Research on Fine-grained Sentiment Classification*

Wang Zhihui$^{[0000-0003-4612-7632]}$, Wang Xiaodong, Chang Tao, Lv Shaohe, and Guo Xiaoting

National Key Laboratory of Parallel and Distributed Processing, College of Computer Science and Technology, National University of Defence Technology, Changsha 410073, China;
{wangzhihui13,xdwang,changtao15,shaohelv,guoxiaoting18}@nudt.edu.cm

Abstract. Aiming at the fine-grained sentiment classification that distinguishes the emotional intensity, the commonly used dataset SST-1 is analyzed in depth. Through the analysis, it is found that the dataset has serious problems such as data imbalance and small overall scale, which seriously restricts the classification effect. In order to solve the related problems, data augmentation method is adopted to realize the optimization of the dataset. The IMDB and other data which are relatively homologous to the original dataset are annotated, and the focus is to expand the categories with fewer numbers. By this way, the problem of data imbalance is effectively alleviated and the original data scale is expanded. Then, based on the Bidirectional Encoder Representations from Transformers (BERT) model, which has good overall performance on natural language processing, the benchmark classification model is built. Through multiple comparison experiments on the original dataset and the enhanced data, the influence of the deficiency of the original dataset on the classification effect is verified. And, it is fully demonstrated that the enhanced data can effectively improve the test results and solve the problem of large differences in performance between different categories well.

Keywords: Fine-grained · Sentiment classification · Data imbalance · Data augmentation · BERT.

1 Introduction

Text classification is a quite important module in text processing. And its application is also very extensive, including: spam filtering, news classification, part-of-speech tagging, sentiment classification and so on [1]. For the sentiment classification, the coarse-grained classification without distinguishing the emotional intensity can already have good effects, and the accuracy can basically

* Supported by National Key Laboratory of Parallel and Distributed Processing, College of Computer Science and Technology, National University of Defence Technology.

reach more than 80%. However, for the more fine-grained multi-category sentiment classification, there are two problems. First, the number of related datasets is small, and the overall size is small. Second, the research on this problem is relatively scarce, making the effect of fine-grained sentiment classification much worse. And it is difficult to meet the needs of practical applications. When expressing emotions, people's emotional attitudes are not simply happy, annoying or neutral. Even in the same emotional polarity, there is a big difference in intensity. In the aspects of public opinion control, hot spot analysis, and object evaluation, the effects of different intensity emotional texts are not uniform. Therefore, the research on fine-grained sentiment classification that distinguishes emotional intensity has important practical significance.

With the vigorous development of deep learning technology, there have been scholars introduced it into the sentiment classification. It has been proved that the deep learning models can achieve excellent results and become the mainstream trend to solve this problem. In related researches, the commonly used datasets are mainly about the data of movie reviews, and they are basically coarse-grained sentiment classification. For the fine-grained sentiment classification, the lack of data makes the performance of the classification model limited. At the same time, subjectivity has a great influence on the division of emotional intensity. Especially for ambiguous texts, it is difficult to give a label that is satisfactory to all parties. And it also has a great impact on classification performance.

In this paper, the fine-grained sentiment classification that distinguish emotional intensity is analyzed and studied. By the data enhancement method, the classification effect is improved. And the problem of uneven effect of different categories is effectively improved. In this paper, the existing fine-grained sentiment classification dataset is analyzed in depth, and the problems of unbalanced data and small data size are found. In order to solve related problems, the data enhancement method is adopted. The original dataset is expanded and optimized by using data that is relatively homologous to the original dataset. By this method, the problem of data imbalance is effectively alleviated. Finally, the benchmark classification model is constructed by using BERT [2]. Through multiple comparison experiments, it is fully demonstrated that the enhanced data can effectively improve the classification results, and solve the problem of large performance difference between different categories. And the validity of our method is verified.

2 Related Work

2.1 Fine-grained Sentiment Classification

The term "fine-grained" in the fine-grained sentiment classification can have multiple meanings. It can refer to different attributes or aspects of the object being judged, such as performance, appearance, quality in the comments of the goods. And it can also refer to the fine-grained intensity of emotion and the finer division of emotional tendencies. The fine-grained sentiment classification studied in this paper refers to the latter, the fine-grained intensity of emotion. Generally, it can be divided into five categories: very positive, positive, neutral, negative, and very negative.

