NANYANG TECHNOLOGICAL UNIVERSITY

Wee Kim Wee School of Communication and Information

H6751 – Text and web mining

Group Project:

COVID-19 Misinformation Detection

Submitted By

Chen Xingyu (G1901287H)

Ho Chong Howe Jerry (G2002332C)

Sun Cheng (G2002908C)

Introduction

Misinformation and rumours have been spreading in tandem with the COVID-19 outbreak. Dubbed by international bodies as an 'infodemic' (Zarocostas, 2020), the spread of misinformation during an outbreak can have serious consequences on a society, ranging from panic buying to sinophobia (Hall et al., 2020; Wong, 2020).

By design, fake news are intended to attract readers and they often employ 'clickbait' in the crafting of their headlines (Bourgonje et al., 2017). These clickbait titles are often stylistically different as compared to actual news headlines and as such, it is possible for the use of language models to discriminate between real news and fake news. As fake news can be socially problematic especially during a pandemic situation, a generalisable model that can automatically detect such news would be more necessary during this time. The automated detection of COVID-19 fake news is relevant in assessing and filtering potentially deceptive news from real news.

Transformer models such as BERT have revolutionised work done in the natural language processing field by improving performance on multiple NLP tasks (Devlin et al., 2018). Such models have also been used in the context of fake news detection (Baruah et al., 2020; Khan et al., 2019). However, there has not been as much work done on the detection of fake news in the current COVID-19 context due to the recency of the event.

To address this, we developed and fine tuned a COVID-19 fake news detection model derived from the BERT (Bidirectional Encoder Representations from Transformers) architecture for this report. Model training was performed on news headlines that were classified as True news and False news. In order to test the robustness of the methods of the proposed method of using BERT for fake news detection, we compared BERT against a multinomial Naive Bayes classifier and a Logistic Regression classifier in terms of their training accuracy as well as their performance on test datasets.

Data

Currently, there are a few sources of information describing COVID-19 -related fake news. For this paper, the following data sources were used:

Training dataset

The first dataset (i.e. FakeCovid) contained 7623 multilingual fact-checked news articles, which were labelled to 11 categories (Shahi & Nandini, 2020). The categories were manually annotated by the scholars or referenced to fact-checking websites (e.g. Poynter and Snopes). The dataset was collected from Apr 1 to July 1 2020 and was downloaded from the author's GitHub page. Only English records were extracted for this report.

The other dataset (i.e. CoAID) contained records from multiple sources and in multiple formats (e.g. URLs, videos, pictures, text and tweet ids). The dataset included both picture information and text information (Cui & Lee, 2020). The dataset was collected up to Jul 1 2020 . For this project, we only

extracted text effective records, excluding the records with ‘Error 404’. Due to the uncertainty of online information from the URLs, some of the records are currently no longer available or changed. Web scraping was operated to collect those effective records again and invalid records were omitted.

The training dataset was created from aggregating the CoAID and the FakeCovid datasets. They contain information articles about COVID-19 from June 2020 and July 2020. Most records in FakeCovid dataset were misinformation, so the records were combined with CoAID dataset to make a balanced dataset. The original features include id, content, title and label, but we decided to not use ‘content’ as a feature as the model already works well enough with the title.

Initially, there were many labels that indicated that the record was fake news. Some of these labels include: Misleading, mostly false, false, labeled satire and many more. As this is a binary classification problem, it is required of us to convert these labels all to a single class which represents fake news, which we eventually did. As for real news, all of it was labeled True, which we converted to ‘Real news’ as to avoid confusion.

Testing dataset

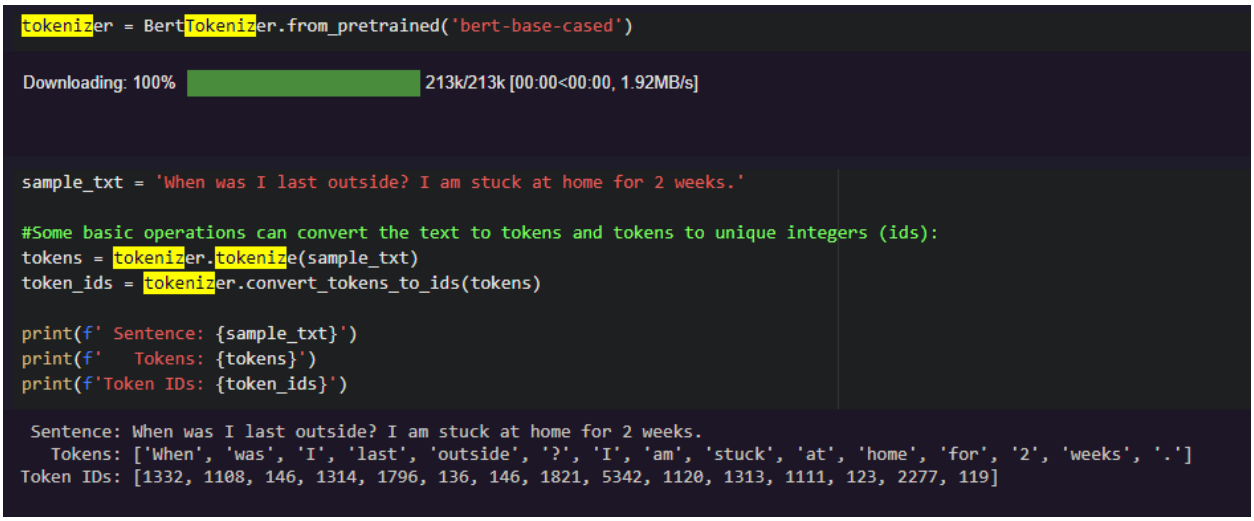
Finally, to test the robustness of the detection model, a separate testing dataset was manually created by scraping the Singapore Ministry of Health website https://www.moh.gov.sg/COVID-19/clarifications , which posted clarifications to various kinds of fake news concerning COVID. The features in this dataset are: title, id and label, which means no preprocessing has to be done as the features are the same as what is being fed into the model. The dataset was balanced with real news taken from various news sources such as CNA, Straits Times etc. This dataset contains 27 real news and 28 fake news. While splitting our first dataset into train and test dataset could also be done to verify the effectiveness of the models, but we wanted records from another source to verify the robustness of our model. This dataset contains 27 real news and 28 fake news.

