Rumour Detection and Analysis on Twitter

Student Number: [redacted]

Abstract

Social media has become one of the most important platforms for information propagation in this generation. However, some of the unverified statements, which also known as rumours (Maddock et al., 2015), are also prevalent on it. Due to the open nature, rumour is unavoidable on social media such as Twitter. Therefore, an efficient rumour detection technique is needed, especially during the covid-19 pandemic to prevent public panic (Depoux et al., 2020). In this report, I will discuss and implement one of the state-of-the-art natural language processing technique, the Bidirectional Encoder Representations from Transformers (Devlin, Chang, Lee, & Toutanova, 2018), on the Twitter dataset and analyse the special pattern of rumour about the covid-19, to prevent the misleading information spreading during this pandemic.

1 Introduction

Rumour is defined as unverified statements (Maddock et al., 2015) which could be misleading information or even a conspiracy theory. The open nature of social media allows users to spread rumours around the world (Wen et al., 2014). Rumours on social media pose an unignorable threat to nowadays society due to its fast propagation speed (Sahafizadeh & Ladani, 2018).

The spreading of rumour may cause significant impact on human affairs (Nekovee, Moreno, Bianconi, & Marsili, 2007) such as the public opinion of a country (Galam, 2003), financial market (Kimmel, 2004), and even causing public panic (Depoux et al., 2020). Take the covid-19 rumour for example, the misinformation about the outbreak of covid-19 created panic among the public, caused "panic buying", the excessive prevalence of buying utilitarian goods (Barnes, Diaz, & Arnaboldi, 2021). This indicates the importance of having an efficient rumour detector during a pandemic such as covid-19 to filter out the rumour on social media.

To address these issues, a rumour detection system is built to recognize the rumour spreading on Twitter and a covid-19 rumour analysis is conducted to distinguish the special pattern and mechanism of rumour during this recent pandemic.

In section 2, I will use Bert (Devlin et al., 2018), a pre-trained & fine-tuning based language model, to contextually embed the tweets on Twitter, and then implement a neural network trained on a labelled Twitter dataset as a classifier for rumour and non-rumour.

In section 3, I will use this trained model to classify an unlabelled covid-19 tweets dataset and analysis the special pattern and sentiment of the covid-19 rumour.

2 Rumour Detection System

In this section, I will discuss the performance of the state-of-the-art sentence-based embedding method on our Twitter rumour detection task. Also, the data pre-processing will be discussed.

2.1 Dataset

In this task, a set of source tweets and their replies are provided. A source tweet and its reply tweets are called an "*event*". Each event is labelled as either a rumour or non-rumour. There are 4641 events in the training set. 3058 of them are non-rumour and 1583 are rumour, which is slightly unbalanced.

2.2 Model Selection

Transfer learning by using a pre-trained language model can effectively improve the performance of a wide range of natural language processing (NLP) tasks (Cer, Diab, Agirre, Lopez-Gazpio, & Specia,

2017) including semantic analysis. Rumour classification can be treated as a kind of semantic analysis, therefore, it can also benefit from incorporating a language model pre-trained on a large corpus for text embedding.

In order to make our model able to represent the whole context of each tweet, I choose to implement the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) due to its better contextual representation than other unidirectional models. BERT is a masked language model, which uses a multi-layer bidirectional transformer to encode a sentence. This makes BERT capable to capture both left and right context. The bidirectional nature of BERT makes it outperforms the unidirectional language model such as ELMo (Peters et al., 2018) and OpenAI GPT (Radford, Narasimhan, Salimans, & Sutskever, 2018).

Another benefit of BERT is that BERT is a fine-tuning based representation model instead of a feature-based approach that require task-specific architecture (Devlin et al., 2018); this means that BERT is task-independent. We do not need to heavily engineer a task-specific architecture. All we need is to add an extra classification layer on top of the BERT model. The contextual embedding generated by BERT will directly feed into the classification layer as inputs. During the training, the parameters of BERT for the words embedding will also be fine-tuned with the parameters of the classification layer. This makes BERT flexible to a variety of NLP tasks and easy to implement.

