News Short Text Classification Based on Bert Model and Fusion Model

Hongyang Cui [1,†], Chentao Wang [2,†], Yibo Yu [3,*,†]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

[2] Tianyi IOT Technology Co., Ltd. Shanghai Branch, Shanghai, China

[3] School of Artificial Intelligence, Nankai University, Tianjin, China

* Corresponding Author Email: 2013839@mail.nankai.edu.cn

[†]These authors contributed equally.

Abstract. Text classification task is one of the most fundamental tasks in NLP, and the classification of short news text could be the basis for many other tasks. In this paper, we applied a fusion model combining Bert and TextRNN with some modified details to expect higher accuracy of text classification. We used the THUCNews as dataset which consists of two columns one for news text and the other for numbers. The original dataset was seperated into three parts: training set, validation set and test set. Besides, we used BERT model which contains two pre-training tasks and TextRNN model which refers to the use of RNN to solve text classification problems. We trained these two models in parallel, and then the optimal Bert and TextRNN models obtained through training and parameter tuning are added with a fully-connected layer to receive the final results by weighting the efficiency of Bert and TextRNN. The fusion model solves the problem of over-fitting and under-fitting of a single model, and helps to obtain a model with better generalization performance. The experimental results show the sharp change in loss and accuracy as well as the final accuracy of the BERT model. The precision, recall-rate and F1-score are also evaluated in this paper. The accuracy of fusion model of BERT and TextRNN is much better than single Bert model and has a gap to 1.76%.

Keywords: Text Classification, Bert, Fusion model

1. Introduction

The short news text classification task is to classify the texts into their corresponding categories by analyzing the contents of the texts. Its input is a short news text of one or two sentences, and its output is the category the text may correspond to, e.g., education, economy, entertainment. The classification of news text through machine learning could automatically classify a great amount of news on the Internet, which improves the efficiency of searching news and the users' experience. Besides, news text classification is also the basis of further analysis of news content and users' preference, which can help operators provide more appealing news.

In the early stage of the research text classification task, Xu et al. employed traditional machine learning algorithms such as naive Bayes to classify news texts, which enables text to be automatically classified by the computer [1]. The results of the experiment illustrated that the effect of text classification is much satisfactory, reaching 96.27%, and the time is relatively short. However, the news categories of the task are really few, and the fields and corresponding dataset involved are also very limited. These shortages truly limit further promotion of the accuracy and the function. With the development of deep learning and neural networks, researchers begin to try to apply them into text classification tasks. For example, some researchers used Text Convolutional Neural Network (TextCNN) to classify political texts and select an appropriate category name for each text through deep learning of text content [2]. And the number of text categories have been significantly increased. However, though the length of text is long, and enough information is learned, the experimental results are not very satisfying, and the f1-micro value is only 78.25%. After the Attention Mechanism and Transformer Network were introduced, researchers proposed the Bert model and tried to apply it to the text classification task. This model can deeply learn the context of each word and have a more

comprehensive analysis of the whole text. The Bert model can more easily deal with data sets with larger data scale and smaller relationship between categories. In an experiment of the classification of Chinese news texts using the BERT model, the F1-score of the whole experiment reaches 0.93, which shows that the accuracy of the experiment is very high [3].

After that, some studies try to further improve the results of text classification through the fusion models. Some researchers combined Bi-directional Long Short-Term Memory (BiLSTM) and Recurrent Neural Network (RNN), which combines the advantages of BiLSTM and RNN and makes up for the problem that RNN tends to ignore the information of context [4]. The model made several tests on a public dataset, and the final results have vividly demonstrate that the accuracy of the fusion model gained a better effect than that of using the model of BiLSTM or RNN did alone. In addition, the fusion of BERT and TextRNN was used for the classification of short texts of Clinical trial screening [5]. Although the training time of the model is relatively short, and the accuracy of the fusion model is a bit higher than that of each single model, the improvement of the experiment is not very ideal. At the same time, the experiment has not proved that the fusion model of BERT and TextRNN is suitable for data sets with more categories or larger scales. Through investigation, it is found that there are not many examples of TextRNN model participating in the work of fusion models, and the fusion model is rarely used in the short text classification task of Chinese news.

