Rumour Detection and Analysis on Twitter

1040203 1128590 904821
COMP90042 Natural Language Processing
The University of Melbourne

1 Introduction

With the advancement of Internet technology, an increasing number of people are willing to share their thoughts and experiences online. Meanwhile, intentionally spreading rumours on social media is far easier and less expensive. It will be much more difficult for us to obtain accurate information if our social media is filled with rumours. More importantly, if a rumour becomes widely circulated and believed, it may cause society to be slow to recognise and respond to problems. As a result, research and development of a classification system capable of accurately and rapidly classifying rumours is required.

The outbreak of the COVID-19 global pandemic has sparked a wide range of opinions on Twitter since the end of 2019. People are looking to Twitter for the most up-to-date information on virus cases, treatments, and mutations. Many rumours about COVID-19, on the other hand, have been disseminated to provoke unneeded alarm and instability.

As a result, the purpose of this research is to first investigate various machine learning approaches and evaluate their performance on the twitter rumour classification problem. The report will then go on to discuss and analyse the prediction results of the best-performing model in order to determine the difference between COVID-19 rumour tweets and non-rumour tweets.

2 Related Work

To handle the rumour classification challenge, various algorithms have been developed, including classic machine learning methods and deep learning methods. Nguyen et al. presented BERTweet (Nguyen et al., 2020), the first public large-scale pre-trained language model for English Tweets, based on BERT (Devlin et al., 2019), a well-known pre-trained model in the Natural Language Processing area based on Transformers (Vaswani et al., 2017). Later, they released BERTweet-COVID19, which includes additional pre-training on the COVID-19 English Tweets corpus. Kaliyar et al. developed FakeBERT (Kaliyar et al., 2021)by integrating various Convolutional Neural Networks with the original BERT in a bidirectional training technique aimed at capturing semantic and long-distance relationships in phrases. Kumar et al. used a convolutional neural network and an ant colony filter-wrapper tuned by a Naive Bayes classifier to create a hybrid model for microblog rumour detection (Kumar et al., 2021).

3 Dataset

The University of Melbourne COMP90042 teaching team provides labelled train and development datasets, as well as unlabeled test and covid datasets. Each event in the dataset contains one or more tweet ids, with the first being the source tweet and the others being reply tweets. We get tweet objects based on these tweet ids using the Twitter API, and only keep events when the source tweet object exists and the language is English. There are two labels for labelled events: rumour and non-rumour. Finally, 1549 training events, 533 development events, 558 test events, and 15618 covid events were collected. This is a class-imbalanced classification task with more non-rumour events than rumour events. Table 1 shows that the label distributions of training data and development data are similar.

Dataset	Instance	Rumour	Nonrumour
Train	1549	318	1231
Dev	533	114	419
Test	558	None	None
Covid	15618	None	None

Table 1: Number of instances and label distribution of datasets.

4 Task 1: Rumour Detection

4.1 Preprocessing

In each event, we first concatenated the text of the source tweet with the texts of the reply tweets. We then eliminated user mentions, URLs, stop words, and emojis from the tweets because we didn't think they provided us with useful information. We then performed stemming and tokenization on the data. Following that, we employed various strategies for various models.

We applied TF-IDF (Manning et al., 2008) to process data and generated features for Logistic Regression and Naive Bayes. For the bidirectional LSTM model, we applied word embedding to transform each token into a vector. To produce the inputs id and attention mask for the BERT model, we used the tokenize method provided by BERT. The attention mask prevents the model from training the parameters for padding tokens, and the inputs id are the vocabulary IDs that map the input tokens to the model. We didn't preprocess the dataset for the BERTweet model because it can be used for processing raw tweets.

4.2 Naive Bayes

Naive Bayes (NB) (Webb, 2010) is a supervised machine learning algorithm that is based on the Bayes theorem with the assumption that features are independent. Based on the training dataset, the NB classifier estimates the prior probability for each class, then uses maximum likelihood estimation to estimate the conditional probability of each feature given each class. The posterior probability can be calculated using the Bayesian formula utilising the prior probability and likelihood value to produce the predicted result. Because we used TF-IDF as features, we used Multinomial NB as a classifier. This model serves as a benchmark for us.

4.3 Logistic Regression

A linear approach for solving binary classification problems is Logistic Regression (LR) (Cox, 1958). LR is a probabilistic discriminative model that does not require feature independence assumption because it optimises posterior probability directly. In simple terms, LR assumes that the data follows a specific distribution and then estimates the parameters of that distribution using maximum likelihood estimates. We investigated the topic of class imbalance and used class weights (CW) to train a new LR model as a contrast to the basic LR model. This is our other baseline model, just like the NB model.