Moreover, the attention principle of transformers in BERT can make the model focus on important words, hence, BERT can reach a better performance than feedforward network and recurrent network based language model (Devlin et al., 2018).

To sum up, due to the transfer learning approach, bidirectional nature, attention principle in transformers, and fine-tuning based architecture, BERT can reach state-of-the-art performance in multiple NLP tasks including semantic analysis. Therefore, in this project, I will implement BERT with an extra classification layer as my rumour detection model.

2.3 Data Pre-processing

Raw texts are usually directly fed into the BERT model without any pre-processing such as word normalisation. BERT uses byte-pair encoding (BPE) to tokenise the words, therefore, the infection of words can automatically handle (e.g. the word "running" is decoded into "run" + "##ing", where "##" token means sub-word span). Also, the stop words don't need to be removed, because the transformers in BERT use the attention principle to focus on words that have an impact on output instead of the common words. In conclusion, lemmatisation, stemming, and stop-word removal are not necessary for BERT model.

However, since our texts are from Twitter, those texts may contain "#hashtag" or "@mention". These words started with special "#" or "@" tokens may be treated as unknown words and deteriorate the performance of my model. To examine how the "#hashtag" and "@mention" affect the performance, I trained my model by using three different pre-processed training sets. The detailed performance comparison will be discussed in the result section.

The tweets dataset is in a special structure that each event contains a source tweet and multiple reply tweets. To decide whether a source tweet is a rumour or not, the context of its reply tweets is also very crucial because some users may question the authenticity of the source tweet in their reply. Therefore, I extracted the text of the source tweet and append [CLS] token in front of them. Then, I concatenated the text of each reply tweets within the same event one by one with the source text by using [SEP] token to separate them. This treatment can maintain the source-reply structure while keeping the full-text context.

2.4 Training

This rumour detection model is trained on Google Colab. First, the text of events is fed into the pre-trained BERT model for embedding. Then the embedding contextual representation of each event is fed into the classification layer (neural network) for training. During the training, the labelled data is used, and the parameters of both the classification layer and the transformers layer will be updated.

The batch size is set to 20 due to the limited GPU memory size. The training curve is shown in Figure 1 and the development curve is shown in Figure 2. Here, f1 score is used as the evaluation metrics because the number of rumour and non-rumour in the training set is slightly unbalanced; therefore, the f1 scores can represent more comprehensively than accuracy.

As we can observe from Figure 1, the training loss decreases very fast and the f1 score of training reaches 100% very soon. This indicates the tendency of overfitting. The overfitting situation can be further verified by the development curve (see Figure 2); we can see that after epoch 4, the development loss starts to increase and the development f1 score becomes saturated. Therefore, I extracted epoch 4 as my final rumour detection model.

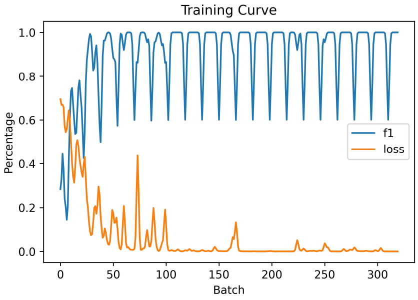

Figure 1. Training Curve*

fluctuating due to small batch size

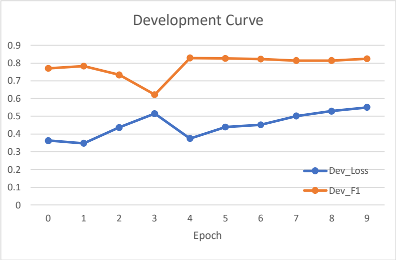

Figure 2. Development Curve

2.5 Result & Discussion

As mentioned in section 2.3, in order to test the impact of "#hashtag" and "@mention", I trained the model on three different pre-processed datasets. Their performance is shown in Table 1. The result shows that only removing "@mention" while keeping "#hashtag" reach the best performance. Then, I investigated the dataset to see the wrong classified events. I found that hashtags usually contain crucial information about events such as the name of a protest or the location of a crime scene. On the contrary, the mentions are usually just some random Twitter usernames which are not helpful for our rumour classification test. Therefore, missing hashtags information will decrease the model performance.