On this basis, we tried to combine BERT and TextRNN, and applied the BERT model, TextRNN model and fusion model to the classification of Chinese short news texts. We compared the accuracy of these three models, tried to enhance the accuracy rate of the fusion model in several ways and analyzed the reasons for the change in accuracy.

2. Method

2.1. Dataset description and preprocessing

In this project, we used the THUCNews provided by Tsinghua University Natural Language Processing Lab [6]. It was generated according to the six years' data in Sina News between 2005 and 2011, comprising 200,000 articles in short journalism from 10 different categories of finance, education, science, society, sports, reality, stocks, politics, games and entertainment. The original dataset was divided into three parts: 180,000 news for training, 10,000 news for validating while 10,000 news for testing. The dataset consists of two columns one for news text and the other for numbers. They were separated by a tab and the number represents a category mentioned above.

The preprocessing is composed of three parts. First, we used split method to separate the content string and label. Second, Bert Tokenize was used for word segmentation of the dataset, and the strategy of short filling and long cutting was adopted according to Pad size, with Bert labels such as [PAD] and [CLS] added. Finally, each sentence was converted to the same length as the input layer and an iterator is then constructed.

2.2. Proposed approach

In this project, we applied a fusion model based on Bert and TextRNN as well as modified some details since some studies have proved the advantage of fusion model [7-9]. Bert model uses deep-bidirectional Transformer component to establish the whole model, so the deep-bidirectional language representation that can fuse the context on both sides is finally generated.

Moreover, Bert's pre-training tasks mainly include two parts: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM means a random selection of 15% of the words in one sentence using for prediction. For those words that were covered in the incipient sentence, a symbol which is the same as [MASK] was used to replace them 80% of the time and a random-selected word was used to replace them 10% of the time, and the original word was left fixed in the remaining, about 10%. This forces the model to be more dependent on context information to make predictions for vocabulary and allows the model an ability of correcting error. NSP is a task offered two different sentences in an article and determines whether in the article the second sentence immediately follows

the first sentence. Combined with MLM tasks, the model can more precisely depict semantic information at the sentence level, or probably the discourse level.

TextRNN refers to the use of RNN to solve text classification problems. See Figure 1 and Figure 2, it has two structures: (1) x–>embedding–>BiLSTM–>concat—>softmax–>y (2) x–>embedding–>BiLSTM–>concat–>LSTM–>softmax–>y

