Identifying Disinformation through BERT

Kwok Chun-kiu

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Hong Kong

15 May 2023

Contents

List of Figures

List of Tables

Identifying Disinformation through BERT

SEEM2460 Introduction to Data Science Course Project Report

KWOK Chun-kiu

AIST, 1155141911@link.cuhk.edu.hk

1 Introduction

As deep generative models, such as ChatGPT, have become ripened, users can generate all sorts of media from art, music and even novels. However, these enhanced powers also paved the way for illegal acts such as deep fakes and fake news.

Fake news has substantially affected people's lives. For example, during the COVID pandemic, myths and rumours about vaccines and the virus had stalled efforts to stop the virus [5]. Dr Van Kerkhove, technical lead for the COVID response in the WHO Health Emergencies Programme, commented that disinformation is "undermining the safe and effective tools" that can mitigate the impact of the virus.

Fake news is also leveraged as a weapon to shudder democracy or the government of a country by the opposition or even foreign powers, as in the recent case of Nigeria and Brazil [7, 13]. In both cases, presidential candidates are either defamed as corrupt or vilified as criminals. These disinformation attacks have significantly impeded the institution's credibility and intensified the polarization of different political views, which in turn causes a rift in the nation.

It is obvious that fake news is becoming more and more of a threat to the society we reside and even to the whole world if we do not eradicate them in time. By analyzing the writing styles, we can leverage AI to identify the cliches in fake news.

2 Methods

[Note: Section 2.1 and 2.2 are paraphrased from my final year project report.]

2.1 Bidirectional Encoder Representations from Transformers (BERT)

We use the BERT model to process the sentences in the dataset. The BERT algorithm is introduced in [8] and is designed to learn unlabelled texts from both left and right contexts. With its implementation, eleven tasks related to natural language processing have been improved. Based on the context of masked tokens, the Masked Language Model (MLM) objective randomly masks some tokens from input and requires the model to predict the ground truth label. 15% of the input tokens in the pre-training phase are uniformly masked 80% of the time; 10% are replaced with random words; and for the remainder, the word remains unchanged [10]. Through MLM, the representation can combine left and right contexts, making it possible to train a bidirectional deep transformation. Additionally, the "next sentence prediction" task is performed to train representations of text pairs. The BERT model extracts text vectors from a token sequence, often derived from the output of the tokenizer, and inserts special tokens such as `[CLS]` and `[SEP]`. By using these tokens, the model can handle a single sentence as well as a pair of sentences in a single token sequence. Transformer blocks are then used to generate token-type embeddings, segment-type embeddings, and positional embeddings (see Figure 1), and these three embeddings are summed as the final embeddings. Custom heads (also known as "poolers" in some literature) can be added to the model to suit different purposes, such as classification or regression. In the present project, we use the classification head which contains two linear layers.

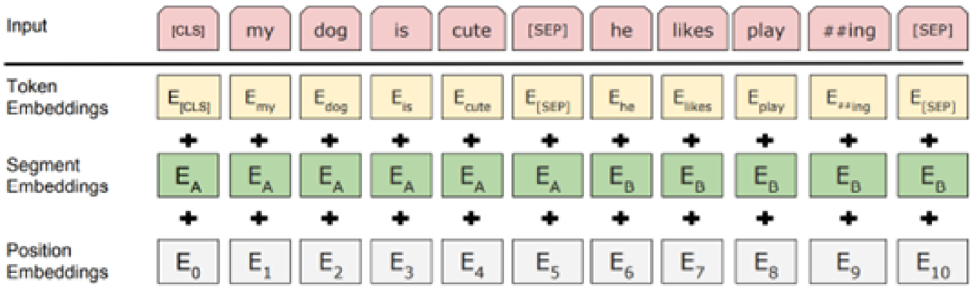


Figure 1: The model architecture of BERT.

2.2 RoBERTa

As a replication study, [11] found that BERT was significantly under-trained, but could match or outperform post-BERT models if the training method was improved. Thus, the Robustly Optimized

BERT Approach (RoBERTa) is introduced. In this recipe, BERT is trained longer with larger batches and datasets, the "next sentence prediction" objective has been scrapped, longer sequences are trained, and dynamic masking patterns are applied to the training data. Using dynamic masking eliminates overfitting issues caused by the original BERT implementation, which generates masking patterns only during data preprocessing. Results show that by introducing dynamic masking, the performance is similar to static masking but with improved efficiency. Also, removing the next sentence prediction task allows the model to match BERT's performance, disproving the belief that the next sentence prediction task is a crucial part of BERT. Training with larger batches also reduces the perplexity of the prediction; in other words, the model can predict texts better. Over eight natural language processing tasks, RoBERTa performed better than BERT by 1-2% points and achieved state-of-the-art on GLUE, RACE and SQUAD datasets, which demonstrates that BERT, after refining the design decisions, remains competitive even after the invention of newer models.

