COMP90042 Group Project: Twitter Covid-19 Rumour Analysis

Abstract

Rumours in social media have become a big concern. Due to complex reasons like Information Asymmetry, it is usually difficult for one to justify the truth of the messages in the Internet. However, around a specific topic, rumours may have some common features, such as certain topics and keywords. Owing to the development of pre-trained language models, we can now better capture these features and identify rumours from non-rumours. In this report, we focus on Covid-19 rumours collected from Twitter, proposing a method based on pre-trained language models to make classification. Besides, we also discuss the patterns of Covid-19 rumours and non-rumours from different perspectives.

1 Introduction

There has been a lot of research on identifying Twitter rumors by traditional machine learning methods, such as decision trees, plain Bayesian, SVM and KNN for Persian rumor identification(Zamani et al., 2017), and logistic regression for medical rumor identification(Dito et al., 2020). However, these models require feature engineering and are difficult to use with large training sets.

One of the popular areas in machine learning is deep learning, such as CNN model(Alsaeedi and Al-Sarem, 2020), which saves time in feature engineering and performs better on large data sets compared to traditional machine learning methods. In addition, many studies have shown that language pre-training models can effectively improve the results of many natural language tasks.These include tasks that analyze and predict sentence connections at the sentence level, and more fine-grained output at the token level.BERT is one representative of such models(Devlin et al., 2018).

The study(Liu et al., 2019) pointed out that the BERT model is undertrained and proposed the RoBERTa model based on it. The Sentiment Knowledge Enhanced Pre-training (SKEP) model (Tian et al., 2020b). significantly outperformed the RoBERTa model. In this project, we first studied the related work of this topic. Then we did some pre-processing on the dataset to remove words, symbols and links that have no actual meaning. After that, different models were used on the dataset for comparative experiments to obtain the optimal model. Finally, we will analyse the characteristics of the Covid-19 rumours and non-rumours from the perspective of topic, hashtags, time trends, sentiments, and their creators.

2 Dataset

We use different datasets for the two tasks in our project.

For **Task 1**, we simply use the provided dataset. The details of the provided dataset are shown in Table 1.

Dataset for Task 1	**# Instances**
task1 training set	1813
task1 development set	595
task1 testing set	558
Dataset for Task 2	**# Tweets**
task2 kaggle dataset	179108
task2 provided dataset	161916

Table 1: Dataset Information

For **Task 2**, we use two different datasets. The first one is extracted from Kaggle [1] every tweet in this dataset is not retweeted. Based on the given twitter ids in this project, the second dataset is obtained by crawling tweets using the Twitter API [2]. The number of tweets contained in both datasets is shown in Table 1, while the specific content for each dataset is displayed in Table 2. Based on

[1] https://www.kaggle.com/datasets/gpreda/covid19-tweets

[2] https://developer.twitter.com/en/docs/twitter-api

the characteristics of the two different datasets, we conduct different analysis on them.

Dataset Content	Kaggle	Provided
tweets	✓	✓
date	✓	✓
source	✓	✓
location	✓	✓
sentiment	✓	✓
hashtag	✓	✗
user name	✓	✓
user verified status	✓	✗
user description	✓	✗
user followers	✓	✗
user friends	✓	✗
user favourites	✓	✗
user created time	✓	✗
retweeted	✗	✓

Table 2: Task 2 Datasets Content

3 Methods

3.1 Data Clean Method

In the language environment of Twitter, there are many words, symbols, and links with no actual meanings such as user names, emojis, and urls. These patterns are usually not covered in the dictionary of pre-trained models and might pose negative influence on our classifier. Thus, we remove these patterns by applying a stopword list and regular expression matching

3.2 Model Structure

In this project, we apply Sentiment Knowledge Enhanced Pre-training (SKEP) (Tian et al., 2020a) in our rumour detection task. SKEP incorporates sentiment knowledge by self-supervised training by sentiment masking and defining three pre-training objectives. First, Sentiment Masking is used to recognize sentiment information of input sentences based on automatically-mined sentiment knowledge. Then, the sentiment information will be remove to produce a corrupted version. Finally, three sentiment pre-training objectives will require the model to recover sentiment information for the corrupted version.

In Sentiment Masking, SKEP proposes to combine Sentiment Word Detection with Hybrid Sentiment Masking. The model will first pick sentiment words with the help of knowledge base and find aspect-sentiment pair from the neighbours of these sentiment words with a maximum distance of 3. For Hybrid Sentiment Masking, steps below will be followed:

- Aspect-sentiment Pair Masking. At most 2 aspect-sentiment pairs are randomly selected to masks.
- Sentiment Word Masking. Those identified as sentiment words will also be masked with a limit of less than 10%.
- Common Token Masking. Common word tokens can also be substituted with '[MASK]' if the number of tokens in step 2 is not enough.