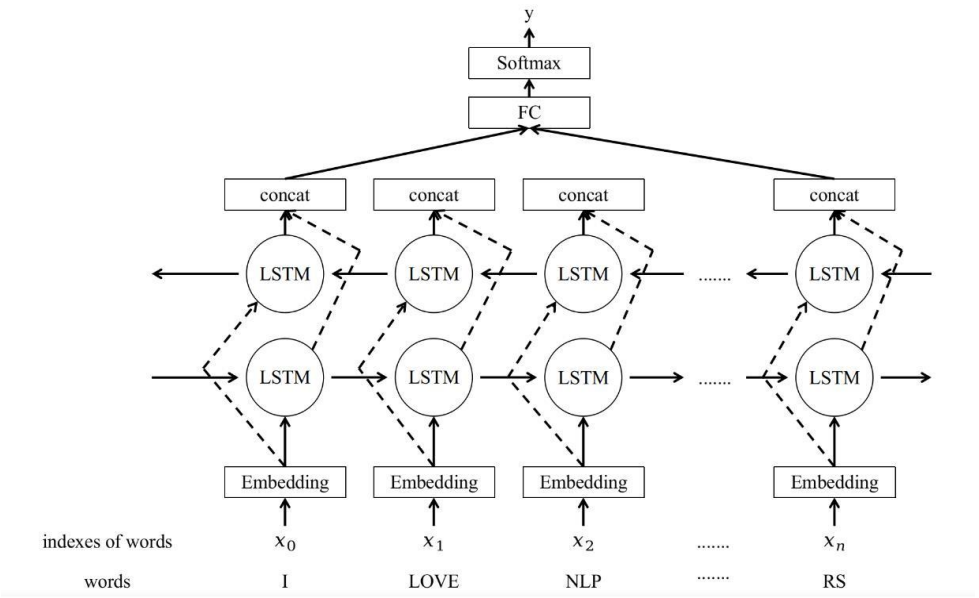

Figure 1. TextRNN structure-1 [10]

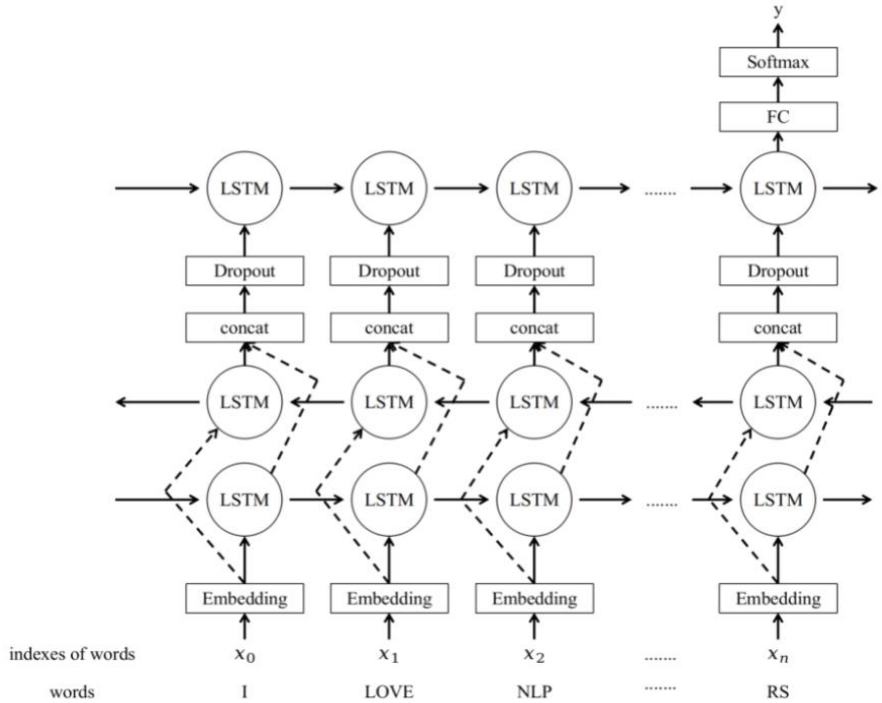

Figure 2. TextRNN structure-2 [10]

However, the fusion model solves the problem of over-fitting and under-fitting of a single model, and helps to obtain a model with better generalization performance. Here, the Bagging integration strategy is adopted in this paper. First, the two models are trained in parallel, and then the optimal Bert and TextRNN models obtained through training and parameter tuning are added with a fully-connected layer to receive the final result by weighting the efficiency of Bert and TextRNN, just like the Figure 3.