4.4 Long Short-Term Memory

The Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) model is a recurrent neural network-based (RNN) (Medsker and Jain, 1999) model. Through the forget gate, input gate, and output gate, LSTM regulates the proportion of information retained by each section. In the training process for long sequences, LSTM solves the problem of gradient explosion and gradient vanishing in the RNN model. Because the text in our task incorporates source and reply tweets, and the sequences are long, we believe LSTM is a better fit than RNN. To enable the model to capture the information before and after each token, we used a bidirectional LSTM (BI LSTM) to train our model.

4.5 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. In order to capture dependencies between words, BERT leverages self-attention networks from Transformers to complete two pre-train tasks: Masked Languaged Model (MLM) and Next Sentence Prediction (NSP). The MLM task plays a critical role in overcoming the unidirectionality limitation. The NSP task is also useful for capturing sentence associations. A pre-train model is empirically more powerful than other models, which is why we explored the BERT model. We need to put a classification layer on top of the contextual representations from BERT to complete the rumour detection task.

4.6 BERTweet

BERTweet is a pre-trained BERT model on 850 million English Tweets. The BERTweet-COVID19 version we used for this report is the outcome of pre-training the original BERTweet model on a corpus of 23M COVID-19 English Tweets. As a result, BERTweet has the same benefits as BERT, but it is more adaptable to tweet-related downstream tasks, particularly for tweets about COVID-19 topics. We also set up a dropout layer to prevent overfitting by setting the outcomes of certain neurons to 0 at random.

4.7 Evaluation

Our evaluation criterion is the F1-score on the development dataset.In our task, the majority class is the rumour, whereas the minority class is the nonrumour tweet. The ratio of the two labels is 4:1. The accuracy will be quite high if the model predicts that all tweets are rumours. As a result, accuracy is a poor criteria for this task, and the F1-score is the harmonic mean of precision and recall. It takes both into account. We believe that the F1-score can better assist us in selecting models.We also evaluated the results of giving class weights during training at the same time. The evaluations on the development dataset are displayed in table 2.

Model	F1	Precision	Recall
NB	0.74	0.65	0.85
LR	0.80	0.90	0.72
LR + CW	0.81	0.84	0.78
BI LSTM	0.88	0.89	0.87
BI LSTM + CW	0.85	0.81	0.91
BERT	0.85	0.78	0.95
BERTweet	**0.94**	**0.94**	**0.95**

Table 2: Model evaluation on the development dataset

NB	LR	BI LSTM	BERT	BERTweet
0.78	0.82	0.79	0.87	**0.92**

Table 3: F1-score on the private test dataset

4.8 Result

According to table 2, with an F1-score of 0.94, the BERTweet model performed the best of all the models we evaluated. We believe that NB has the worst performance because it assumes that all features are independent of each other. The independent assumption cannot be satisfied since each word has a relationship with others. LR outperforms NB, but not as well as other models, because it does not require the independent assumption. LR is a linear classifier that struggles with non-linear problems, we believe it's difficult for LR to find a hyperplane that perfectly separates the two classes in this task. The BI LSTM considers the information in the bidirectional text, but only captures the surface bidirectional representation, and only depends on the training set, while BERT uses the Self-Attention and pre-trained models, which can not only capture the relationship in word pairs and can introduce more information from pre-trained models, Bertweet offers a pre-trained model for tweet data, so it has the best performance in our task. At the same time, we discovered that using class weights can enhance model recall while lowering precision, therefore the F1-score will not definitely improve as a result of using class weights.

The final results on the private test dataset are shown in Table 3. The results basically meet our evaluation on dev. BERTweet provides us with the highest F1-score, which is 0.92.

5 Task 2: Rumour Analysis

We classified the covid dataset using BERTweet, the best performing model on the development dataset, and analysed the natures in source tweet user, source tweet, and reply tweets for each label. There are 12187 rumours and 3431 non-rumours in covid dataset using the BERTweet.

5.1 Users Analysis

We investigated the users of the source tweet, focusing on three user attributes: "verified", "followers_count", "favourites_count", "friends_count". These four characteristics were chosen because they are more likely to relate to whether the tweet is a rumour or not. The results are in table 4.

The first feature “verified” indicates whether the user is verified or not. The percentage of rumours coming from verified users is 77.73%, while the percentage of non-rumors coming from verified users is 79.78%. Despite the similar percentages, the non-rumour user verified rate is greater. We believe that the user's verified status has little to do with rumours and non-rumors based on the results.

The second feature, "followers_count", refers to the amount of followers a user has. We computed the average number of followers. The rumour's average number of followers is 4626821, whereas the non-rumour's average number of followers is 4949139. We discovered that non-rumor users have more followers on average.

