Rumour Detection and Analysis on Twitter

COMP90042 Project Report (Student ID: 830683)

Abstract

Microblogs such as Twitter is now growing to be a platform to share news. However, due to its freely available sign-ups and lack of curation, it is difficult to guarantee the validity of information. To combat this, we use BERT with Feedforward Neural Network to build a rumour classifier system. The resulting classifier is used to label a COVID-19 related dataset. Based on the labels, general topics, keywords, and key hashtags are extracted to gain insights on characteristics that constitute rumour and non-rumour tweets. It was found that a) non-rumour tweets have a more diverse set of topics, b) health advice is more commonly found in non-rumour tweets, and c) rumour tweets tend to include more negative words.

1 Introduction

In its early inception, social media is mainly used to share about people's lives with their friends and family. Nowadays, social media is a burgeoning news sharing platform, also known as microblogging [1]. One example of such social media is Twitter, which has a monetizable Daily Active Usage (mDAU) of 192 million users worldwide in Q4 2020 [2]. This has helped to propel information sharing, especially to different parts of the world. However, this also means that misinformation is prolific due to lack of curation and its large outreach. Therefore, it is essential to have a solid rumour detection system to help mitigate the spread of incorrect information.

The aim of this project is to build a rumour classification model that can label tweets to be *rumour* or *non-rumour*. The focus will be on text analysis of tweets. The model will then classify an unlabelled dataset of COVID-19 related tweets. The labels are further used to determine the general characteristics of rumour and non-rumour tweets, such as common topics, keywords, and key hashtags. Similarly, the analysis will be focused on text analysis of tweets and hashtags.

2 Methodology and Data

All codes used in this project is run on Google Colab [3] to harness its computing capabilities.

2.1 Task 1: Rumour Classification

Our aim is to classify each tweet into either a *rumour* or a *non-rumour* tweet where a rumour tweet is neither necessarily untruthful nor truthful. Therefore, this task can be modelled as a binary classification problem. In terms of features used, we focus on performing a text analysis on parent tweets $P = \{p_1, p_2, \ldots, p_i\}$ and their corresponding reply tweets $R_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,n}\}$.

The text analysis is performed using BERT. BERT is a pre-trained language model that results in a bidirectional language representation. It can be fine-tuned for downstream tasks such as classification or language generation [4]. BERT is known for its state-of-the-art performance and is the predecessor to many cutting-edge language processing technologies [5][6].

The training dataset contains tweet sets, each holding a record set of parent tweet and its replies. Each thread is labelled as either *rumour* or *non-rumour*, respectively encoded as 1 and 0. Tweet texts are pre-processed to remove usernames (words that start with '@'), '#' in hashtags, and URLs.

Following this, the texts are tokenized using AutoTokenizer based on BERT base uncased model since the model is intended for use in a downstream classifier task. The parent tweet is inputted as the first sentence and the concatenated reply tweet is inputted as the second sentence. Padding is based on *'max_length'*, set as its maximum allowed value of 512, and longer texts are truncated. The PyTorch [7] tensor of encoded, tokenized tweets (token ids) is saved along with the attention mask and seg ids (denoting whether a token belongs to parent (0) or replies (1)). The information is stored as a class object which is then loaded using PyTorch's DataLoader.

The classifier uses the same BERT base (uncased) pretrained model for the binary task classification. Token ids, attention mask, and seg

ids are fed into the BERT model. Afterwards, the resulting contextualised embedding of [CLS] is passed to a feedforward network with an input dimension of 768 and a single scalar output. The classifier is GPU-enabled for better performance.

A binary cross-entropy loss function is used as it is a binary classification. Adam optimizer is used as it performs well with problems requiring large data such as text analysis [8].

During training, the data from DataLoader is fed in batches of 16 (Table 1). This is based on the recommendation of BERT's authors [4] and resource constraints from Google Colab.

Table 1.
Hyperparameters used in BERT model.