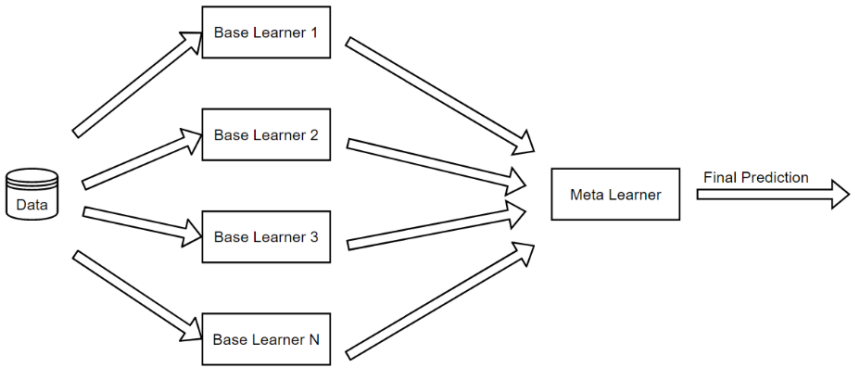

Figure 3. Fusion model

3. Implementation details

In this paper, the accuracy rate, loss value, confusion matrix and other indicators are used to evaluate the employed models. And the model applied a configuration of 12-layer, 768-hidden, 12-heads, 110M-parameters. Meanwhile, the experimental environment adopted in this paper is shown in Table 1.

Table 1. Experimental environment in this study.

Operating System	Linux Ubuntu 20.04 LTS
GPU	RTX 3060
Nvidia	CUDA 11.3
Programming Language	Python 3.9.12
Python Package	Pytorch 1.12.0 Matplotlib NumPy Sci-kit learn

In the process stage, seeing the Figure 4, we first loaded the model, and parameters such as the number of iteration rounds, batch processing size, length of each sentence, and learning rate were set. Loading the dataset starts the Bert model and the training procedure is activated. In this stage, we adopted the strategy of testing on the validation set once each 100 sessions and the test result was output. If the loss measured by the validation set does not decrease after more than 1,000 training sessions, it is considered that there has been no improvement for a long time and the training has reached the optimal level. The training will be stopped and then tested on the test set.

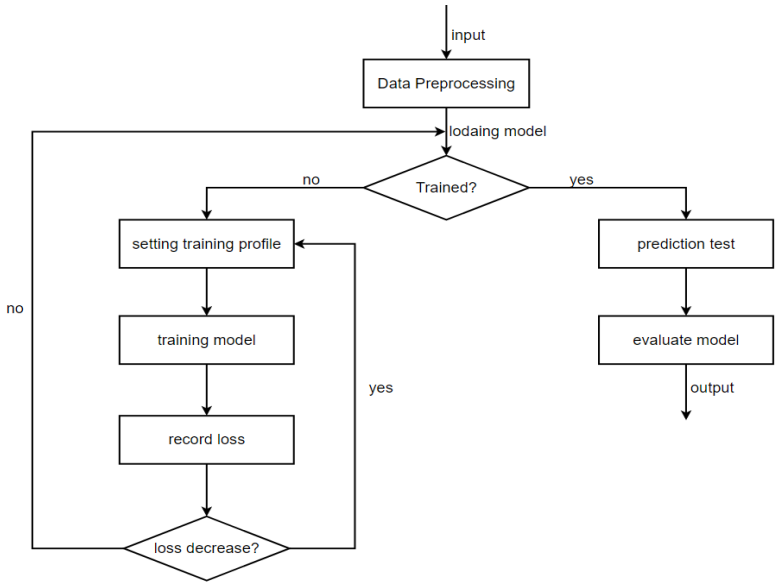

Figure 4. Processing stage

In the test stage, this paper randomly selects some news, outputs Chinese classification, and writes all the Chinese classification results of the test set into another file. Matplotlib was used to visualize the changes of loss and accuracy during training. Finally, the accuracy and loss rate of the test set are measured, and the confusion matrix is output.

