BERT-based Rumour Identification and Analysis for Twitter Posts

Student ID: 1133751
COMP90042 Natural Language Processing
The University of Melbourne

1 Introduction

With an increase in the adoption of social media as a news source, it has become very easy for villains to share false information with a large audience. This can lead to the spread of so-called ”fake news” and rumours that may manipulate the public’s opinion. To combat this issue, there is a need for automated solutions to perform rumour identification.
Identifying rumours is a challenging problem because of their dynamically evolving nature and the ambiguity concerning what should be considered a rumour and what not (Adriani, 2019). This makes the task of automatically identifying rumours challenging and requires the use of elaborate methods to differentiate rumours from non-rumours.

2 Related Work

In recent years, the topic of rumour identification and analysis has attracted significant attention and has been subject to studies in shared tasks such as RumourEval 2017 (Derczynski et al., 2017) and RumourEval 2019 (Gorrell et al., 2019). Previous work on the topic of rumour identification for tweets can be divided into four major categories: *text-based*, *feature-based*, *propagation-based* and *stance-based* methods.
Text-based methods leverage the textual contents of the tweets. Tweets often have unique characteristics with regard to their language and syntax which can be used for rumour detection. An example of this method is Zhou et al. (2019), who used a GRU-based architecture to perform rumour detection at an very early stage of the propagation.
Feature-based methods extend this approach to also include non-textual features such as user profile data for rumour detection. In Enayet and El-Beltagy (2017), word sentiments and phrases with support & denial terms were used for rumour detection, while in Liu and Wu (2020), a variety of both text and user-based features from users’ text response and profiles were extracted before training a CNN classifier for rumour identification.
Propagation-based methods aim at incorporating the information propagation path for rumour classification. A recent approach was proposed by Ma et al. (2018) where an extended tree-structured recursive neural network (RvNN) was used to model information propagation. In Liu and Wu (2018), a novel architecture that uses Recurrent and Convolutional Neural Networks to classify propagation paths of events on Twitter was used. Ma and Gao (2020) were able to incorporate user interactions into the propagation classification via a tree transformer model.
Stance-based methods try to determine the rumour stance of a post with respect to the previous tweet and the source tweet. A system that leverages BERT for stance identification was published by Fajcik et al. (2019). Another architecture that uses a combination of a CNN and BERT architecture was presented by Tian et al. (2020).

3 Dataset

The task of the COMP90042 2021 project was to develop a rumour detection system and analyze the nature of rumours that are being propagated on Twitter. The dataset for this task was published by the COMP90042 teaching team and consists of a set of source tweets and their replies (incl. corresponding metadata) that has been extracted from the Twitter API. The training set contains a total of 4641 events which have been labeled as either *RUMOUR* or *NON-RUMOUR*. To evaluate the performance of the system, an additional development set has been made available.
The training set is imbalanced and the majority of the tweets (approx. 66%) belong to the non-rumour class, whereas the remaining data (approx. 34%)

belongs to the rumour class.

4 Experimental Setup

All systems used in this research were implemented in Python and make use of the Transformers library (Wolf et al., 2020) which provides access to various Transformer-based architectures such as BERT, RoBERTa, and DistilBERT.

The experiments reported in this paper were performed on a VM instance within the Google Cloud Platform running Debian 10 with a NVIDIA Tesla K80 GPU.

5 Task 1 - Rumour Detection

The objective of task 1 was to build a binary classifier that can reliably predict whether a given tweet represents a rumour or not.

For this task, we have implemented three classification systems: A BERT-based implementation that uses the textual representation of tweets (we refer to this architecture as *PureBERT*) and an extension of this architecture that combines the textual features with tabular data (we refer to this architecture as *MultimodalBERT*).

In addition to that, a third system which is built on a pre-trained language model for English Tweets (Nguyen et al., 2020) has been implemented. We refer to this model as *BERTweet*.

5.1 PureBERT

The *PureBERT* system leverages the BERT language model that was published by Devlin et al. (2018). BERT was pre-trained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia and leverages the transformer architecture to provide contextualized representations for downstream tasks.

The *PureBERT* model used in this research was implemented using Tensorflow (Abadi et al., 2016) and leverages the textual contents of the source tweets and as well as of their corresponding replies / retweets for rumour classification.

Pre-processing routine

Prior to training the models on the given data, we have employed the following pre-processing procedure to clean the data and remove any Twitter-specific tokens:

1. For every Twitter event chain in the dataset, extract the source tweet text and corresponding replies and concatenate them
2. Remove URLs and user mentions from the tweet texts
3. Convert tweet texts to lower-case

For our experiments with *PureBERT*, we have used a variety of pre-trained BERT models that have been made available on the HuggingFace Model Hub [1]. The following models were used for this task:

- `bert-base-uncased`
- `bert-large-uncased`
- `talkheads_ggelu_bert_en_base`[2]

The models `bert-base-uncased` and `bert-large-uncased` refer to the models that were published in the original BERT paper (Devlin et al., 2018), while `talkheads_ggelu_bert_en_base` refers to an improved BERT architecture proposed by Shazeer et al. (2020) and Shazeer (2020).