Batch size	16
Learning rate (Adam)	2e-5
Number of epochs	2

The batches are fed into the classifier to obtain loss. Parameter gradients are computed based on the loss, which is used to update the parameters after each batch training. After all batches have been trained, the model is tested against the development set. This constitutes as one epoch. The best development accuracy from all epochs is saved as the final model.

2.2 Classifier Dataset

We are provided with a total of 5802 tweet sets, comprising parents and the reply tweets. 80% is used as training data while 20% is split evenly between dev and test set (Table 2). The ratio of rumour to non-rumours is approximately 1:2.

Table 2.
Rumour distribution among train, dev, and test set.

	Total count	**Rumour**	**Non-Rumour**
Train	4641	1583	3058
Dev	580	187	393
Test	581	No label	No label

2.3 Task 2: COVID Analysis

The aim of this task is to gain insights on COVID-19 related tweets and general characteristics of both rumour and non-rumour tweets. First, the COVID-19 Twitter dataset is labelled using the rumour classifier. Separately, the parent texts are concatenated with their replies and pre-processed to remove usernames, '#' in hashtags, and URLs. Afterwards, the texts are lemmatized using NLTK's WordNetLemmatizer [9]. The lemmatized texts are tokenized and further cleansed against an extended NLTK stopwords encompassing distinctive COVID-19 terms such as *covid* and *coronavirus*.

We focus on gaining insights that are text-oriented such as tweet texts and hashtags used. First, we aim to perform topic analysis to identify clusters of common topics alongside the keywords for rumours and non-rumours. To do this, we use Latent Dirichlet Allocation (LDA) model available from Gensim [10]. LDA is a generative probabilistic model that aims to bucket discrete data such as text corpora into broader topics based on probabilities [11]. The default values of passes=10 and chunksize=100 is used. To determine the optimum number of topics, k, LDA is initially run using multiple values of k and their coherence points are observed using Gensim's CoherenceModel. Before being fed into LDA, the tokens are converted into a bag-of-words using Gensim dictionary.

To complement the LDA model, an analysis on hashtag frequencies is conducted using TF-IDF measuring tool in Scikit-learn [12]. A high TF-IDF number indicates a strong relationship between a word and its document [13], making it a keyword for the document. In this part, hashtags are compiled to create two documents: rumour and non-rumour hashtag. The two documents are combined to create a hashtag corpus, then TF-IDF computation is performed on the corpus.

Since the dataset is labelled using a model with approximately 82% accuracy, it can be assumed that the resulting analysis is fairly accurate, barring some expected inaccuracies.

2.4 COVID-19 Twitter Data

The dataset provided contains 254681 tweets comprising of 17458 parent and 237223 reply tweets. Figure 1 shows the total distribution over time. It is apparent that despite some early traffics, most COVID-19 tweets come starting March 2020. This could be related to the declaration of global pandemic by WHO [14]. This project focuses on tweet text and hashtag features of the dataset. Other features are not considered in the interest of scale and resources saving.

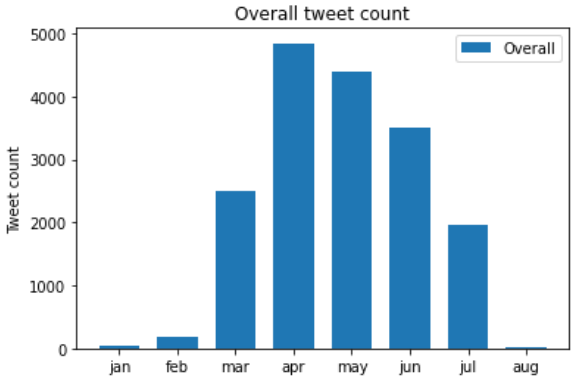

Figure 1: COVID-19 related tweet distribution

3 Results and Analysis

For both tasks, an attempt to perform MaxMatch and reversed MaxMatch algorithms to split hashtags into words was made, however due to the hefty processing power and RAM required, the program repeatedly stopped executing midway. Therefore, the decision was to keep hashtag texts.