The third feature, "favourites_count", denotes the amount of users followed by the user. The average number of rumour users following is 44694, while the average number of non-rumour users following is 23937. According to the results, the number of rumour users following is higher than that for non-rumour.

The last feature “friends_count” implies the numbers of the user's friends (mutual following). The average of the rumour's friends is 1640 and the average of the non-rumour's friends is 461. The results seem that the number of friends of the user who tweets a rumour is triple of the user who tweets the non-rumour.

Feature	Rumour	Nonrumour
Verified	77.73%	**79.78%**
Followers	4626821	**4949139**
Favourites	**44694**	23937
Friends	**1640**	461

Table 4: Users Analysis

5.2 Tweet Analysis

We analyzed source tweets and reply tweets separately for our tweet analysis.

5.2.1 Hashtags

We examined the hashtags of tweets in the covid data separately by rumour and non-rumour.

In source tweets, we discovered some overlaps, such as #coronavirus, #covid19, and #breaking, which we believe is reasonable given that these are covid-related data. There are #trump, #trumppressconference, #wuhanvirus in rumour source tweets, and #lockdown, #stayhome, #wuhan in non-rumour source tweets. We believe this is consistent with our assumptions. For example, #wuhanvirus and #wuhan are two comparable terms, #wuhanvirus is associated with racism and rumour, whereas #wuhan is just associated with the covid outbreak in Wuhan.

In reply tweets, we noticed there are some hashtags in rumour retweets that involves political slogans like #maga which is the abbreviation for "Make America Great Again" and #gop which is the Republican Party in America. These hashtags could potentially incite the emotions of the public. In non-rumour retweets, hashtags are more focused on cheering and ways to contain the epidemic, such as #WHO, #precautions, #stayhomestaysafe, #lockdown and #indiafightscorona.

5.2.2 Topics

We extracted 5 topics, each having 10 topic words, using Non-negative Matrix Factorization (NMF) (Lee and Seung, 2000) to process the TF-IDF matrix of rumour data and non-rumour data, respectively.

We found that rumour-related topics in source tweets are primarily related to Trump and the White House, but non-rumour-related topics are primarily related to covid new cases, test and health.

Rumour-related topics in retweets involves potential racist words like "wuhanvirus", "originated" and political-related words like "ccp" and "pence". However, non-rumour-related topics are more focus on authenticity of information, for example, "quoted", "real", "cdc" and "article".

Rumour Topics
trump, coronavirus, says, people, donald, response, pandemic, rally, americans, china
cases, deaths, new, florida, us, coronavirus, record, reports, day, reported
covid, anime, part, trump, death, rally, died, america, hoax, people
president, trump, said, campaign, america, donald, country, first, testing, says
house, white, briefing, task, force, press, coronavirus, live, secretary, cnn

Table 5: Rumour Topic for source tweet

Nonrumour Topics
cases, new, confirmed, reported, total, reports, number, day, breaking, florida
coronavirus, says, trump, us, people, pandemic, world, health, president, house
covid, people, today, update, health, amp, pandemic, patients, test, died
positive, tested, tests, test, players, april, breaking, coronavirus, negative, people
deaths, us, total, uk, reported, italy, number, coronavirus, may, apr

Table 6: Nonrumour Topic for source tweet

Rumour Topics
covid, people, coronavirus, get, like, one, many, cases, amp, going
coronavirus, hoax, get, like, wuhan, go, would, people, pence, flu
trump, coronavirus, president, pandemic, response, americans, hoax, donald, us, amp
china, wuhan, chinese, virus, wuhanvirus, racist, world, ccp, lab, originated
deaths, death, cases, florida, us, number, new, million, numbers, flu

Table 7: Rumour Topic for reply tweets

Nonrumour Topics
covid, people, get, amp, like, one, know, mask, many, would
coronavirus, china, wuhan, people, virus, get, world, like, one, chinese
trump, president, americans, america, donald, us, administration, response, pence, testing
deaths, cases, new, death, number, reported, florida, numbers, per, total
quoted, article, tweet, read, tcm, thank, kerri, exist, real, cdc

Table 8: Nonrumour Topic for reply tweets

5.2.3 Sentiment

We used VADER (Hutto and Gilbert, 2014) to conduct rumour and non-rumour for source tweet and tweet reply sentiment analysis, respectively. The VADER model is a pre-trained rule-based sentiment analysis model that calculates scores for positive, negative, and neutral sentiment. The compound is calculated by the three sentiments. The sentiment is more positive when the compound is closer to 1, the more negative it is when the compound is closer to -1, and the more neutral when the compound is closer to 0. We discovered that, despite sentiments similar, the degree of negative sentiment in rumour source tweet is larger than that of non-rumour source tweet. In retweets, reply of rumour tweets are more negative than reply of non-rumour tweets. The difference is in compound value of retweets also proves this point, the sentiment of retweets are overall neutral indicating calm discussions.