To provide input to the *PureBERT* model, we first tokenize the input tweets via the `BERTTokenizer`. For each tokenized input text. we construct the following:

- **input ids:** a sequence of integers identifying each input token to its index number in the *PureBERT* tokenizer vocabulary
- **attention mask:** a sequence of 1s and 0s, with 1s for all input tokens and 0s for all padding tokens

We subsequently fine-tune the pre-trained BERT models on the given input for 7 epochs with AdamW optimizer (Loshchilov and Hutter, 2017) using a learning rate of $3e^{-5}$, an epsilon value of $1e^{-8}$. We set the weight decay parameter to $1e^{-2}$.

5.2 MultimodalBERT

The *MultimodalBERT* model is built using an experimental framework called Multimodal-Toolkit which has been made available on Github[3]. It allows to incorporate numerical and categorical features for downstream classification.

The textual data used by the *MultimodalBERT* system has been subject to the same pre-processing routine as described in section 5.1. In addition to that, a variety of hand-crafted context metadata

[1] HuggingFace Model Hub `https://huggingface.co/models`

[2] Pre-trained model from Tensorflow Hub `https://tfhub.dev/google/collections/transformer_encoders_text/1`

[3] Multimodal Transformers Repository `https://github.com/georgian-io/Multimodal-Toolkit`

Tweet-level features	**User-level features**
Number of retweets	Number of total user posts
Number of favorites	Number of liked tweets
Occurrence of question mark	Number of followers
Occurrence of URLs in tweet	Number of followings
Number of embedded URLs	User verification status
Occurrence of media in tweet	Presence of geo-location

Table 1: Metadata features for *MultimodalBERT*

features have been produced.

Metadata Features

The context metadata used by *MultimodalBERT* can be categorized into tweet-level and user-level features. This approach was inspired by (Gao et al., 2020) and is based on the observation that rumours may have different properties than non-rumours (e.g. rumours may be more likely to include links with unverified information). The full set of features that have been extracted is described in Table 1. In our experiments, we have used *MultimodalBERT* with the `gating_on_cat_and_num_feats` configuration and trained the model for 7 epochs using the default configuration settings.

5.3 BERTweet

The *BERTweet* model was published by Nguyen et al. (2020) and has been pre-trained on a corpus of 850M English Tweets. We put forward the hypothesis that the language that is used in tweets is fundamentally different from traditional text in terms of length and the use of informal language. Hence, a language model that is pre-trained on a large corpus of tweets should achieve a better performance compared to a generic language model pre-trained on more ”formal” language like Wikipedia.

We have implemented the *BERTweet* model using the PyTorch framework (Paszke et al., 2019). Since BERTweet comes with its own tokenizer `BertweetTokenizer` that supports raw Twitter data, no pre-processing has been applied to the tweet texts at all.

To provide input to the *BERTweet* model, we first tokenize the input tweets via the tokenizer. For each tokenized input text, we again construct *input ids* and *attention masks* as described in section 5.1.

We subsequently fine-tune the pre-trained BERTweet model on the given input for 7 epochs with AdamW optimizer (Loshchilov and Hutter, 2017) using a learning rate of $5e^{-5}$ and an epsilon value of $1e^{-8}$. We set the weight decay parameter to $1e^{-2}$.

Development performance

Model / Features	Precision	Recall	F1 Score
PureBert$_{\text{base}}$	76.47	79.73	75.21
PureBert$_{\text{large}}$	77.14	86.17	81.41
PureBert$_{\text{talking heads}}$	76.77	80.85	78.76
BERTweet	78.68	82.89	80.73
MultimodalBERT	72.44	87.63	79.32

Final evaluation performance

Model / Features	Precision	Recall	F1 Score
PureBert$_{\text{base}}$	83.52	80.85	82.16
PureBert$_{\text{large}}$	81.25	82.98	82.11
PureBert$_{\text{talking heads}}$	85.39	80.85	83.06
BERTweet	86.17	86.17	**86.17**
MultimodalBERT	70.91	82.98	76.47

Table 2: Evaluation scores in %

5.4 Results

The results of the proposed models evaluated on the development set and the COMP90042 CodaLab competition are shown in Table 2. From the evaluation scores, it is evident that increasing the number of layers and number of parameters enhances the performance of BERT (*PureBert$_{base}$* vs. *PureBert$_{large}$*). We were able to further improve the performance of BERT with respect to the F1 score by leveraging an updated BERT architecture *PureBert$_{talking\ heads}$*. The *MultimodalBERT* model performed worst in our experiments. This may indicate that the hand-crafted features used by this model were not sufficient indicators for rumours. The model using the pre-trained *BERTweet* model outperforms all other methods.

The above results support the hypothesis that the language that is used in tweets is fundamentally different from regular natural language and hence a custom model fine-tuned exclusively on tweets is more suited for the classification of Twitter posts than a language model trained on regular English Language.