2.3 Python Libraries

For data analysis and visualization, we utilize the Pandas library to investigate the dataset and visualize the results through Seaborn. Pandas is a data analysis library which allows us to import, process and export data tables. [1] To be concrete, it allows data filtering and SQL-like operations such as table join. Seaborn is an advanced data visualization library based on the more basic plotting library matplotlib. [3] It allows users to easily plot heat maps, scatter plots and other advanced plots. It can also conduct simple regressions without using any other machine learning libraries.

We also utilize the PyTorch library to compile deep learning models for classifying the articles into corresponding classes, "true" or "fake". PyTorch is a full-fledged deep learning library, which allows users to build and load custom data sets and compile deep learning networks. It also allows users to simply download data sets and pre-trained models. We will also use the Hugging Face library, which contains an application programming interface for using off-the-shelf transformers easily. This can ease our effort, so we need not build the transformer from scratch.

Finally, we use the scikit-learn library to calculate the metrics for the models. In the project, we will only use the F1 score, but there are many more metrics available such as the AUC-ROC score, another popular metric for quantifying classification loss, and the Davies-Bouldin index, a popular metric for quantifying clustering accuracy. [2]

3 Experiment

3.1 Dataset

We obtain our data source from the deep learning community Kaggle. This website contains a wide variety of data sets from simple toy data sets to professional data sets and allows users to compete with one another to win prizes. For the investigation of this project, we will use the data set "Getting Real about Fake News" and "Fake and Real News Dataset". [6, 14] The two datasets combined contain a total of 57,897 samples of news articles with its title listed. Non-English articles are filtered out for simplicity. The strings are preprocessed by removing punctuation marks and numerals and converting all alphabets into lowercase. After processing, we have 57,255 rows of records for exploration. Table 1 shows an excerpt of the dataset.

Index	Title (excerpt)	Article (excerpt)	Label
13575	`senior saudi prince freed in billion settlemen...`	`riyadh reuters senior saudi arabian prince mit...`	1
41247	`breaking reporter says secret service took deb...`	`health professionals have been speculating for...`	0
31013	`breaking news on clinton informant s gag order...`	`this is huge the doj has authorized the fbi in...`	0
13855	`family of lebanon pm hariri visit french presi...`	`beirut reuters french president emmanuel macro...`	1
42393	`would you like to live for free on a luxury cr...`	`just when you think you ve heard everything th...`	0

Table 1: An excerpt from the preprocessed training dataset. Strings are truncated automatically by Pandas for readability.

As mentioned before, the articles are labelled according to their veracity: if the article contains false information, the label is 0; otherwise, the label is 1. In this combined dataset, the number of fake news articles is 67% more than that of the true articles. Figure 2 shows the number of news articles counted by their labels.

To count the lengths of the text, we first remove punctuation marks, replace hyphens with whitespace, and then split the string by whitespace. The title lengths range from 0 to 69. As shown in Figure 3, the peak is around 10 to 15. In fact, 50% of the titles are 10 to 15 words long, having an interquartile range of 5.

Before truncation, the article lengths range from 0 to 24356. 50% of the text length lies between 195 and 550, which means an interquartile range of 255. Since our model can only process 256 tokens in a sequence, we use 256 as the truncation length. After truncation, most of the text is 257 tokens long. The number exceeds 256 because we have not removed redundant tokens at the moment; they will be automatically

