Sentiment Classification Based On BERT

Jinbin Cai[1], **Fei Chen**[1], **Siyao Chen**[1]
[1]Xiamen University, Xiamen, China

Abstract

In recent years, social networks such as Weibo, social networks, forums, wikis, and online shopping platforms have gathered a large number of users. These users are not only the browsing and recipients of online information resources, but also the providers and transmitters of said resources. This information includes objective reports on people, things, and events, as well as subjective expressions of people, things, and events. How to automatically analyze and process subjective emotional expressions from different social networks has become a difficult problem to be solved urgently. Among them, the theoretical basis of sentence-level text sentiment analysis technology involves many research fields such as natural language information processing, artificial intelligence, information retrieval, probability and statistics, and is used in practical applications for customer experience, market research, customer insight, and digital analysis. Even media evaluation is a key solution, closely related to Internet social media analysis such as blogs, Weibo and WeChat, and has become a current research hotspot. This article conducts indepth research on sentence-level sentiment classification. Under the framework of supervised learning, the Bert model is introduced, the improvement research of sentence-level sentiment analysis is carried out, and some novel sentiment classification methods are proposed. And conducted experiments on the hotel review data set. Experiments show that the model has achieved good research results.

1 Introduction

With the rapid development of computer and network technology, human has entered the information society. Along with it, profound changes have taken place in people's work and life. One of the most important changes is that human activities are increasingly networked. More and more people work, shop, entertain and make friends through the network. In the network society, people's social way and scope have also undergone great changes. The changes of the way are mainly reflected in using blog to publish log, using social network to chat with friends, using micro blog to exchange news, etc. the change of scope mainly reflects that the social scope begins to break through the geographical limit and enter into a broader virtual space. At this time, a large number of users, such as wiki, social networking and so on, gathered. The depth and breadth of users' participation in network activities has reached an unprecedented scale.

Emotion analysis technology is an important basic research topic in social media computing. Its purpose is to classify the sentiment tendency of social media data (Internet information resources) "Support, opposition or neutrality" and emotion classification "joy, anger, sadness, fear, panic" and so on, because most of the network information resources are based on sentence level text, it can also be called sentence level sentiment analysis technology. In the face of the explosive increase of Internet information and the urgent needs of users, how to quickly mine, analyze, express and manage the emotional category information in Internet information resources, so that users can quickly find the information they need, has become the key technology to realize social media computing.

The full name of Bert (Devlin et al. 2018) is bidirectional encoder representation from transformers. In traditional language models, CNN or RNN are usually used. In order to meet the needs of downstream tasks, researchers have developed a large number of network structures. The result is that once the downstream tasks change, or even just change the data sets with different distributions, the final result will be obvious performance loss. Considering the powerful generalization ability of human language, this processing method is obviously inconsistent with human normal language understanding behavior. The appearance of Bert has completely changed this situation. Instead of using traditional CNN or RNN, Bert is based entirely on transformer. In squad1.1, the top level test of machine reading comprehension, Bert has achieved amazing results: it has surpassed human beings in all two indicators, and achieved SOTA performance in 11 different NLP tests, including pushing the glue benchmark to 80.4% (absolute improvement of 7.6%), and accuracy of 86.7% (absolute improvement of 5.6%), which has become a milestone model achievement in the history of NLP. In this paper, Bert is applied to sentence level sentiment classification, and the accuracy of sentence level sentiment classification in hotel review data set is more than 95%.

2 Related Work

Thanks to the development of computer hardware technology, deep learning related fields have also been greatly de-

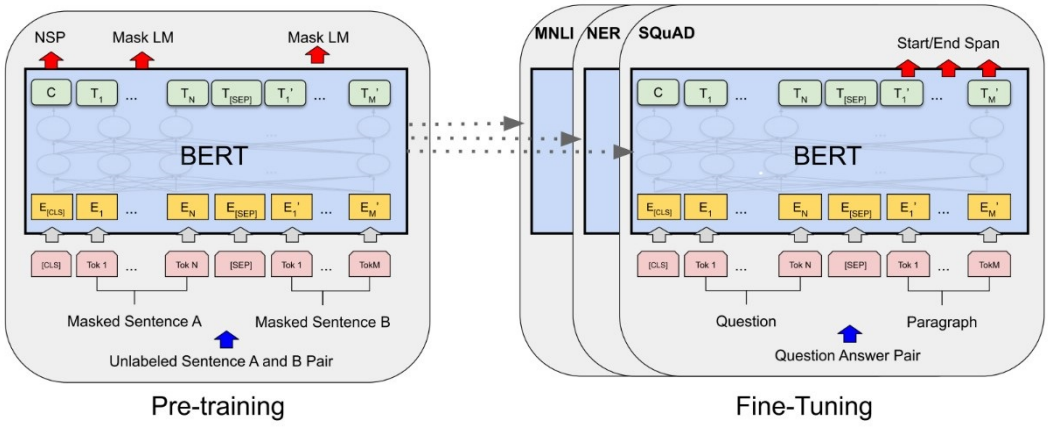

Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different downstream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special special separator token (e.g. separating questions/answers).