The final model I submitted to the Codalab was therefore trained on the data which keep the hashtags. The final result is shown in Table 2.

	Hashtag Removed	*Mention Removed*	*Both Removed*
Dev F1-Score	0.804	0.833	0.81
Dev Precision	0.806	0.854	0.812
Dev Recall	0.802	0.813	0.807

Table 1. Comparison of Different Data Pre-processing

	Test (final)	*Test (ongoing)*	*Development Dataset*	*Train Dataset*
F1-Score	0.8241	0.8287	0.833	0.996
Precision	0.8367	0.8621	0.854	0.995
Recall	0.8119	0.7979	0.813	0.996

Table 2. Final Model Performance

As Table 2 indicates, my model performed similarly on both the test and develop dataset. Also, my BERT model outperforms the LSTM model, which only reached 77% f1 on the development set.

3 Covid-19 Rumour Analysis

In this task, I will classify unlabelled Covid-19 tweets by using the rumour detection model trained in the previous task. Then, I will analyse the sentiment, topic, hashtag, and user information of the Covid-19 rumours.

3.1 Data Distribution

There are 17458 events in the Covid-19 tweets dataset; 1621 of them are classified as "rumour" while 15837 of them are classified as "non-rumour".

3.2 Hashtag Analysis

Here, I investigate the most common hashtags in both the source tweets and the reply tweets.

As Figure 4 indicates, the frequent hashtags of source tweets in both rumour and non-rumour are just normal tags for news or events. The rumour tends to use "#Breaking" tags more often, and there are more location or event-related hashtags such as "#India", "#China", "#MLB". On the other hand, the frequent hashtags of reply tweets show a more explicit pattern of rumour. The rumour tends to use strong biased and emotional hashtags such as "#ChinaVirus" and "#ChinaLiedPeopleDied", or appalling hashtags such as "#SecondWave" or "#CoronavirusOutbreak" (see Figure 4). This can prove that people tend to react emotionally to fake news and become biased when affected by rumours.

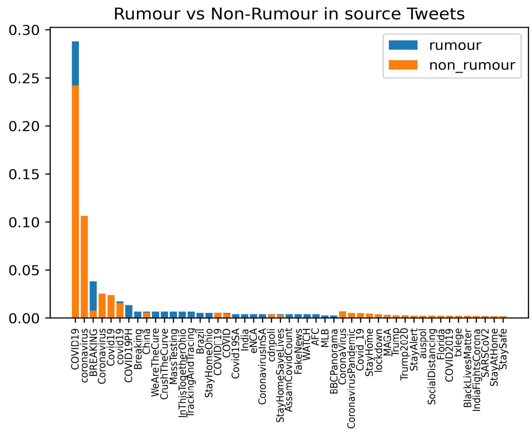

Figure 3. Source Tweets Hashtags Distribution

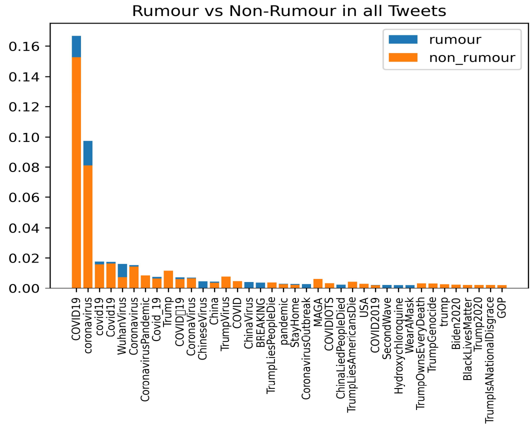

Figure 4. Reply Tweets Hashtags Distribution

3.3 Source User Analysis

As we can see in Table 3, the source users of a rumour tweet and a non-rumour tweet share the same pattern. The source account of rumour tends to have slightly fewer followers and friends, and shown in fewer public Twitter lists. Also, the source account of rumour tends to issue more tweets than a normal account.

