harder to interpret and many topics share the same words. Rumour topics are about former U.S. President Trump, China, COVID situation in each U. S. state(confirmed and death cases). The non-rumour topics include social distance, wearing face masks, Fauci, WHO, and confirmed cases. We separated rumour tweets into three parts to see if the topic or trend of ru-

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05
0	trump	trump	case	trump	trump
1	test	people	report	america	mask
2	crisis	case	death	china	penny
3	health	death	trump	million	wear
4	event	texas	state	number	china
5	tell	american	florida	president	people
6	time	rally	record	country	president
7	social	million	break	help	american
8	distance	florida	total	death	spread
9	away	record	number	think	mike
10	positive	state	china	american	white

Figure 6: Example Output of Topic Model

mour tweets changes over time. 16 percent of rumour tweets are from January, February, and March, 52 percent of rumour tweets are from April and May, and 32 percent of tweets are from June, July, and August. The first part of COVID's rumour tweets mostly discusses China and Trump, there is also a topic about Italy. For the first three months of 2020, the coronavirus was first discovered and made known in Wuhan, China. And later, Italy also had a huge number of confirmed cases. For the second part of rumour tweets, the topics had changed to countries of other parts of the world such as Brazil, Austria, and Australia which indicates the widespread of the COVID-19 pandemic. Topics like masks and restaurants start to appear which indicates the prevention methods for COVID-19. In the last part of rumour tweets, the topics are more similar to the topics of the entire rumour tweet set which discuss Trump and the U.S. with many U.S. states appearing in the topic model.

8.5 Sentiment Analysis with VADER

VADER is a text sentiment analysis model that maps documents to positive and negative polarity. The model does so by mapping words in each document to sentiment scores (emotion intensities) based on a dictionary. Thus the sentiment score of a text is the sum of the intensity of each word within the text. VADER is used to perform sentiment analysis on rumour and non-rumour tweets because it is Sensitive to sentiment expression in social media (Elbagir and Yang, 2019). It is easy to acquire, deploy and produce satisfactory results.

Sentiment analysis was performed on rumour and non-rumour tweets as well as the replies to rumour and non-rumour tweets. From Figure.7, we can see that rumour tweets convey more negative sentiment whilst non-rumour tweets have roughly the same number of positive and negative sentiments. rumour contains 1.7 times more negative sentiment than positive sentiment tweets. But non-rumour tweets have 1.1 times more positive sentiment than negative sentiment tweets. The replies to rumour tweets contain 1.8 times more negative than positive sentiment tweets. And the negative sentiment tweets in replies to non rumour tweets are only 1.3 times higher than positive sentiment tweets. Overall, both the rumour and replies to rumour tweets convey more negative sentiments than positive and neutral sentiments. And non-rumour is the only group that has more positive sentiments than negative sentiments.

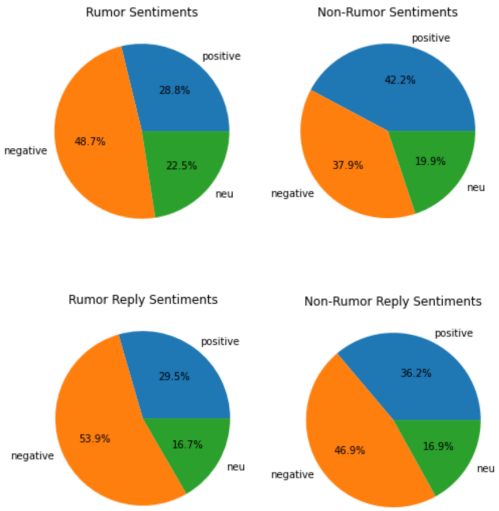

Figure 7: shows the distribution of positive, negative and neutral tweets for rumour, non-rumour and their replies

9 Conclusion

In conclusion, we have built the best classifier using the combination of BERT-uncased, attention and statistics. This model obtains the highest F1 equal to 0.894 among other models. We also extensively analysed different aspects of COVID rumour and non-rumor tweets. In the future, we hope to shift focus to "white-box" models and see the performance of Logistic Regression and Naive Bayes.

10 Contribution

Team Member1: Task 2 of the report. Written all the codes for task 2 and wrote the second part of the report which is about task 2.

Team Member2: Wrote introduction, models, critical analysis for the report. Coded dataset file and models file for task 1.

Team Member3: Wrote Result, Evaluation, Pipline for the report. Coded the Pytorch-lightning

framework.

References

Ola Altiti, Malak Abdullah, and Rasha Obiedat. 2020. Just at semeval-2020 task 11: Detecting propaganda techniques using bert pre-trained model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1749–1755.

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*.

Shihab Elbagir and Jing Yang. 2019. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, page 16.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Murimo Bethel Mutanga and Abdultaofeek Abayomi. 2022. Tweeting on covid-19 pandemic in south africa: Lda-based topic modelling approach. *African Journal of Science, Technology, Innovation and Development*, 14(1):163–172.