SKEP also defines the following sentiment pre-training objectives.

- Sentiment Word (SW) prediction. This objective is designed to recover all the masked sentiment words in the sentences.
- Word Polarity (WP) prediction. Word Polarity calculates the polarity (negative or positive) of the masked sentiment token, which is critical for sentiment analysis.
- Aspect-sentiment Pair (AP) prediction. Aspect-sentiment pairs may contain more information than only sentiment words have. This objective is designed to recover all the masked aspect-sentiment pairs, providing stronger capability for SKEP.

Overall, training objective L can be described as:

$$L = L_{SW} + L_{WP} + L_{AP} \quad (1)$$

4 Experiments

In this section, experiments are conducted to show how the performance varies on different pre-trained models. In this paper, we compare the performance of SKEP (Tian et al., 2020a) with BART (Lewis et al., 2019). Our codes are based on Pytorch and Paddle framework. Meanwhile, pre-trained weights of Roberta are extracted from Hugging Face. All our experiments are run on Nvidia 3080 Laptop and Nvidia A100. We use F1, Precision, and Recall scores for evaluation

4.1 Hyper Parameters

Here, hyper parameters are shown in Table 3.

Items	Range
batch size	4
steps	5000
warm up	1000
lr	2e-5
optimizer	Adam
input length	512

Table 3: Hyper Parameters

Model	F1	P	R
BART-base	91.55	94.89	88.44
BART-large	90.65	91.97	89.36
SKEP-large	**95.26**	**95.26**	**95.26**
Public Board	94.00	-	-
Private Board	89.74	-	-

Table 4: Model Performance. The above three rows are the performance on development set, while the last two rows are our final result for Kaggle submission

4.2 Results

Table 4 shows all our experiment results. Here, a base model is a 12-layer model with a 6-layer encoder and a 6-layer decoder, while a large model has 24 layers in all, 12 layers for encoder and 12 layers for decoder. Limited to the interface of Baidu Paddle framework, we only test SKEP-large here.

It is clear that SKEP outperforms all other models in our experiments, showing its superiority in classification tasks. Compared to BART-base, SKEP-large achieves an improvement of 3.71 points in F1 score and shows better balance in precision and recall scores. As SKEP is augmented with sentiment information, it can be inferred that pre-trained weights contribute to the performance improvement in this task. This is reasonable because rumours tend to be negative so that they can confuse the public and spread fright.

Meanwhile, it is also interesting to see that BART-large achieves a lower F1 score than BART-base. The reason is that BART-large has much more weights than BART-base, which means that BART-large can better capture the patterns and fit the dataset better. However, the give dataset in our Task 1 has rather limited instances and can not represent the actual distribution of Covid-19 rumour tweets. In this case, BART-large easily becomes over-fitted on the provided dataset.

For almost the same reason, although we achieve an F1 score of 94.00% in Public Board, our model performs much poorer on the other 60% of test dataset, which shows a drop of F1 score.

5 Discussion

5.1 Kaggle Dataset Analysis

5.1.1 Rumour Topic Analysis

For comparison, we have listed the 9 hottest topics among rumoured and non-rumoured tweets below. Each figure in the Table 5 represents the percentage of the word in its corresponding category. The symbol ✗ means that the word is not among the hottest 9 topics within corresponding category, not that the word does not appear in this category. There is quite a bit of overlap between the nine most popular topics in rumoured and non-rumoured tweets since word e.g. 'death', 'people', 'pandemic' are common topics in COVID19 data. The hottest topics among rumour and non-rumour tweets are trump and case respectively. It's also intuitive that the word 'lie' appears in the trending topic among rumoured tweets.

Hot Topic	**Rumour %**	**Non-Rumour %**
trump	1.64	✗
people	1.03	0.74
death	0.60	0.83
pandemic	0.49	0.56
flu	0.41	✗
world	0.37	✗
president	0.36	✗
lie	0.34	✗
new	0.34	1.26
case	✗	1.79
positive	✗	0.50
mask	✗	0.43
health	✗	0.42

Table 5: Hot Topics Comparison

As illustrated in Figure 1, topics in rumour tweets and in non-rumour tweets are examined by the top 9 word frequency for comparing topics in these two classes. Additionally, to have a better understanding of the overall word frequency distribution, we plot the word cloud map as shown in Figure 2.