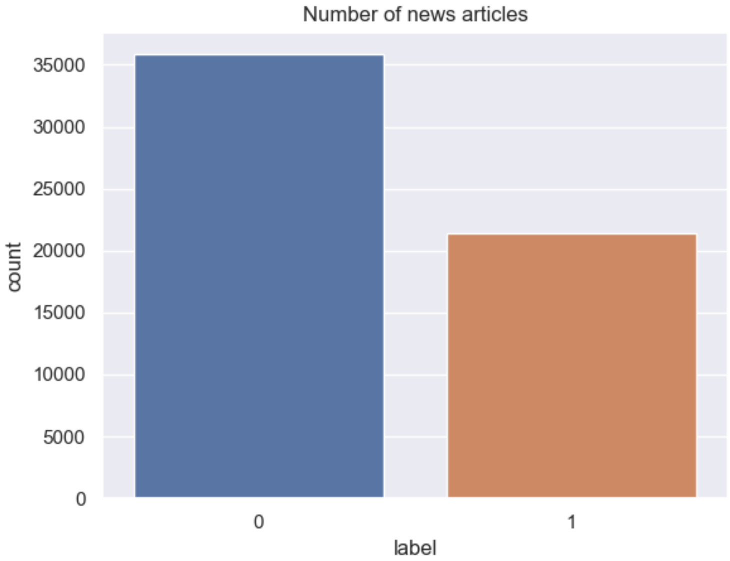


Figure 2: The number of news articles classified by their labels.

removed by the tokenizer during training.

3.2 Classification Task

The project is a binary classification task; that is, we expect the model to predict whether a given article is a piece of news is fake or not. We propose using state-of-the-art architectures such as BERT, a Transformer based model, to complete the classification task. We classify the samples using the traditional classification approach (that is, using SoftMax functions). To investigate the effect of each feature on the model's performance, we also conduct an ablation study. Through this project, we aim to investigate the writing styles or frequent words in disinformation. The combined data set contains a total of three attributes, two of which are the title and text of the article and the veracity of the article as the label. Since the label is the dependent variable, we do not contain it in the input features.

In the project proposal, we proposed using triplet loss to learn a better representation of the articles; however, after reading [9], we find it unsuitable in this case as the number of samples in each class is not small, nor is the number of classes big enough for us to implement triplet loss. If we implement triplet loss in this case, not only will it cause minimal improvement, but also it will add an unnecessary burden to the model. A better usage of triplet loss would be when classifying news articles into different sub-topics or fields of interest, such as 'Politics - Donald Trump' or 'Economics - Interest rate'.