.

veloped. R. Collobert et al. (Collobert et al. 2011) established a unified model by using neural network. The model does not use man made input features and prior knowledge, but extracts more essential features from a large number of unlabeled data for various tasks of natural language processing. R. Moraes et al. (Moraes, Valiati, and Neto 2013) proved that artificial neural network (ANN) is better than support vector machine in document level sentiment analysis. Johnson et al. (Johnson and Zhang 2014) proposed a CNN based model BOW-CNN, and directly used high-order word vector as input, and achieved good results. Tang et al. (Tang, Qin, and Liu 2015) first learned sentence features from word embedding using CNN or LSTM, and then encoded the semantic relations of sentences by GRU. The effectiveness of this method was verified on IMBD and Yelp datasets. Xu et al. (Xu et al. 2016) proposed a LSTM model with cache to obtain the overall semantic information in long text. Yang et al. (Yang et al. 2016) proposed a multi-layer attention model (HAN) which can be used in document level emotion classification tasks. The model includes two levels of attention mechanism: word level and sentence level, which can better express document features. For sentence level sentiment classification task, Socher proposed RAE (Socher et al. 2011) to obtain the reduced dimension vector representation of sentences, to represent the MV-RNN (Socher et al. 2012) associated with each word and the matrix in the tree structure, and to use tensor based composite functions to better obtain the RNTN (Socher et al. 2013) of interaction between elements. Dos Santos et al. (Dos Santos and Gatti 2014) proposed a CharsCNN which uses two convolution layers to extract relevant features from words and sentences. Wang et al. (Wang et al. 2015) used the hidden state parameter of LSTM to obtain the relationship between words, and classified the sentiment of twitter text. Wang et al. (Wang, Jiang, and Luo 2016) proposed a hybrid model of CNN and RNN by using CNN to learn coarse-grained local features and RNN to learn long-distance dependence. Tang et al. (Tang, Qin, and Liu 2016) adopted multi-layer LSTM, added attention module of extra memory on each layer, and constructed end-to-end memory network, which had significant effect in object-level emotion analysis task.

3 Proposed Method

In the sentence level sentiment classification, the model based on Bert mainly processes the comment sentence into the sequence input form of Bert, then takes the representation vector of [CLS] as the expression vector of the whole sentence to represent the emotional information of the whole sentence, and then obtains the classification result through pooling operation. The main architecture of the model is discussed below.

3.1 BERT

The model structure of BERT is a multilayer bidirectional Transformer encoder based on the original implementation described by Vaswani et al. (Vaswani et al. 2017). In this section, we will introduce BERT and its application in sentence-level sentiment classification.

Our framework has two steps: pre-training and downstream task fine-tuning. In the pre-training process, the model is trained on unlabeled data in different pre-training tasks. In order to fine-tune, first initialize the BERT model with pre-trained parameters, and then use labeled data from downstream tasks to fine-tune all parameters. Each downstream task has a separate fine-tuning model, even if they are initialized with the same pre-training parameters. The question and answer example in Figure 1 will be used as a running example of the BERT model in this section.

The Transformer in BERT uses two-way self-attention, while the Transformer in GPT uses limited self-attention. Each marker can only focus on the context on its left. We noticed that in the literature, the two-way Transformer is usu-

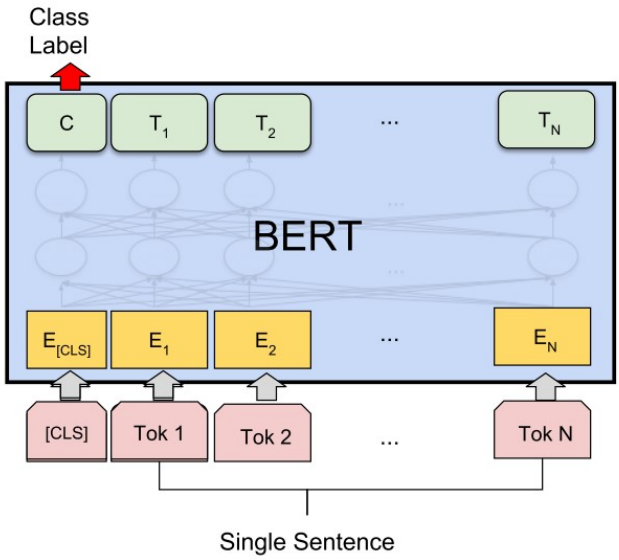

Figure 2: BERT in Sentiment Analysis

ally called the "Transformer encoder", and only the version marked with the left context is redefined as the "Transformer decoder" because it can be used for text generation.