4. Results

4.1. Results of Bert Model

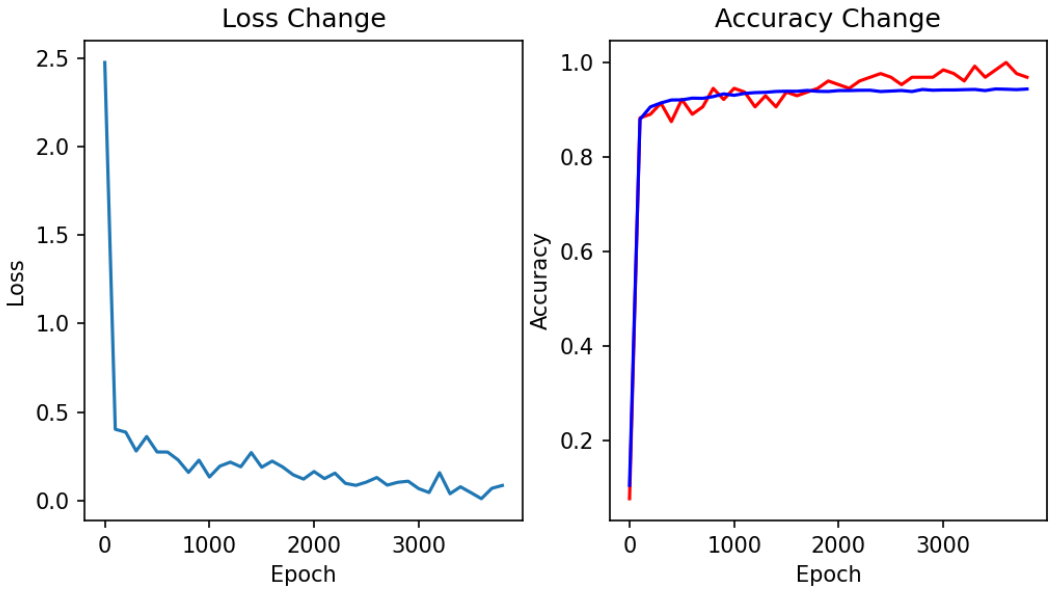

Figure 5. The loss and accuracy of the model

This section records the training process of Bert model, displaying training results, external output, and test set effect. From Figure 5, it can be clearly seen the sharp change in loss and accuracy as well as the final accuracy of the training model reaching 94.01%, which is a relatively high data for the current news short text classification technology.

Moreover, in the Table 2, the precision, recall rate and F1 value are also evaluated in this paper. It can be seen that the classification result of stock news is not so good, only about 90%, and f1-scoer is only about 0.8978.

Table 2. Evaluated metrics for the results.

	Precision	Recall	F1-score	support
Finance	0.9364	0.9280	0.9322	1000
Realty	0.9540	0.9530	0.9535	1000
Stocks	0.9037	0.8920	0.8978	1000
Education	0.9630	0.9630	0.9630	1000
Science	0.8994	0.9030	0.9012	1000
Society	0.9377	0.9490	0.9433	1000
Politics	0.9181	0.9310	0.9245	1000
Sports	0.9820	0.9810	0.9815	1000
Game	0.9702	0.9440	0.9569	1000
Entertainment	0.9461	0.9660	0.9560	1000

4.2. Result of Fusion Model

In this section, we ran the fusion model consisting of Bert and TextRNN, and recorded the result under the same condition. Seeing Table 3, it can be observed that fusion model is much better than Bert model and has a gap to 1.76%.

Table 3. Comparison between Bert and Fusion model.

	Fusion model	Bert model
Condition	Learning rate: 7e-5 Pad_size: 32 Batch_size:128	
Accuracy	95.73%	94.01%

In addition, parameters of Fusion Model are adjusted expecting for better accuracy. It can be obviously concluded that the model performs well in all aspects when the Pad Size is 32 and 40.

