Detection for Fake news based on Bert Fine-tuning

Jiajun Guo
Master of Applied Statistics
University of Michigan
Ann Arbor, Michigan 734-450-1356
Email: gjiajun@umich.edu

***Abstract*—Detecting fake news is essential for social media. However, with the increasing amount of information, it becomes tough to detect fake news through human power. To solve this problem, an automatic detecting tools based on deep learning is built in this project. In this project, the author used a pretrained Bert-based model and fine tune it on redasers/difraud dataset from hugging face. The detection model finally achieved the accuracy of 92.5% in the test dataset.**

1. Introduction

Fake news has become a serious problem in social media. It causes misguidance in public opinion and undermine society's harmony. Hence detecting fake news is an essential task in managing social media. However, with the rapid-growing of the amount of information, it becomes more and more time-consuming to make detection in this ocean of information through human power. As a result, there is a pressing need for automatic algorithms to make detection with precision and speed. In these years, deep-learning model become a popular and powerful tool to deal with large amount data and complicate tasks. It is especially widely applicated in NLP, where deep learning methods is used to process language data. It is a potential effective solution to our fake-news problem. This project aims to develop a automatic algorithms based on deep-learning model that classify fake news with precision and speed.

In 2017, transformer [1] structure is proposed. It captures correlation between units through"self-attention", which assign attention weight between units by calculating a weight matrix. It has become the first choice in NLP by its high computing efficiency and strong representation ability. Bert is a transformer-based language model proposed in 2019 by Devlin [2]. One advantage of Bert is that after Bert is pretrained, we could directly use the pretrained weight and fine-tune the model by make slight task-specific modification to the model and train the model using small amount of data without training the model from scratch. By its convinience,Bertt has been frequently used in a wide range of NLP tasks, such as question answering, text classification. Considering the effectiveness and convinience of Bert model, this project will utilized a pre-trained Bert model and fine-tune it on the fake-news dataset.

2. Methodology

2.1. Problem Formulation

This problem is a binary classification problem. The object is to build a model, which takes the input of a sequence and output a label indicate the class. The groud truth label is $[1, 0]$ for truth and $[0, 1]$ for fake. We expect the model's output as close to the ground truth as possible. Meanwhile, before being inputted into the model, the text string need to be processed by tokenizer, where the string is seperated into a sequence of word ids and be able to processed by model.

2.2. dataset

This project utilized the open dataset redasers/difraud on hugging face, which can be accessed on https://huggingface.co/datasets/redasers/difraud. This dataset contains various deceitful and truthful texts from a number of independent domains and tasks. Among them, the fake news domain is used in the project. This subset contains total 20456 texts, with 8832 fake news and 11624 true news. Considering the limitation of time and device, this project will only utilized part of this dataset, with 4000 training data, 400 validating data and 400 testing data. The batch size in this project is set to 1.

2.3. Model

This project utilized distilled bert model. Bert is a language model based on transformer. It takes in a sequence and output an sequence which is an aggregate representation of the full context. It's designed to learn bidirectional representation from unlabeled text by pretraining tasks, where it's given left and right context to predict current text. This enhances language models'ability to learn contextual information in language. Distilled bert [3] is a distilled version of bert base model. It's obtained by pruning the original Bert model, and enhanced by model distilling process. As a result, this model is lighter and faster than Bert, while still having a good benchmark. The model used in

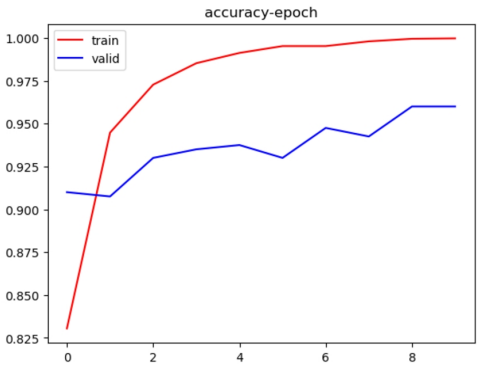

Figure 1. training accuracy and validating accuracy in each epoch

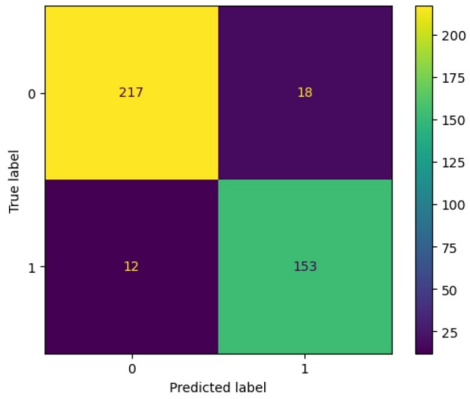

Figure 2. confusion matrix in predicting test dataset

this project is a distilled bert model for sequence classification on https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english, whose structure is modified into classification tasks by adding a full-connection layer on the bottom of base model and change the output into a label vector.

2.4. Optimizer

This project used Adam as optimizer. The Adam is efficient for stochastic optimization which only requires first-order gradient. It takes less memory and compute adaptive learning rate, making it one of the popular methods in training neuro network. The initial learning rate in this project is set to 10^{-4}.

3. result

3.1. training & validating

Considering the less training required in fine-tuning, the epoch number is set to only 10. After each epoch, model is validated on validating set and the accuracy on both training set and validating set in this epoch is calculated and recorded. The accuracy change with epochs as fig.1 shows.

By figure, the model's accuracy is already high in the first epochs, and have no significant increase in the following epochs. This implies that the model's loss converged in only one epoch, which gives evidences for the high efficiency and strong representation ability of the distilled bert model. The model can be fine-tuned only by small amount data and little training time.

In 10 epochs, the model got the highest accuracy of 96% in epoch 8. The model in this epoch is saved as the final version of this model. It will be tested by test dataset in the following section.

3.2. testing

The test is run on a 400 size test dataset. Figure 2 displays the confusion matrix. The matrix shows a good benchmark. This shows that distilled bert is a suitable model for detecting fake-news.

However, this benchmark still have space for improvement. Firstly, not all data is applied in training due to restriction of time and device. If full dataset is applied, author believe that model will attain a higher benchmark. Secondly, in real life, the judgement for fake news not only based on the news' context. There are many other factors such as authors, publishers, contributes to judging fake news. Applying those data may further improve this accuracy, which would requires a new model structure to integrate multi-type data.

4. conclusion

This project applied distilled bert model in automatic detecting fake news. By the result, distilled bert turns out to be suitable for this problem by giving a high accuracy in the test dataset. This would be an applicable methodology in fake news detecting. Meanwhile, more improvement is expected for this project. We hope there will be further research into this problem.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:203626972