	Rumour	*Non-Rumour*
Account Created Year	2010.77	2010.88
#Followers	4258754	5464000
#Friends	7185	7934
Raito: having profile image	100%	99.9%
Ratio: is verified	76%	74.2%
#Public lists this account in	14650	15533
#Tweets issued by this user	134357	114158

Table 3. Source User Average Information

3.4 Rumour Topic & Semantic Analysis

To analysis the semantic of rumour tweets and non-rumour tweets, I implemented the Stanza Sentiment Analysis tool created by Stanford NLP Group (Qi, Zhang, Zhang, Bolton, & Manning, 2020). This tool uses CNN classifier to classify sentence into three categories: "negative", "neutral", and "positive". The result indicated that both source and reply tweets of rumour tend to have more negative sentiments than non-rumour, and the difference is more obvious in the source tweets (see Table 4).

	Source Tweets		Reply Tweets	
	Rumour	Non-rumour	Rumour	Non-rumour
Negative	32.3%	29.5%	30.4%	29.9%
Neutral	62.8%	64.0%	64.8%	65.2%
Positive	4.9%	6.6%	4.8%	4.9%

Table 4. Semantic Analysis

Then, I investigated the topic of the rumour tweets. Most of the rumour topic of the source tweets are about covid-19 outbreak, new case, fake information about lockdown, fake policy of the government, or fake statistics of covid-19 such as the fatality rate. Those rumour sources tend to be appalling, panic-causing or political orienting, such as hate speech, discrimination, or even racism.

The reply tweets also show specific patterns, such as questioning the authenticity, disputing or debating with the source user, showing overreacted emotions, and sometimes including swear words.

4 Conclusion

The BERT model with one classification layer reached 83% f1 score on an unseen test set with 86% precision and 80% recall. This indicates the predicted rumour is likely to be a true rumour while some of the rumours are still undetected. The undetected rumours tend to have neutral sentiment with just plain descriptions; therefore, are not well captured by only the text content. I believe if the reply tree structure is incorporated into my model, the recall can be further improved.

This report also showed that the user information difference between rumour source and non-rumour source is not obvious; this means that nowadays, the rumours are mostly spread by normal users rather than by zombie accounts. The semantic analysis shows that the rumours tend to be more negative and hostile; and the hashtag analysis shows that the reply tweets of rumour tend to be more biased and emotional.

After understanding the nature of Covid-19 rumour, the negative sentiment of source tweets and the biased replies can be used as metrics for Twitter user to manually distinguish the rumours.

References

Barnes, S. J., Diaz, M., & Arnaboldi, M. (2021). Understanding panic buying during COVID-19: A text analytics approach. *Expert Systems with Applications, 169*, 114360.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., & Larson, H. (2020). The pandemic of social media panic travels faster than the COVID-19 outbreak. *Journal of Travel Medicine, 27*(3). doi:10.1093/jtm/taaa031

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Galam, S. (2003). Modelling rumors: the no plane Pentagon French hoax case. *Physica A: Statistical Mechanics and its Applications, 320*, 571-580.

Kimmel, A. J. (2004). Rumors and the financial marketplace. *The Journal of Behavioral Finance, 5*(3), 134-141.

Maddock, J., Starbird, K., Al-Hassani, H. J., Sandoval, D. E., Orand, M., & Mason, R. M. (2015). *Characterizing online rumoring behavior using multi-dimensional signatures.* Paper presented at the Proceedings of the 18th ACM conference on computer supported cooperative work & social computing.

Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications, 374*(1), 457-470.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Sahafizadeh, E., & Ladani, B. T. (2018). The impact of group propagation on rumor spreading in mobile social networks. *Physica A: Statistical Mechanics and its Applications, 506*, 412-423.

Wen, S., Haghighi, M. S., Chen, C., Xiang, Y., Zhou, W., & Jia, W. (2014). A sword with two edges: Propagation studies on both positive and negative information in online social networks. *IEEE Transactions on Computers, 64*(3), 640-653.