Data preprocessing

As the label is not encoded yet (Fake news vs real news), we carried out label encoding where the real news was labeled as 1, and fake news labeled as 0. Our training dataset consists of roughly 5200 records - 2845 fake news and 2384 real news, which is quite balanced. A balanced dataset is important as we want our models to see sufficient positive and negative cases to improve its predictive power.

```
Name: class, dtype: int64
Fake news    2845
Real news    2384
Name: label, dtype: int64
```

Lastly, the title feature was tokenised using the BertTokenizer for the BERT model.



Methods

We used the pre-trained base cased BERT model and fine tuned it on the training set. To perform validation, we split the dataset using scikit-learn's train_test_split, with a 90:5:5 ratio. The model was trained on 90% of the dataset, and 10% of the dataset used for testing and validation. For training of parameters, AdamW optimizer is chosen. The loss function that we used is the standard nn.CrossEntropyLoss(). The model is trained on 3 epochs, and it took approximately 6 hours to train it.

Additionally, we wanted to compare BERT's performance with other machine-learning approaches, and thus implemented a basic Multinomial NB model as well as a logistic regression model. These models will be trained and tested on the same training dataset. For the Multinomial NB model, scikitlearn's CountVectorizer was used to tokenise the title feature. To perform validation, we split the dataset using scikit-learn's train_test_split, with a 90:5:5 ratio. No additional preprocessing was done to support the comparison between models.

Finally to test the robustness of both models, the testing dataset was used to test the models.

Results

In this section, we discuss the results obtained by the fake news detection models.

BERT

By using the BERT model on the test dataset obtained from train_test_split, the model accuracy is at 98% (see Table 1). The BERT model also performed very well with a high accuracy, f1 score, precision and recall. Out of 260 records, the model predicted 256 correctly, with only 2 false positives and 2 false negatives.

Table 1. Classification report of BERT model (validation dataset)

	precision	recall	f1-score	support
Fake news	0.99	0.99	0.99	144
True News	0.98	0.98	0.98	116
accuracy			0.98	260
macro avg	0.98	0.98	0.98	260
weighted avg	0.98	0.98	0.98	260

The BERT model also performed well on the testing dataset obtained from another source (see Table 2). The model accuracy on this dataset is at 95%. Out of 55 records, the model predicted 52 correctly, with 1 false positive and 2 false negatives.

Table 2. Classification report of BERT model (testing dataset)

	precision	recall	f1-score	support
Fake news	0.96	0.93	0.95	28
True News	0.93	0.96	0.95	27
accuracy			0.95	55
macro avg	0.95	0.95	0.95	55
weighted avg	0.95	0.95	0.95	55

Multinomial NB

The multinomial NB results performed poorer relative to BERT with generally lower metric scores for both training and testing data.

By using the NB model on the test dataset obtained from train_test_split, the model accuracy is also high at 97% (see Table 3). The NB model also performed very well with a high accuracy, f1 score, precision and recall. Out of 260 records, the model predicted 251 correctly, with only 6 false positives and 3 false negatives.

Table 3. Classification report of NB model (validation dataset)

	precision	recall	f1-score	support
Fake news	0.96	0.98	0.97	133
True News	0.98	0.95	0.97	127
accuracy			0.97	260
macro avg	0.97	0.97	0.97	260
weighted avg	0.97	0.97	0.97	260

By using the NB model on the test dataset obtained from train_test_split, the model accuracy is at 76% (see Table 4). Out of 55 records, the model predicted 42 correctly, with only 4 false positives and 9 false negatives.

Table 4. Classification report of NB model (training dataset)

	precision	recall	f1-score	support
Fake news	0.83	0.68	0.76	28
True News	0.72	0.85	0.78	27
accuracy			0.76	55
macro avg	0.76	0.76	0.76	55
weighted avg	0.76	0.76	0.76	55

Logistic Regression

The Logistic Regression (LR) model performed slightly poorer relative to BERT with lower metric scores for training data.