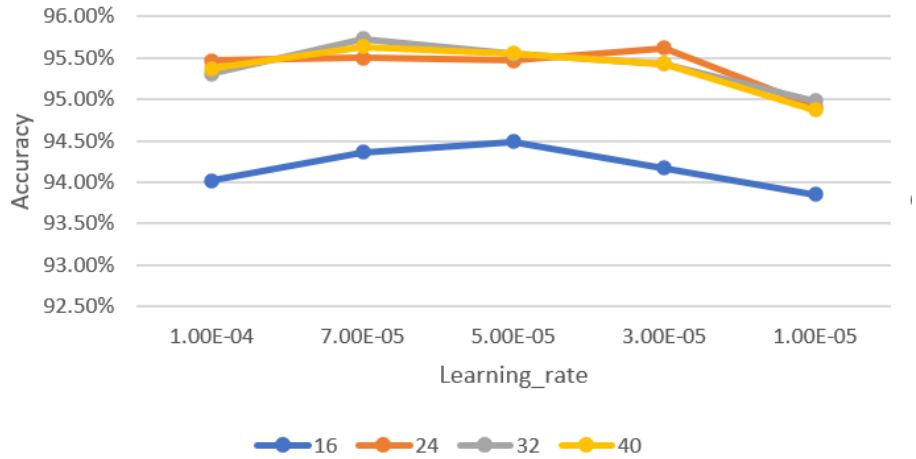

Figure 6. Precision change situation

It can be found from the Figure 6 that the overall accuracy increases with the improvement of the learning rate. However, under the condition of fixed number of iterations, when the learning rate is too large, it can be known that the under-fitting phenomenon occurs according to the change of the accuracy, and it repeatedly jumps around the optimal solution and fails to converge to the optimal solution. The learning rate is too small, the learning time is too long, and the phenomenon of over-fitting occurs.

The model has high accuracy on the training set, but the performance on the test set decreases. In terms of pad size, as it represents the the input sentence's length, when the pad size is too small, the model is forced to delete part of the content in the preprocessing stage to adapt to the size of the input layer, and some information is lost, resulting in the sample with a lower accuracy than the Pad size under the same learning rate at each learning rate. When the pad size is too large, the training time and video memory demand will be greatly increased. Due to the hardware resource limitation, only the maximum value of 40 is tested. Theoretically, when the pad size is slightly larger than the average text length, better results can be achieved, which can not only effectively express sentence information, but also not lead to waste. The experimental results are also consistent with this inference, that is, the accuracy increases with the increase of Pad size, but the improvement is not significant after 32, while the training time increases exponentially.

5. Conclusions

In this work, we proposed to combine traditional Bert model with TextRNN to test the effect on Chinese news text classification using a large number of short text and we slightly tuning parameters to find a better result. We evaluated both Bert model and fusion model in different aspects. Massive experiments were carried out and the results demonstrated that the fusion model had a better performance than simple model did in this project. In the future, we may combine more models and make comparisons in order to obtain a better model.

References

[1] Xu Baoxin, Huai Libo, Cui Rongyi. Naive Bayes algorithm application in the classification of news based on MapReduce [J]. Journal of Yanbian University (Natural Science Edition), 2017,43(01): 55-59.DOI:10.16379

[2] Li Yue, Tang Kun. Policy text classification based on TextRNN [J]. Electronic Design Engineering, 2022,30(12): 43-47.DOI:10.14022

[3] Duan Dandan, Tang Jiashan. Wen Yong, Yuan Kehai. Chinese Short Text Classification Algorithm Based on BERT Model [J]. Computer Engineering, 2022,30(12): 43-47.DOI:10.14022

[4] Gong Weiyin, Wei Xuqin. News Text Classification Method Based on BiLSTM-RNN Model [J], Computer Knowledge and Technology, 2021,17(21): 105-107.DOI:10.14004

[5] Yang Fei-hong, Wang Xu-wen, Li Jiao. BERT-TextRNN-based classification of short texts from clinical trials [J]. Chinese Journal of Medical Library and Information Science, 2021,30(01):54-59.

[6] Natural Language Processing and Computational Social Science Lab. Thuctc [R], 2022 http://thuctc.thunlp.org/

[7] Mohammed, Adam AQ, et al. Multi-model ensemble gesture recognition network for high-accuracy dynamic hand gesture recognition [J]. Journal of Ambient Intelligence and Humanized Computing (2022): 1-14.

[8] Yu, Jun, et al. Multi-model Ensemble Learning Method for Human Expression Recognition [R]. arXiv preprint arXiv:2203.14466 (2022).

[9] Khan, Aisha Urooj, et al. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering [R]. arXiv preprint arXiv:2010.14095 (2020).

[10] CSDN. Text-RNN [R]. 2020, https://blog.csdn.net/beilizhang/article/details/109005461