5.1.2 Hashtag Analysis

Table 6 shows 10 most popular tags among two classes. The hashtag overlap accounts 14.46% and

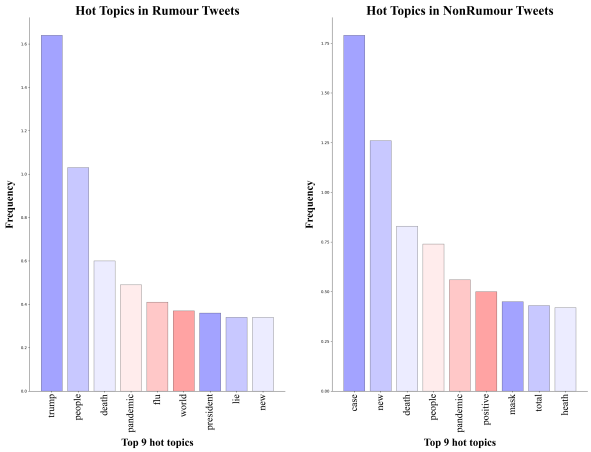

Figure 1: Hot Topic Comparison

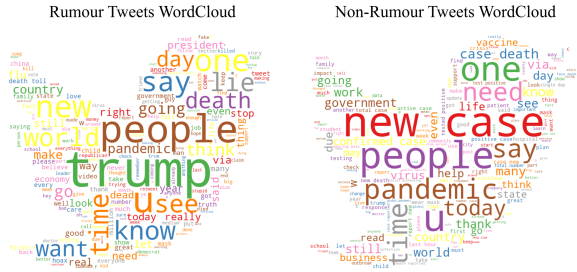

Figure 2: WordCloud Comparsion

1.79% in rumour and non-rumour tweets respectively.

5.1.3 Rumour Trends Analysis

This dataset covers a time span of 26 consecutive days (2020.8.23-2020.8.30). For the purpose of exploring the evolution of rumours over time, we divide the 26 days into four periods, consisting of three seven-day periods and one five-day period. The word frequency comparison diagram is shown in Figure 3. Most topic trends follow a similar pattern such as the word 'case', 'new' and 'death'. The highest figure is reached in the first week, followed by a drop and rise in the second and third weeks, and a decline again in the fourth. There is a significantly consecutive rise in the topic 'vaccine' in the first three weeks and a decline in the fourth. The topic 'pandemic' reached its peak in the third week (2020.08.09-2020.08.16) as well.

5.1.4 Sentiment Analysis

We applied a sentiment classifier from hugging face API [3] to each tweets. According to the Figure 4 , the proportion of negative emotions in rumour tweets is 69.6%, which is 12.2% higher than that in non-rumor tweets.

[3]https://huggingface.co

Hot Hashtag	Rumour	Non-Rumour
COVID19	✓	✓
Trump	✓	✓
coronavirus	✓	✓
GOP	✓	✗
pandemic	✓	✓
TrumpVirus	✓	✗
USA	✓	✗
MAGA	✓	✗
Russia	✓	✗
China	✓	✗
India	✗	✓
CoronavirusPandemic	✗	✓
Odisha	✗	✓
CoronaVirusUpdae	✗	✓
WearAMask	✗	✓
vaccine	✗	✓

Table 6: Hot Topics

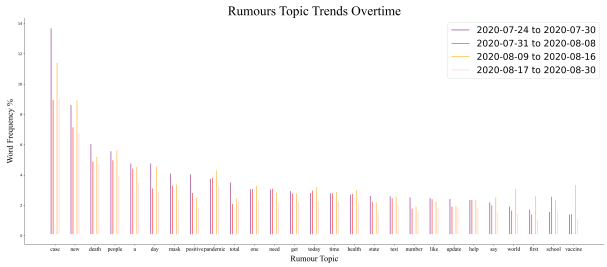

Figure 3: Topic Trends

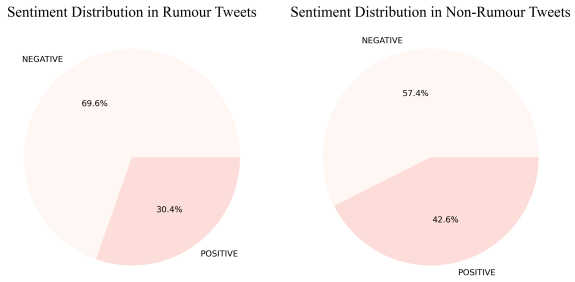

Figure 4: Sentiment Comparsion

5.1.5 User Analysis

According to Figure 5, users' verified rate for non-rumor tweets is 13.7%, however, for rumor tweets this figure is only 4.1%, which is a nearly three-fold decrease. Among the rumour Twitter accounts, CGTN is the account with the most followers. With 13892839 followers. CGTN has tweeted four ru-

mor tweets. Seven of the top 20 tweeted accounts have more than one million followers, while none of the 20 Twitter accounts with the most rumour tweets has more than 100K followers.

