A BERT-Based Ensemble Learning Approach for Sentiment Classification in Twitter*

Shupei Li[1] and Ziyi Xu[1]

LIACS, Leiden University, Leiden, Netherlands
{s3430863, s3649024}@umail.leidenuniv.nl

Abstract. To be continued.

Keywords: Sentiment analysis · BERT · Ensemble learning

1 Introduction

2 Related Work

3 Data

4 Methodology

4.1 BERT

BERT [1] is a popular NLP model that has achieved the remarkble performance on various tasks, such as text classification, question answering, etc. Compared to traditional RNN models, it encodes sequences in both directions instead of following a left-to-right or right-to-left routine, which is more closer to how humans understand the meaning of the text. Its bidirectional encoding ability is accomplished by the transformer architecture, which is a stack of encoders using multi-head attention mechanism. Specifically, encoders process the entire sequence at once and use layer-wise tensor operations to learn relationships between words in a sentence. This design not only encodes the inputs bidirectionally but also eliminates the possible local bias, for it gives equal importance to the local context and the long-distance context. It is worth mentioning that the training process of BERT is more efficient than RNN due to the feasibility of parallelization.

BERT requires a special format of input called WordPiece. The WordPiece tokenizer splits words into tokens and adds special tokens at the beginning as well as the end of the sentence. Preprocessed inputs provide three aspects of information for BERT model: tokens, sentence segments, and the absolute position. We perform the tokenization operation on both the training set and the test set before modeling.

* Text Mining, Master CS, Fall 2022, Leiden, the Netherlands

We also consider two variants of BERT in our project – RoBERTa [2] and DistilBERT [3]. RoBERTa is an optimized version of BERT model. Authors of RoBERTa find that the original BERT is actually undertrained after reproducing BERT experiments. Their solution is increasing the training epochs of BERT and carefully select hyperparameters, which significantly improves the model performance in practice. On the contrary, DistilBERT is a compressed version of BERT. The main idea of DistilBERT is reducing the model size via distillation technique. Distillation consists of a teacher model and a student model. The teacher model is trained on a large dataset and is fine-tuned to maximize the accuracy. However, many features learned by the teacher model are redundant for a specific task. Therefore, we can train a much smaller student model to focus on key features and imitate the output of the teacher model. Experimental results show that DistilBERT is cheaper to train while maintaining a comparable performance to BERT.

4.2 Ensemble Learning

Ensemble learning refers to methods that combine multiple models to achieve better performance in machine learning. It encourages base models to learn different aspects of the data to reduce errors and avoid being entrapped in local optima. In the project, sentiment analysis in Twitter is a multi-class classification problem. And we develop a classification voting ensemble model integrated BERT and its variants. Figure 1 illustrates the architecture of our model.

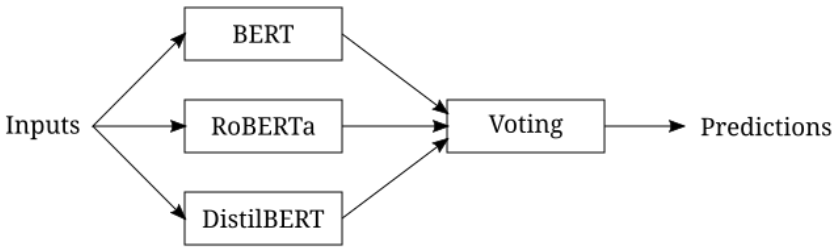

Fig. 1. The Architecture of the Proposed Model

Each base model outputs the class prediction for each record. Then, the voting classifier adopts the plurality voting strategy to generate the final prediction. In other words, the final prediction is the class label received the majority votes from base models. If all classes have the same votes, the voting classifier will choose a class label randomly as the output.

5 Experiments

5.1 Experimental Setup

Software, hardware, baselines, metrics.

5.2 Results

6 Discussion

7 Conclusion

8 Contributions

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv **abs/1810.04805** (2019)
2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)
3. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv **abs/1910.01108** (2019)