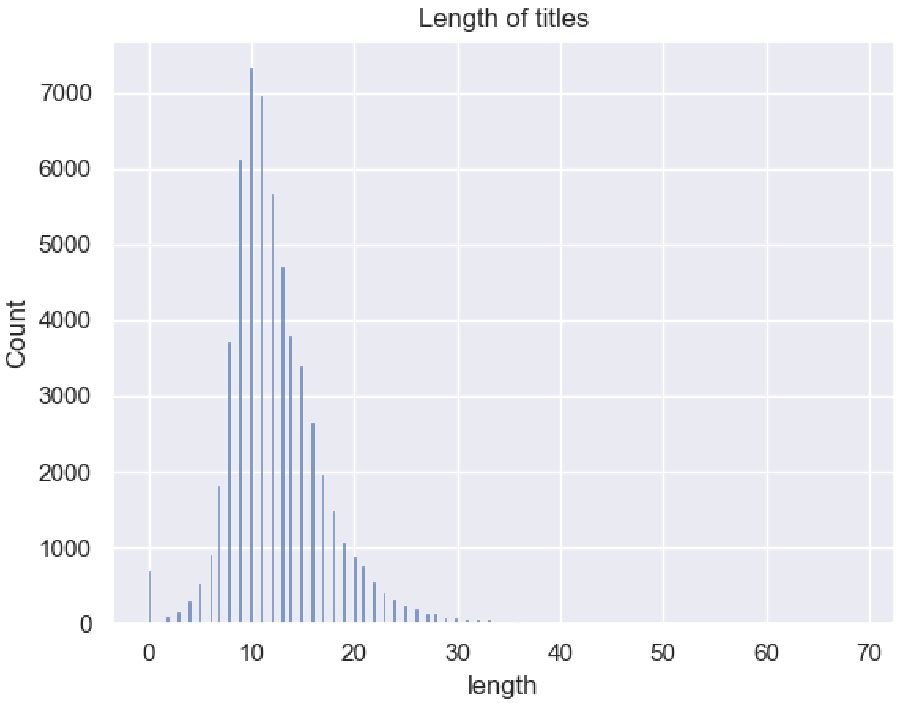


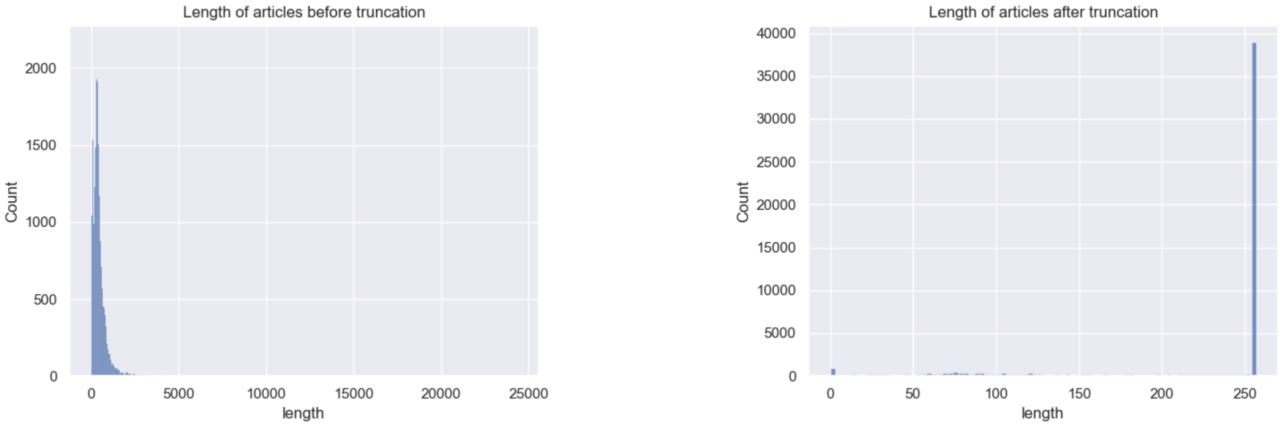
Figure 3: Distribution of title lengths.

3.3 Ablation Study

An ablation study is a method to investigate the effect of input features on the model's performance. [12] For research purposes, this method allows us to investigate the causal relationships between the input features and the output. From a practical perspective, it also allows us to remove redundant features beforehand to increase efficiency and save computing resources. Since there are two textual features, we can either combine them or only use one of them for prediction. Therefore, we can conduct an ablation study to investigate the indicative power of the titles and the articles.

3.4 Configuration of the Model

The BERT model used in this project is directly taken from the Hugging Face library, which provides convenient access to Transformers hosted on their website. To train the model, we use linear warmup and decay to tune the learning rate from 5e-5 down to 0. For the optimizer, we use the AdamW optimizer to do backpropagation. The model is trained for 10 epochs; each time the validation loss is lower than the best, we save the model for a later prediction stage.



(a) before truncation

(b) after truncation

Figure 4: Distribution of news article lengths

3.5 Evaluation

To evaluate our predictions, we use accuracy in line with common practice. We also calculate the F1 score to reveal the actual performance of the model. The F1 score is calculated by the following formula:

$$F1 := \frac{2TP}{2TP+FP+FN},$$

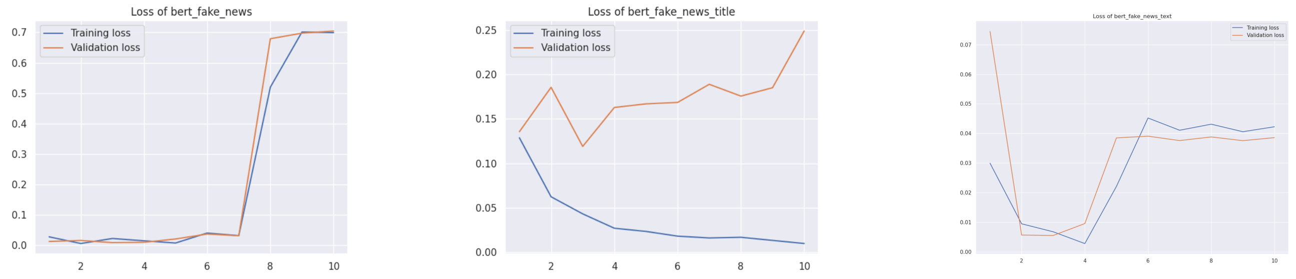
where TP, FP and FN denote the number of true positives, false positives and false negatives respectively. This can be easily evaluated using the scikit-learn library mentioned in Section 2.3. Using the F1 score allows us to have a better grasp of the models' power of classification whenever the accuracy seems suspiciously high, especially in critical usages such as medical diagnosis.