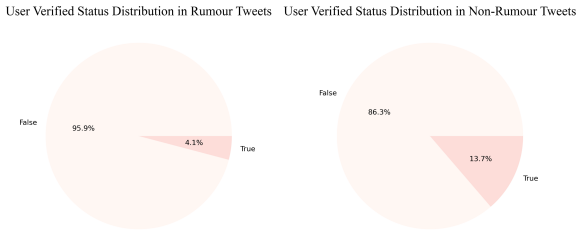

Figure 5: User Verified Status Comparsion

5.1.6 Location Analysis

Figure 6 compares the 7 countries with the most tweets. USA and India are the regions with the most rumored and non-rumored tweets, respectively.

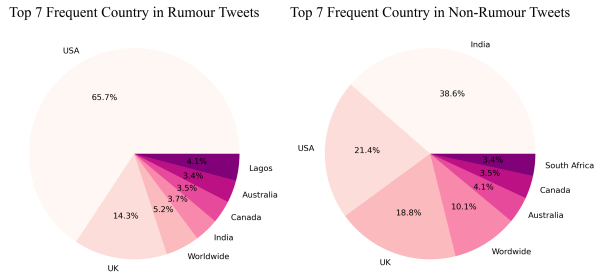

Figure 6: Location Comparsion

5.2 Provided Dataset Analysis

5.2.1 Rumour Topic Analysis

Popular rumour topics and non-rumour topics are listed in the Table 7.

5.2.2 Rumour Trends Analysis

There are 178 consecutive days in the dataset. These tweets are divided into six 30-day periods and one 28-day period. As shown in Figure 7, during the 6 months, Trump has been the hottest topic of discussion in every month. The word 'case' has been growing for the first five months and we see a slight decrease in the last month. The topic 'pandemic' increased nearly fivefold between the first and second month, peaking in the second month.

5.2.3 Sentiment Analysis

Tweets about rumours have 84.1% negative emotions, which is 7.5% higher than tweets without rumor.

Hot Topic	Rumour %	Non-Rumour %
trump	3.19	1.14
people	1.13	1.18
hoax	0.95	✗
death	0.89	1.01
lie	0.73	✗
president	0.65	✗
flu	0.60	✗
pandemic	0.56	0.48
china	0.54	✗
case	✗	0.77
virus	✗	0.57
state	✗	0.46
country	✗	0.43
new	✗	0.42

Table 7: Hot Topics Comparison

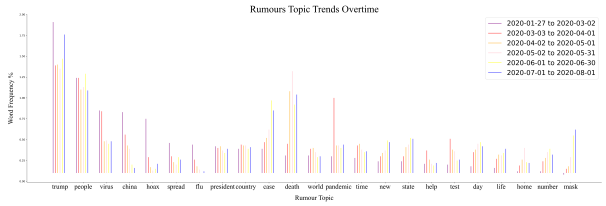

Figure 7: Topic Trends

5.2.4 Rumour Label Evolution Analysis

17 % Twitter's initial and last retweeted tweet classifications (rumor vs non-rumor) have changed.

6 Conclusion

SKEP model outperforms all other models in our experiments, achieving an F1 score of 95.26 on the dev set. We believe this is benefit from SKEP's strength in sentiment analysis, which also works in rumor analysis, possibly because rumours tend to be negative so that they can confuse the public and spread fright.

In our analysis of Covid19 data, we found that 'Trump' and 'lie' appear more frequently in rumours while 'case' appear more frequently in non-rumors. Most of the topics in rumours are decreasing over time while vaccines and pandemics are increasing. None of the rumour spreaders had a large number of followers, but some accounts with a large number of followers, such as CGTN, also posted some rumours.

References

Abdullah Alsaeedi and Mohammed Al-Sarem. 2020. Detecting rumors on social media based on a cnn deep learning technique. *Arabian Journal for Science and Engineering*, 45(12):10813–10844.

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*.

Fatima Mohammed Dito, Haleema Adnan Alqadhi, and Abdulla Alasaadi. 2020. Detecting medical rumors on twitter using machine learning. In *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, pages 1–7. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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*.

H. Tian, C. Gao, X. Xiao, H. Liu, B. He, H. Wu, H. Wang, and F. Wu. 2020a. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020b. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Somayeh Zamani, Masoud Asadpour, and Dara Moazzami. 2017. Rumor detection for persian tweets. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, pages 1532–1536. IEEE.