In this article, BERT is used for sentence-level sentiment classification. Our input representation can clearly represent a single text sentence in a tag sequence. The input representation is constructed by summing the tag embedding, sentence embedding, and position embedding corresponding to a given tag. The model structure is shown in Figure 2.

3.2 Pre-training

The pre-training process generally follows the language model pre-training process in the previous literature. For the pre-training corpus, we use Chinese Wikipedia (2,500M words). For Wikipedia, we only extract text paragraphs, and ignore lists, tables, and headings. Then split the BERT Wikipedia training corpus into train_wiki.txt and test_wiki.txt. The corpus comes from: https://github.com/brightmart/nlp_chinese_corpus.In the prepared file, it is written in the following format. Each line is a string, which corresponds to two sentences with context.

BERT paper, the recommended model parameters: a reference model (transformer_block = 12, embedding_dimension = 768, num_heads = 12, TotalParameters = 110M), visible parameters in which a total of 110 million, in addition, there is also greater than the reference model. A high-performance model with a parameter amount of 300 million. To train and use a model with such a large parameter, ample computing resources are required! But after my actual test, combined with the needs of sentence-level sentiment classification I am currently studying, I found that this is actually parameter comparison excess, we pre-trained BERT used parameters (transformer_block = 6, embedding_dimension = 384, num_heads = 12, TotalParameters = 23M), the parameters reduced to 20 million, but even

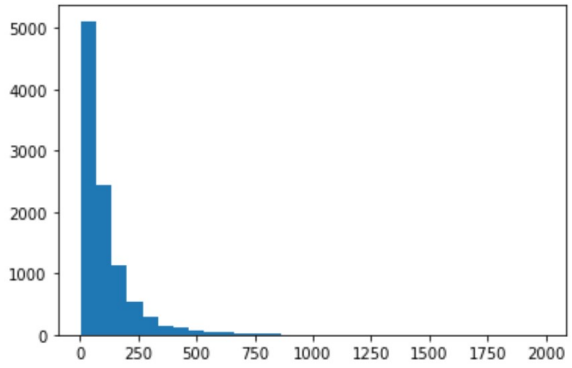

Figure 3: Sentence length

so, using a 11GB of video memory 2080Ti, it also takes a week to train the BERT on the Wikipedia corpus. At the same time, the model we use today is modified on the basis of the open source project https://github.com/huggingface/pytorch-transformers.

4 Experiment

In this section, we conduct an empirical evaluation of Bert on the hotel reviews dataset.

4.1 Dataset

The training data set is from hotel reviews. The sentence length is shown in Figure 3. Most of the sentence text lengths are below 300. All the training data are 10000 in total, and the data set is divided according to the proportion of training set: verification set = 8:2 by leaving method.

```
有几次回到酒店房间都没有被整理。两个人入住，只放了一套洗漱用品。
负样本，输出值0.22
----------
早餐时间询问要咖啡或茶，本来是好事，但每张桌子上没有放“怡口糖”（代糖），又显得没那么周到。房间里卫生间用品补充，有时有点漫不经心个人觉得酒店房间禁烟比较好
负样本，输出值0.34
----------
十六浦酒店有提供港澳码头的SHUTTLE BUS，但迭仔没有订了普通房，可能是会员的关系 UPGRADE到了DELUXE房,风景是绿色的河，感观一般，但房间还是不错的，只是装修有点旧了另外品尝了酒店的自助晚餐，种类不算多，味道OK，酒类也免费任饮，这个不错最后就是在酒店的娱乐场赢了所有费用，一切都值得了！
正样本，输出值0.97
----------
地理位置优越，出门就是步行街，也应该是耶路撒冷的中心地带，去老城走约20分钟。房间很实用，虽然不含早餐，但是楼下周边有很多小超市和餐厅、面包店，所以一切都不是问题。
正样本，输出值0.98
----------
实在失望！如果果晚唔系送朋友去码头翻香港一定会落酒店大堂投诉佢！太离谱了！我地吃个晚饭消费千几蚊 ，买单个黑色衫叫Annie果个唔知系部长定系经理录左我万几蚊！简直系离哂大谱的 ！咁样的管理层咁大间酒店真的都不敢恭维！
负样本，输出值0.24
----------
酒店服务太棒了，服务态度非常好，房间很干净
正样本，输出值0.94
```