Fine-grained Sentiment Classification Dataset In sentiment classification, there are many datasets available, but these datasets usually only contain two categories, positive and negative or subjective and objective. The public dataset of fine-grained sentiment classification is very rare. Currently, only the SST-1 dataset of Stanford Sentiment Treebank has been widely used. This situation has caused great difficulties in the research of fine-grained sentiment classification. The Stanford Sentiment Treebank is an extension of the movie review dataset constructed by Pang et al. [3]. The original dataset includes a total of 10662 movie review data and each movie review is a short sentence. The source of the data is the ROTTEN TOMATOES website. It provides film-related reviews, information and news. For the original dataset, half of the data is annotated positive and the other half is negative. Each piece of data is extracted from a long movie review and reflects the overall view of the reviewer. Then, the data is re-processed using the Stanford Parser and annotated by the Amazon Mechanical Turk. The processed data contains two datasets. One is fine-grained SST-1 datasets, it is composed of five categories: very positive, positive, neutral, negative, very negative. The other is the SST-2 dataset, it only contains two categories: positive and negative. In the study of sentiment classification problems, these two datasets are widely used as criteria for testing model classification ability.

Research on Fine-grained Sentiment Classification Most classification models have a classification accuracy of 80% or higher on the two-category dataset. Compared with the sentiment classification datasets with only two categories, the fine-grained sentiment classification effect studied in this paper is much worse. After the SST-1 data set is proposed, Socher et al. experiment with two traditional machine learning methods, Support Vector Machine and Naive Bayes [4], with accuracy of 40.7% and 41.0%, respectively. Then, the recurrent neural network is used, the standard recursive neural network [5] achieved a accuracy of 43.2%. Meanwhile, they propose the Recursive Neural Tensor Network (RNTN) and achieved a accuracy of 45.7%. Irsoy et al. [6] propose the Deep Recursive Neural Network (DRNN). This model refreshes the best effect on this dataset, and the accuracy reached 49.8%. Convolutional neural network [7] and recurrent neural network are commonly used in the field of natural language processing, and have achieved good results in this dataset. The accuracy of the Convolutional Neural Network (CNN) model proposed by Kim et al. [8] reaches 47.4%. And the Dynamic Convolutional Neural Network (DCNN) proposed by Kalchbrenner et al. [9] achieves a accuracy of 48.5%. Yin et al. [10] propose a Multichannel Variable-size Convolutional Neural Network (MVCNN), and it achieves the accuracy of 49.6%. Tai et al. [11] conduct experiments with Long Short-Term Memory (LSTM), Bidirectional LSTM and Tree-LSTM, and obtained accuracy of 46.4%, 49.1%, and 51.0%, respectively. The Linguistically Bidirectional LSTM (LR-Bi-LSTM) model proposed by Qian et al. [12] achieve the accuracy of 50.6%. In addition, Zhou et al. [13] combine convolutional neural network and recurrent neural network to solve classification problems. The Bidirectional LSTM with Two-Dimensional Convolutional Neural Network (BLSTM-2DCNN) model achieved the best performance on this dataset, and the accuracy reaches 52.4%.

2.2 Pre-training Model - BERT

BERT is a pre-training model released by Google in 2018. It has made breakthroughs in 11 natural language processing (NLP) tasks, including text classification tasks. And it is known as Google's strongest NLP model. For a specific task, it is only necessary to fine tune the BERT. It is a multi-layer, bidirectional Transformer [14] encoder based on fine-tuning. And it is trained using two unsupervised predictive tasks, Mask LM [15] and Next Sentence Prediction. Currently, the BERT model has been open sourced and released a variety of models of 12 and 24 layers for researchers to apply and further improve.

3 Research on Fine-grained Sentiment Classification Based on SST-1 Dataset

This paper first analyzes the SST-1 dataset in depth. Through analysis, it is found that the data is not balanced and the overall scale is small. In order to solve the problem of the original dataset, the data enhancement method is proposed to expand and optimize the dataset. By this way, the problems of data imbalance and small overall size are effectively alleviated. Then, the benchmark classification model is constructed by using BERT, and the effectiveness of the method is verified by comparison experiments.

3.1 Dataset Analysis

The SST-1 dataset is divided into three parts: training set, dev set and test set. The training set contains 8854 data, the dev set contains 1101 data, and the test set contains 2210 data. The specific composition of each part is shown in Figure 1.