4 Results

Model	# Params (M)	# Samples	Title	Text	Accuracy (%)	F1 Score (%)
BERT	110.0	57,255	▨	▨	99.76	99.67
		56,574	▨		96.08	94.87
		56,516		▨	99.85	99.80
RoBERTa	125.0	57,255	▨	▨	99.31	99.09
		56,574	▨		95.69	94.44
		56,516		▨	99.54	99.39

Table 2: Model results on the classification task. Grey cells indicate the input features used in training the model.

Table 2 shows the results of the BERT model on the classification task. All results are over 90%. The models trained by text ("text models") match the performance of the models trained by both features ("title-text models"). Using titles did not improve the performance of the classification, which implies they may not be descriptive of the articles. As the text provides more context, it allows the model to have a more clear grasp of the topic of the article. Figures 5 and 6 shows the training curves of the two models using different feature configurations. The losses for title-text models and text models show a surge, which occurs when the transformer deviates from a local minimum. However, as we save the model with the least validation loss and call the same model for prediction, this does not affect the result.

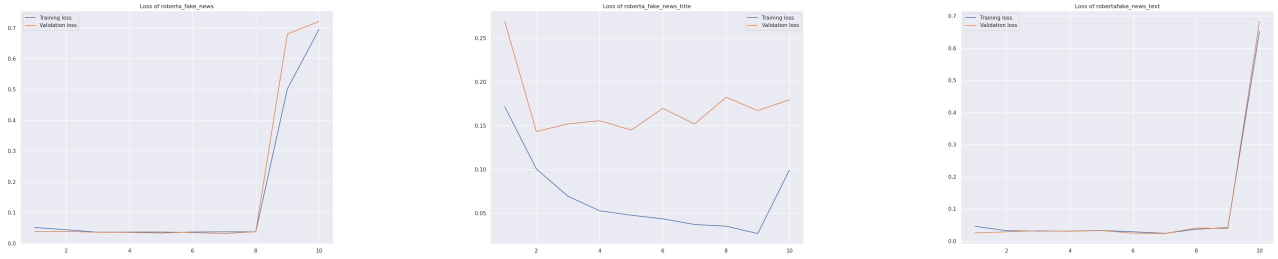


(a) Title and text

(b) Title only

(c) Text only

Figure 5: Training curves of the BERT model.



(a) Title and text

(b) Title only

(c) Text only

Figure 6: Training curves of the RoBERTa model.

5 Conclusion

We successfully classified the news articles with an accuracy as high as 99.85% and an F1 score of 99.80%. This shows that we can automate fake news classification and combat the ever-increasing volume of Internet disinformation using deep learning. A further direction for this project would be to utilize other higher-level technology such as federated learning to facilitate global cooperation on countering disinformation.

While the result is outstanding, this project is not without its limitation. One limitation is that we have only around 57,000 samples, which is far less than the number of websites now on the Internet [4]. This is certainly not representative of the Internet since the Internet now has 1.13 billion websites, which contain a huge amount of data difficult for us to preprocess and store on personal computers, nor to train on Kaggle GPU clusters. Another limitation is that the mentioned models are not online algorithms, which means the model cannot improve itself using new data. As new content is constantly uploaded to the Internet, the use of BERT and RoBERTa may not produce fresh predictions.

References

[1] Pandas. URL: https://pandas.pydata.org/.

[2] scikit-learn. URL: https://scikit-learn.org/stable/.

[3] Seaborn. URL: https://seaborn.pydata.org/.

[4] Top website statistics for 2023. URL: https://www.forbes.com/advisor/business/software/website-statistics/.

[5] D. Bardsley. Disinformation still hampering fight against covid-19, who warns, February 2023. URL: https://www.thenationalnews.com/world/2023/02/09/disinformation-still-hampering-fight-against-covid-19-who-warns/.

[6] C. Bisaillon. Fake and real news dataset. URL: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset.

[7] J. Courtois. The scourge of fake news in brazil's presidential election, September 2022. URL: https://www.france24.com/en/americas/20220918-the-scourge-of-fake-news-in-brazil-s-presidential-elections.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

[9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. arXiv:2004.11362.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL: http://arxiv.org/abs/1907.11692, arXiv:1907.11692.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL: http://arxiv.org/abs/1907.11692, arXiv:1907.11692.

[12] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks, 2019. arXiv:1901.08644.

[13] C. Mwakideu. Nigeria 2023 election: Countering fake news. URL: https://www.dw.com/en/nigeria-2023-election-countering-fake-news/a-64651081.

[14] M. Risdal. Getting real about fake news. URL: https://www.kaggle.com/datasets/mrisdal/fake-news.