6 Task 2 - Rumour Analysis

To understand the nature COVID-19 rumours and how they differ to their non-rumour counterparts, we have first used our best-performing system *BERTweet* to classify the COVID-19 dataset. We subsequently investigated the characteristics of COVID-19 rumours and non-rumours with regard to their topics, hashtags and associated sentiment.

	Rumour topics	**Non-rumour topics**
1	*covid people virus nigeria state amp time health follow let*	*covid people amp like go good help think time thank*
2	*case new total recovery confirm report number bring today day*	*trump american president coronavirus amp america lie response hoax donald*
3	*death number toll report cause figure coronavirus count covid week*	*coronavirus people like spread think pandemic say know go die*
4	*coronavirus spread china man corona lockdown think rule need good*	*death die number covid cause rate flu people count toll*
5	*test positive player day result member week testing covid negative*	*china wuhan chinese world virus wuhanvirus ccp country lab pandemic*

Table 3: Topics of COVID-19 rumours and non-rumours (based on NMF)

6.1 Topics discussed in rumour and non-rumour tweets

To explore the topics that are being discussed in the tweets, a TF-IDF matrix was constructed on the given tweet texts. Subsequently, Non-Negative Matrix Factorization (NMF) was used to extract topics out of the TF-IDF matrix. Using NMF is beneficial because it decreases the impact of high-frequency words and hence helps us to obtain more specific topics.

Table 3 gives an overview of the top-5 topics discussed in rumour and non-rumour tweets. We can observe that rumour tweets are predominately referring to topics that are rather difficult to verify such as remote locations (e.g topic 1: *virus nigeria*) or single persons (e.g. topic 5: *test positive player*). Non-rumour events seem to refer to mostly fact-based topics that are easy to verify, such as topic 4: *death number covid*.

6.2 Hashtag usage in rumour and non-rumour tweets

With regard to the usage of hashtags, we observed that there is large overlap between rumour and non-rumours (e.g. hashtags like #covid, #coronovirus, #trump are present in both classes). Nonetheless, we also identified several differences between the classes: Hashtags like #hydroxychloroquine or #breaking are frequently being used by rumour-spreading tweets, while hashtags like #trump2020 or #biden2020 are mostly present in tweets that are attributed to the non-rumour class.

6.3 Expression of sentiment in rumour and non-rumour tweets

In order to evaluate whether rumour and non-rumour tweets express different sentiments, we leveraged VADER (Valence Aware Dictionary and sEntiment Reasoner) from the NLTK library (Bird et al., 2009). VADER is a lexicon and rule-based sentiment analysis tool that can be used to predict a sentiment score ranging from -100% for negative, 0% for neutral and 100% for positive sentiment. We applied sentiment scoring on the source tweets and reply tweets for both rumour and non-rumour topics. The resulting (aggregated) scores are shown in Table 4.

Rumour tweets	Source tweets	Reply tweets
Compound	-9.37%	-15.02%
Positive	6.50%	5.50%
Neutral	82.26%	87.05%
Negative	9.50%	7.45%
Non-rumour tweets	Source tweets	Reply tweets
Compound	- 2.24%	-23.05%
Positive	8.71%	7.33%
Neutral	81.37%	83.16%
Negative	9.39%	9.49%

Table 4: Sentiment scores

The predicted compound sentiment for both the rumour and non-rumour events is negative, which indicates that COVID-19 is predominantly associated with negative sentiment. The source tweets of rumours are also significantly more negative than non-rumours with a compound sentiment score of -9.37% vs. -2.24%.

It is also interesting to observe that the reply tweets of both rumours and non-rumours are significantly more negative that the source tweets (-15.02% and -23.05% respectively). This indicates that a huge proportion of messages that are being shared in online communities may be of ”toxic” nature.

7 Conclusion

In this paper, we have shown three rumour detection systems that make use of the BERT architecture. We studied their performance characteristics and used these systems to participate in the COMP90042 CodaLab competition where we achieved an F1-score of 86.17% with our best-performing system *BERTweet*. We used this systems to analyze tweets related to COVID-19 and to study the nature of rumours propagated on Twitter.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

Roberto Adriani. 2019. The evolution of fake news and the abuse of emerging technologies. *European Journal of Social Sciences*, 2(1):32–38.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*.

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*.

Omar Enayet and Samhaa R El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 470–474.

Martin Fajcik, Pavel Smrz, and Lukas Burget. 2019. But-fit at semeval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jie Gao, Sooji Han, Xingyi Song, and Fabio Ciravegna. 2020. Rp-dnn: A tweet level propagation context based deep neural networks for early rumor detection in social media. *arXiv preprint arXiv:2002.12683*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yang Liu and Yi-Fang Brook Wu. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jing Ma and Wei Gao. 2020. Debunking rumors on twitter with tree transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5455–5466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. 2020. Talking-heads attention. *arXiv preprint arXiv:2003.02436*.

Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. 2020. Early detection of rumours on twitter via stance transfer learning. In *European Conference on Information Retrieval*, pages 575–588. Springer.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, and Sam Shleifer. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. 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 1614–1623.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.