The average length of the data is roughly the same, and the number of words per data is about 19. However, as can be seen from Figure 1, there is a large difference in the number of different categories, and there is a serious imbalance in the data. For the training set, the negative and positive categories account for the largest proportion of more than 25%. While the number of very positive and very negative categories is small, and the proportion of neural category is relatively small. The distribution of data in the dev set and the test set is basically consistent with the training set. In the process of data annotation by the relevant personnel of our research group, according to the quantitative statistics, the ratio of the strong emotional review data and the weak emotional

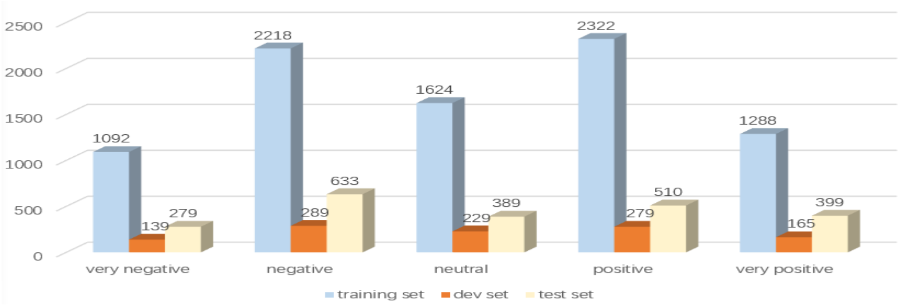

Fig. 1. Composition of SST-1 dataset.

review data and the neural is about 10:2:1. Therefore, the distribution of data in each category in the dataset is different from the actual distribution. When judging a movie, the audience usually publishes reviews to express their emotions when there is a clear emotional attitude. Therefore, the actual distribution is that the proportion of the data of the strong emotional attitude is more, while the data of the weak emotional attitude is less.

3.2 Optimization Method Based on Data Enhancement

Problem Analysis and Method Establishment Data imbalance is a common problem in machine learning, and it will affect the effect of the model. Usually, the category with more data can achieve better results. At the same time, the data of its adjacent category will also be affected and the classification effect will be reduced. Therefore, whether the data is balanced or not plays an important role.

It can be seen from the analysis in Section 3.1 that there is a serious imbalance problem in the SST-1 dataset. And it will inevitably affect the classification effect. For this problem, common solutions include oversampling, undersampling, changing classification algorithms, and cost-sensitive learning. Meanwhile, we can see that the SST-1 dataset not only has the problem of data imbalance, but also the overall size is small. Take the SST-2 dataset as an example, the total amount is about 10000, and there are nearly 5000 data in each category. However, the SST-1 dataset has five categories, but the total amount is only about 10000. For fine-grained sentiment classification, the difference between emotions of the same polarity and different intensity is much smaller than the coarse-grained. During the annotation process, the group members also have different opinions on the labels of some data, and the classification is more difficult. Therefore, we think that we can use data enhancement methods, use more data to train classification models and improve the model's ability to classify fine-grained sentiment. Referring to the situation of other multi-class datasets, and taking into account the difficulty of the classification, the current goal is to expand the amount of data to 100000. Meanwhile, the expansion target will be dynamically changed according to the change of classification effect during the expansion process. In order to maintain consistency, other attributes of the original data set should be as unchanged as possible, and the data is expanded on this basis. The categories with small number are the key of the data expansion. This method not only solves the problem of data imbalance, but also increases the scale of the dataset and provides data support for subsequent research.

Expansion Method In order to maintain data consistency, the movie review dataset is used when expanding the data. There are two selected data sources. One is the English IMDB dataset, and the other is the Chinese Douban movie review dataset. For raw data, each piece of data is a piece of text that expresses the emotions of the reviewer. In this paper, the same processing method as the original data set is adopted. A short sentence that reflects the overall opinion of the reviewer is selected from each review. Then, the short sentence is annotated to become new data. The specific process is shown in Figure 2.

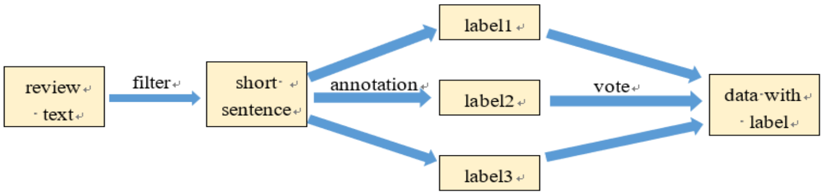

Fig. 2. Process of annotation.

3.3 Benchmark Classification Model Construction

As a powerful pre-training model, BERT has achieved good results in many NLP tasks. To verify the effectiveness of the proposed method, this paper uses this model to fine-tune and build a benchmark classification model. For the processing of data, it is usually necessary to convert the text into a vector. Then the text features are extracted through various types of network structures to obtain higher-level vectors. Finally other structure is added to obtain the final result according to specific tasks.