3.1 Rumour Classification

The performance of BERT + Feedforward Neural Network against test and dev data is detailed on Table 3. Four feature combinations were considered for each dataset: (1) only parent tweet, without (a) and with pre-processing (b), and (2) parent + reply tweets, without (a) and with pre-processing (b).

Table 3.
Classifier performance against test and dev set. 1. Tweet only (a, b), 2. Tweet + replies (a, b); a. without pre-processing (raw), b. with preprocessing

	Train Set			Dev Set
	F1-score	**Recall**	**Precision**	**Accuracy**
1.a.	0.7721	0.8830	0.6860	0.8665
1.b.	0.8	0.8085	0.7917	0.8818
2.a.	0.8272	0.8408	0.8144	0.875
2.b.	0.8141	0.8617	0.7714	0.8783

In general, F1-Score in train set result is positively correlated to the accuracy in dev set. The classifier performs better in dev set compared to train set, which is expected since the dev set is used to fine tune the classifier.

The recall rate shows that the classifier can correctly predict between 80-86% of the relevant dataset (Recall = $\frac{TP}{TP+FN}$). However, the rate at which the classifier correctly predicts rumours (Precision = $\frac{TP}{TP+FP}$) is slightly lower at 68-81%. Overall, the classifier has a good balance of Recall and Precision based on the F1-Score (77-82%, where F1-Score = $\frac{2TP}{2TP+FP+FN}$).

For the parent-only model, the classifier performs better with pre-processing. However, the parent-reply model has a better F1-Score when it is not pre-processed. A likely explanation is since it has more tweets being incorporated, the noise provides additional contexts to BERT instead of polluting it. Therefore, model with the best F1-Score is used for topic analysis (Section 3.2).

The result from final evaluation further solidifies that the raw parent + reply tweets model yields the best performance (F1: 0.82490, Recall: 0.8515, Precision: 0.8).

3.2 Topic Analysis

The classifier predicted 91.6% of the COVID-19 tweets to be non-rumour (Table 4). Hashtags are more proliferate among non-rumours compared to rumours. The distribution of records over time generally follows the same pattern as the overall trend (Figure 2).

It was found that topic coherence is optimal when k is between 60 and 70 (Figure 3). It was decided to pick k = 60 for analysis as at that point, coherences for both rumour and non-rumour are trending upwards.

Table 4.
Rumour and hashtag distribution in COVID-19 dataset

	Rumour	**Non-rumour**	**Total**
# of records	1466	15,992	17,458
# of hashtags	5560	175,356	180,916
Hashtags per record (avg)	3.76	10.97	10.36

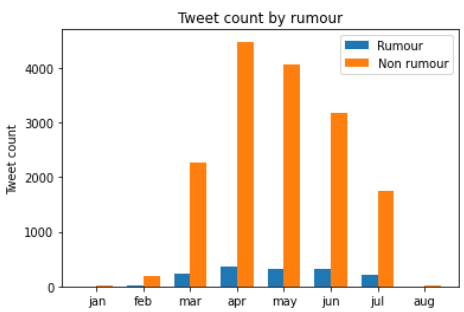

Figure 2: COVID-19 rumour distribution

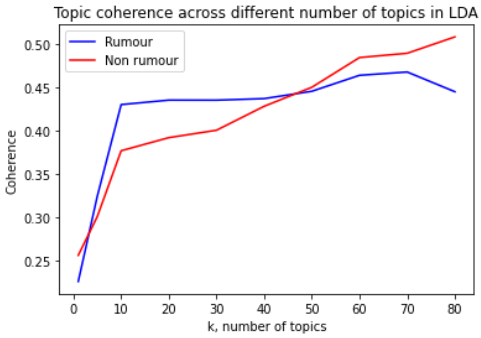

Figure 3: Variation of coherence score against number of topics in LDA

Table 5.
Comparison in keywords for mutual topics.