Sentiment	Rumour	Nonrumour
compound	-0.0980	**-0.0186**
negative	**0.1339**	0.0899
neutral	0.7846	**0.8241**
positive	0.0804	**0.0856**

Table 9: Source tweet sentiment scores

Sentiment	Rumour	Nonrumour
compound	-0.3197	**-0.0733**
negative	**0.2009**	0.1568
neutral	0.6551	**0.6815**
positive	0.1426	**0.1554**

Table 10: Reply tweet sentiment scores

5.2.4 Word Cloud

Based on rumour and non-rumor source and reply tweets, we created word clouds. Word cloud (Oesper et al., 2011) is a graphic representation of words that highlights the most frequently used words. It's easy to see from the word clouds, aside from covid-related words, President Trump appears frequently in rumour, whereas phrases like government and breaking appear frequently in non-rumour.

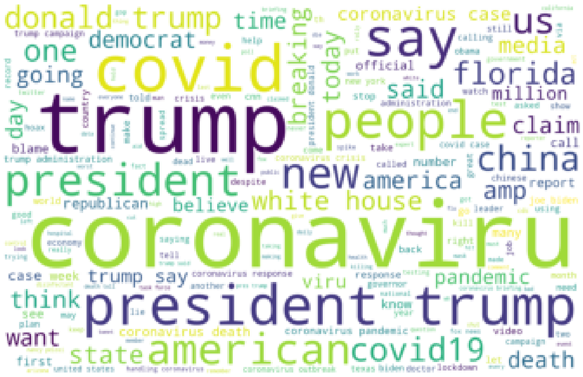

Figure 1: Rumour word cloud

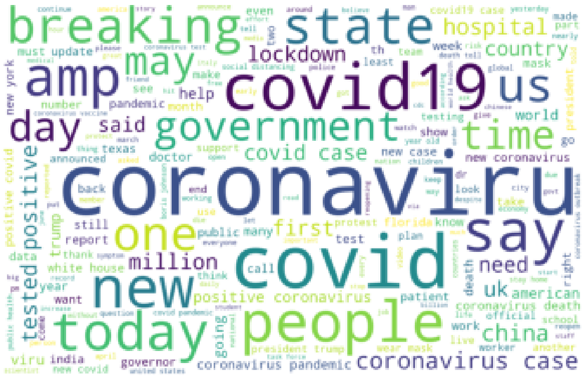

Figure 2: Non-rumour word cloud

6 Conclusion

In this project, we explore different machine learning models and compare their performance on rumour classification. Then we use the BERTweet model, which has the best performance (F1-score is 0.94), to predict on covid data and analyse the natures in rumour tweets and non-rumour tweets. We then analyze users, source tweets and reply tweets.

This exercise provides us with valuable tweet analysis experience. In the future, we'll aim to conduct more comprehensive feature engineering in the future, and integrate more features other than text, as well as try out more various models.

References

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Rohit Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80.

Akshi Kumar, Mps Bhatia, and Saurabh Raj Sangwan. 2021. Rumour detection using deep learning and filter-wrapper feature selection in benchmark twitter dataset. *Multimedia Tools and Applications*, pages 1 – 18.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

L.R Medsker and L.C Jain. 1999. *Recurrent neural networks: design and applications*. CRC Press.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Layla Oesper, Daniele Merico, Ruth Isserlin, and Gary D Bader. 2011. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source code for biology and medicine*, 6(1):7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Geoffrey I. Webb. 2010. *Naïve Bayes*, pages 713–714. Springer US, Boston, MA.

Contribution

All members contributed to this project.

1040203 was responsible for the crawling of tweet objects and the dataset preprocessing progress. In the experiment phase, he built and fine-tuned the NB, LR, BI LSTM, BERTweet model. In the rumour analysis part, he was responsible for analysing hashtags, topics, sentiments and word cloud of source tweet. He was responsible for writing report session 3, 4.1, 4.2, 4.3, 5.2.

904821 was responsible for researching rumour-detection related works in the early stage of this project. He also participated in the crawling of tweet objects and the dataset preprocessing progress. In the experiment phase, he built and fine-tuned the BERT model. In the rumour analysis part, he was responsible for analysing hashtags, topics, sentiments and word cloud of retweets. He was responsible for writing report session 1, 2, 4.5, 4.6, 5.2.

1128590 was responsible for the crawling of tweet objects and the dataset preprocessing progress. He also help built the BERT model. He was responsible for writing session 4.7, 4.8, 5.1, 6.