Text Features Extraction Based on the above process, this paper first uses the BERT to extract features from the original text. In the selection of the model, the pre-training model of 24-layer scales are used for its better effect. With this model, we can obtain high-level sentence vectors that can better represent the semantics of text. In the process of generating sentence vector, this paper uses the tool - bert-as-service. The output of the penultimate layer is selected to prevent the model from being too close to the two pre-training tasks of the BERT. Finally, fine-tuning is performed on this basis. For the 24-layer model, the output corresponds to a 1024-dimensional vector for each text.

Construction of Classification Model After completing the extraction of the text features and obtaining the higher level sentence vector, the next step is the construction of the classification model. Since the obtained sentence vector can represent the semantics of the text well, so three fully connected layers are selected for the specific classification model. For the first fully connected layer, the 1024-dimensional sentence vector is compressed to 256 dimensions. Then the 256-dimensional vector is further compressed to 12 dimensions. At the same time, to prevent over-fitting of the model, dropout is added in the process. For the last layer, the 12-dimensional vector is reduced to 5 dimensions. Finally, the specific category to which the text belongs is obtained according to the largest component of the 5-dimensional vector.

4 Experiment Results and Analysis

In order to verify the effectiveness of the data enhancement method, we performed experiments on the original dataset using 24-layer BERT model and corresponding classification model according to the method in Section 3.3. By analyzing the experiment results, it is verified that the imbalance of the dataset does have an impact on the classification results. In this paper, we adopt the data enhancement method to solve the problem of data imbalance, and uses the same classification model and the same test set to perform experiment. Through experiments, better classification results are obtained on the optimized dataset. And the problem of uneven effect in each category is effectively improved.

4.1 Classification Effect on SST-1 Dataset

Firstly, the classification model based on 24-layer BERT is used to solve the classification problem of SST-1 dataset. The training set contains a total of 8544 training data and five categories. Through training, the model achieved the accuracy of 48.8% on the dev set. The accuracy of each category and the confusion matrix is shown in Table 1.

In the dev set, from the overall point of view, the negative and positive categories have the most number of data, and the accuracy reaches more than 60%. The accuracy of these two categories far exceeds the accuracy of other categories and the whole. Although the neutral category is not the least, the effect is the worst of all categories.

Then, the two categories with the highest accuracy are analyzed. In the training set, both negative and positive categories have the largest number of data, and they are more fully trained, and the model is biased towards these two categories. Therefore, these two categories have the highest accuracy and also affect the effect of adjacent categories.

Table 1. Confusion matrix of dev set.

real value predicted value	very negative	negative	neural	positive	very positive
very negative	35.5%	10.1%	4.1%	0.6%	1.2%
negative	**52.2%**	**64.4%**	38.5%	11.3%	6.5%
neural	6.5%	9.7%	14.9%	7.0%	2.4%
positive	5.1%	15.1%	**39.5%**	**66.5%**	**49.4%**
very positive	0.7%	0.7%	3.1%	14.6%	40.4%

Although the number of neutral category is more, the effect is the worst. There are three main reasons. First, during the training process, the category data is relatively small, the model is not fully trained on this category of data. Second, the discrimination of the neutral category itself is difficult. The third and most important reason is that the two adjacent categories are the most. The model is biased towards the two adjacent categories during the training process, and more data is judged as two adjacent categories.

4.2 Effect Comparison after Data Enhancement

To compare the effects on the original dataset, we train the model on the original dataset, the accuracy on the test set is 50.6%, and the specific results of the model is shown in Table 2.

Table 2. Confusion matrix of test set(24-layers BERT).

real value predicted value	very negative	negative	neural	positive	very positive
very negative	30.1%	9.3%	3.3%	0.2%	0.3%
negative	**59.5%**	**66.8%**	**42.2%**	8.4%	3.3%
neural	6.5%	14.1%	22.4%	6.9%	2.5%
positive	3.2%	9.5%	30.6%	**69.4%**	**51.4%**
very positive	0.7%	0.3%	1.5%	15.1%	42.6%

Classification Effect on Big-scale Equilibrium Data Then, the classification model is used to train on the large equilibrium training set, and the classification effect is tested on the original test set. Through training, the model achieved the accuracy of 52.3% on the test set. And the effect has basically reached the current optimal performance. The problem of large difference between different categories is effectively improved and the effectiveness of the proposed method is confirmed. The specific classification effect is shown in Table 3.