Mutual key topics	Non-rumour keywords	Rumour keywords
Social distancing	'gathering', 'peaceful', 'stayhome',	'deceased', 'london', 'documentary'
Masks	'positive', 'science', 'respect'	'enterprise', 'imprisonment', 'coronavirusuk'
Testing	'contact', 'tracing', 'available', 'free'	'hospital', 'households', 'virus', 'symptoms'
China	'pandemic', 'global', 'outbreak',	'lab', 'theory', 'ccp', 'cell', 'alleged'

Table 6.
Top 10 key hashtags for each label based on TF-IDF.

	Top 10 key hashtags
Rumour	'trump', 'trumpvirus', 'wuhanvirus', 'maga', 'trumpliesamericansdie', 'china', 'covidiots', 'usa', 'trumpgenocide', 'trumpownseverydeath'
Non-rumour	'wuhanvirus', 'breaking', 'trump', 'china', pandemic', 'stayhome', 'trumpvirus', 'fakenews', 'florida','secondwave'

Relevant key topics from rumour and non-rumour tweets are extracted with their keywords from the LDA result. Mutual topics are identified and inspected (Table 5). Upon inspection, keywords for non-rumour tweets tend to carry neutral to positive connotation (*peaceful*, *stayhome*, *respect*, *free*) while its counterpart tend to be more negative and speculative (*deceased*, *imprisonment*, *alleged*). Other common topics include politics (*trump*, *boris*, *government*), status updates (*toll*, *case*, *cdc*), and vaccination (*vaccine*, *development*).

In non-rumour tweets, broad topics include health news (*dr fauci*, *medicare*, *frontline*) and other affairs impacted – but not directly relate – by the pandemic (*education*, *economy*). Health advice is also more prevalent in non-rumour tweets.

In rumour tweets, conversations regarding current affairs are more centred around politics. Some keywords are unique to rumours, such as *chloroquine*, *alleged*, *scam*, *law*, and *mongering*.

3.3 Hashtag Analysis

Due to COVID-19 specific hashtags being high in TF-IDF score for both labels, hashtags that include *coronavirus* and all variants of *covid_19* (e.g., *covid-19*, *covid19*) are excluded from the key hashtag list (Table 6).

We observe that 7 out of 10 key hashtags in rumours are related to Trump. They are also more opinionated and negative (*genocide*, *covidiots*, *death*) than their non-rumour counterparts. Additionally, more topics such as pandemic news (*wuhanvirus*, *china*, *pandemic*), advice (*stayhome*, *secondwave*), and other news (*breaking*, *fakenews*, *florida*) are captured among non-rumours. This supports the previous observation in Section 3.2, where (a) rumour keywords tend to be more negative, and (b) conversations in rumour tweets are centred around politics as opposed to non-rumours where topics are more diverse.

4 Conclusion

We explored different models that can be used to build a rumour classifier using an ensemble of BERT and Feedforward Neural Network. The model was used to classify COVID-19 related tweets and a further text analysis on the tweets, including hashtags, was performed to gain insights on characteristics that constitute rumours vs non-rumours.

References

[1] Hermida, A., 2010. Twittering the news: The emergence of ambient journalism. *Journalism practice*, *4*(3), pp.297-308.

[2] Twitter, Inc. (2021). Q4 and Fiscal Year 2020 Letter to Shareholders. Available at: https://s22.q4cdn.com/826641620/files/doc_financials/2020/q4/FINAL-Q4'20-TWTR-Shareholder-Letter.pdf (Accessed: 14 May 2021).

[3] Bisong, E., 2019. Google colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform (pp. 59-64). Apress, Berkeley, CA.

[4] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[6] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[7] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

[8] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[9] Loper, E. and Bird, S., 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

[10] Radim Rehurek, and Petr Sojka 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). ELRA.

[11] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, pp.993-1022.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.

[13] Ramos, J., 2003, December. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).

[14] Cucinotta, D. and Vanelli, M., 2020. WHO declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, *91*(1), p.157.