By using the LR model on the test dataset obtained from train_test_split, the model accuracy is also high at 97% (see Table 5). The LR model also performed very well with a high accuracy, f1 score, precision and recall. Out of 260 records, the model predicted 253 correctly, with only 3 false positives and 4 false negatives.

Table 5. Classification report of LR model (validation dataset)

	precision	recall	f1-score	support
Fake news	0.98	0.97	0.97	133
True News	0.97	0.98	0.97	127
accuracy			0.97	260
macro avg	0.97	0.97	0.97	260
weighted avg	0.97	0.97	0.97	260

By using the LR model on the test dataset obtained from train_test_split, the model accuracy is at 95% (see Table 6). Out of 55 records, the model predicted 52 correctly, with only 1 false positives and 2 false negatives.

Table 6. Classification report of LR model (training dataset)

	precision	recall	f1-score	support
Fake news	0.96	0.93	0.95	28
True News	0.93	0.96	0.95	27
accuracy			0.95	55
macro avg	0.95	0.95	0.95	55
weighted avg	0.95	0.95	0.95	55

Discussion

1. Performance difference between models.

In terms of the training data, BERT has outperformed both the LR and NB models by 1%. This demonstrates that on small datasets, state-of-the-art models can perform better than simpler methods.

The drop in the accuracy of the models on unseen data is expected. The robustness of fake news detection models has been discussed in the literature. Past work have used traditional machine learning methods as well as deep learning to detect fake news, but it is found that models trained on a dataset do not often do well in other datasets as there are factors such as dataset bias or that the models were developed and designed for those datasets (see Khan et al., 2019).

However, what was unexpected was that the BERT model's drop in accuracy was far lower as compared to the NB model while the LR model as well as the BERT model on the unseen data.

There are a few reasons for this difference in model performance. Firstly, being able to maintain the order and contexts of the words could be useful for a model's generalisability to unseen and new datasets. Deep neural networks like BERT process actual sequences of words (coded as integers) as they appear in the documents thereby maintaining the order and contexts of words. In contrast, the LR and NB models rely on a Bag-of-words method where the order and contexts of words are lost and semantic similarities between words cannot be represented. Secondly, the good performance of LR could be attributed to a few reasons, the first is that the dataset size is relatively small and that being a binary classification problem, LR is likely to perform well on tasks like this.

2. Time consumption

The BERT model training took around 6 hours for 3 epochs, and the NB model took only 3.79 ms, while the LR model took 103 ms. This is because the BERT model training is based on a Recurrent Neural Network process and transforming process. Both of them have high time complexity. Therefore, the running time for training a BERT model can be expected to be really long. Furthermore, for this report, the BERT model is actually fine-tuning a pre-trained model with 109M parameters on cased English text with the training dataset. As such, there is a great size difference in the training as compared to the NB and LR models which were just trained on the training dataset.

Limitation

There are a two limitations with this study that is of note

1. More data needed

As with all machine learning models, having access to more data allows for the model to be trained better. In our case, we were only able to gather close to 5000 records, and our models could have performed better if there was more properly labelled data available. However, it also comes with a tradeoff, as this would also mean the training times for the model would be much longer.

Also, it is easy to scrape data related to COVID-19, but it is hard to retrieve labeled data. This is due to the fact that COVID-19 is a relatively new thing, and not much labeled data is available. This also means that as time passes, models will only get more accurate as more data, labeled and unlabeled, are available for training. Additionally, a larger dataset could also show a larger difference between machine learning models and deep learning.

2. The changing nature of misinformation

Due to the rapidly evolving nature of the COVID-19 pandemic situation, some advisories such as the mask advisories have been revised in light of new research concerning the contagiousness of the virus. These could have an impact on the fake news detection models. Future work can examine how fake news detection models can be sensitive towards such changes.

Conclusion

Deep learning is a great way to predict which articles are misinformation when it comes to COVID-19. This is also evident in other forms of text analytics: email spam-filtering, positive and negative sentiment analysis etc. However, it also depends on the model that is used. In our case, BERT worked the best as seen from the results generated above, whereas other models such as multinomial NB and LR did not work as well.

References

Baruah, A., Das, K., Barbhuiya, F., & Dey, K. (2020). Automatic Detection of Fake News Spreaders Using BERT. *CLEF*.

Bourgonje, P., Schneider, J. M., & Rehm, G. (2017). From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism*, 84–89.

Cui, L., & Lee, D. (2020). CoAID: COVID-19 Healthcare Misinformation Dataset. *ArXiv*.

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

Hall, M. C., Prayag, G., Fieger, P., & Dyason, D. (2020). Beyond panic buying: Consumption displacement and COVID-19. *Journal of Service Management*.

Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection. *ArXiv*.

Shahi, G. K., & Nandini, D. (2020). FakeCovid–A Multilingual Cross-domain Fact Check News Dataset for COVID-19. *ArXiv*.

Wong, Y. (2020). Global Health Security–COVID-19 and Sinophobia in Singapore. *RSIS Commentaries, 046-20*.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676.