compared with the original data set, the difference between the different categories of effects is reduced, and the overall accuracy is also improved. Meanwhile,

Table 3. Confusion matrix (data size: 10000).

real value predicted value	very negative	negative	neural	positive	very positive
very negative	**54.8%**	17.9%	8.5%	1.2%	1.3%
negative	30.5%	**51.2%**	27.5%	4.3%	2.8%
neural	9.0%	24.0%	**39.6%**	13.7%	3.0%
positive	3.9%	5.1%	20.8%	**59.8%**	37.8%
very positive	1.1%	1.9%	3.6%	21.0%	**55.1%**

the effect of the neutral is still the worst, below 40%. For a review, it is usually biased towards some emotion, which makes it more difficult to judge this category. And this is one of the research points to further improve the classification effect.

While the overall effect and the differences between the different categories have improved, the classification effects of the positive and negative have declined. The reason for this phenomenon is mainly the data with blurred boundaries. After training on the unbalanced data set, the model is more biased towards these two categories. However, after the training data is balanced, the model's bias towards these two categories is significantly reduced. Because the training data is not enough, the model can not classify data well enough. When the data with blurred boundary is encountered, it can only be judged as two adjacent categories in an average but less accurate manner. Therefore, the classification results of data with fuzzy classification boundaries are different from before and the effect of these two categories is reduced.

5 Conclusion

According to the fact that the effect of fine-grained sentiment classification problem is relatively poor compared with coarse-grained sentiment classification, this paper finds out the problem of unbalanced data and small data size by analyzing the commonly used dataset SST-1. In order to solve related problems, this paper uses data enhancement method to optimize the dataset. A more balanced and larger dataset is constructed and the problem of data imbalance is effectively alleviated. This paper builds a benchmark classification model based on BERT. And the effectiveness of the method is verified by comparison experiments, the classification effect is improved, and the problem of large difference between different classification effects is effectively improved. However, the current data size is still not large enough, and there is still a big gap between the overall classification effect and the effect of coarse-grained sentiment classification. In the future, other data sources will be utilized to further expand the data. At the same time, practical applications will be combined and more detailed evaluation indicators will be developed to explore the fine-grained sentiment classification problem in more depth.

References

1. Huang Z, Tang X, Xie B, et al. Sentiment Classification Using Machine Learning Techniques with Syntax Features. International Conference on Computational Science & Computational Intelligence. (2015).
2. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805. Last accessed 14 May 2019.
3. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the ACL. Ann Arbor, 115–124 (2005).
4. Ding Z , Xia R , Yu J , et al. Densely Connected Bidirectional LSTM with Applications to Sentence Classification. https://arxiv.org/abs/1802.00889. Last accessed 14 May 2019.
5. Socher R , Pennington J , Huang E H , et al. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011. Edinburgh, 151–161 (2011).
6. Hadji I , Wildes R P . What Do We Understand About Convolutional Networks?. https://arxiv.org/abs/1803.08834. Last accessed 14 May 2019.
7. Cardie C . Deep recursive neural networks for compositionality in language // International Conference on Neural Information Processing Systems. Montreal, MIT Press, 2014: 2096-2104.
8. Kim, Y. Convolutional Neural Networks for Sentence Classification // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014. Doha, 1746–1751 (2014).
9. Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences. Annual Meeting of the Association for Computational Linguistics, ACL 2014. Baltimore, 655–665 (2014).
10. Yin W, Schütze H. Multichannel Variable-Size Convolution for Sentence Classification // Proceedings of the Nineteenth Conference on Computational Natural Language Learning, CoNLL 2015. Beijing, 204–214 (2015).
11. K.S. Tai, R. Socher, C.D. Manning. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. Annual Meeting of the Association for Computational Linguistics, ACL 2015. Beijing, 1556–1566 (2015).
12. Qiao Qian, Minlie Huang, Jinhao Lei, and Xiaoyan Zhu. Linguistically regularized lstms for sentiment classification. https://arxiv.org/abs/1611.03949. Last accessed 14 May 2019.
13. Zhou P, Qi Z, Zheng S, et al. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. Computational Linguistics, COLING 2016. Osaka, 3485–3495 (2016).
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. Annual Conference on Neural Information Processing Systems, NIPS 2017. Long Beach (2017).
15. Taylor W L . "Cloze Procedure": A New Tool For Measuring Readability. The journalism quarterly, **30**(4), 415–433 (1953).