Figure 4: Test result

4.2 Experimental Result

Dynamic learning rate and early stop: before, we divided the corpus into training and test sets. Our training method is to train each epoch with a training set. The performance of the model is measured by AUC. After the current epoch training is completed, use the test set to measure the current training result and record the current epoch's AUC. If the current AUC is not improved compared with the previous epoch, then the learning rate will be reduced. The actual operation is to reduce the current learning rate by 1/5, until the AUC of 10 epoch test sets has not improved, the training will be terminated. Our initial learning rate is 1e-6, Because we train on the basis of Wikipedia pre-training corpus, which belongs to the downstream task, we only need to fine tune the pre-training model.

In practice, using the method of sigmoid(x), we found that although the training set and the test set are both very high, but after inputting some random reviews from various online hotels, we find that the generalization ability is not good. This is because our training data set is very small, even if we distinguish the training set and the test set, but the overall data form is relatively single. In order to improve the generalization performance of the model, I tried another model structure.

I try to use the mean-max pooling method to convert the sequence of hidden layers into a vector. In fact, it is to find the mean value and maximum value respectively along the sequence length dimension, and then put them together to form a vector. After that, they are also mapped to a value and then activated. At the same time, we also use the method of weight decay, which is actually L2 normalization. In pytorch, there is an interface that can be called directly. We will talk about it later. In fact, the function of L2 regularization is to prevent the value of parameters from becoming too large or too small. We can imagine that since our training data is very few, we can use the method of L2 normalization, Therefore, when we use the model to infer, some combination models of words and words or sentence structure are never seen. If the value of parameters in the model is large, the response of the model to the sentence will be too large when encountering some special sentences or words, which will cause the final output value to deviate from the reality. In fact, we hope that the model will be more calm, so we add

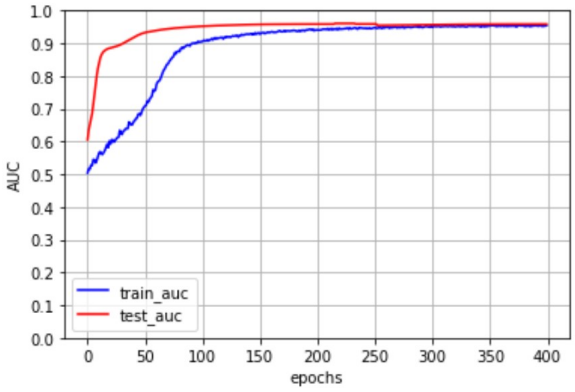

Figure 5: Experimental Result

L2 normalization.

In addition, our pre-trained Bert has six transformer blocks, and we only use three in emotional analysis. Because there are too many parameters behind, it is easy to cause over fitting. Therefore, after the third transformer block, we cut out the hidden layer for pooling, and the latter transformer blocks are not used. Finally, dropout mechanism is used. Dropout is set to 0.4. Because the model parameters are too many, 40% of the parameters are disabled during training to prevent over fitting Through the above methods, the AUC of both the model training set and the tester has reached more than 0.95. Moreover, through the actual test, the model can basically distinguish the emotional polarity of the sentence. The specific results are shown in Fig. 4 and Fig. 5.

5 Conclusions

We propose a model based on Bert for sentence level emotion classification. We first encode context words with pre-trained Bert to capture their information, then fine tune the model on the hotel review data set, and finally classify sentences by using the representation vector of mask [CLS]. A large number of experiments show that our model performs well on the hotel review dataset.

References

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12(ARTICLE): 2493–2537.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dos Santos, C.; and Gatti, M. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 69–78.

Johnson, R.; and Zhang, T. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058* .

Moraes, R.; Valiati, J. F.; and Neto, W. P. G. 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 40(2): 621–633.

Socher, R.; Huval, B.; Manning, C. D.; and Ng, A. Y. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 1201–1211.

Socher, R.; Pennington, J.; Huang, E. H.; Ng, A. Y.; and Manning, C. D. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 151–161.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Tang, D.; Qin, B.; and Liu, T. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1422–1432.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30: 5998–6008.

Wang, X.; Jiang, W.; and Luo, Z. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2428–2437.

Wang, X.; Liu, Y.; Sun, C.-J.; Wang, B.; and Wang, X. 2015. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1343–1353.

Xu, J.; Chen, D.; Qiu, X.; and Huang, X. 2016. Cached long short-term memory neural networks for document-level sentiment classification. *arXiv preprint arXiv:1610.04989* .

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.