Sentiment Classification Based on RoBERTa and Data Augmentation

Xiaoyi Wang[1,2], Siyuan Xue[1,2], Jie Liu[2,3*], Jing Zhang[1,2], Jincheng Wang[4], Jianshe Zhou[2]

[1]School of Literature, Capital Normal University, Beijing 100048, China
[2] China Language Intelligence Research Center, Beijing 100048, China
[3] School of Information Science, North China University of Technology, Beijing 100144, China
[4] Beijing Institute of Control and Electronics Technology, Beijing 100038, China
2220101049@cnu.edu.cn, xuesiyuan987@163.com, liujxxxy@126.com

Abstract: Sentiment analysis is one of the hot research topics in natural language processing, aiming to analyze the subjective sentiment expressed in text through inductive reasoning. Common deep learning-based sentiment classification models often require the integration of word embedding techniques. However, conventional approaches to generating word embeddings struggle to effectively capture the bidirectional semantic features of text. RoBERTa, with its unique attention mechanism and masked technique, enables the capture of more comprehensive features in the text. In this paper, we introduce the RoBERTa model to the task of Chinese comments sentiment classification and address the limited data issue by employing data augmentation methods to expand the dataset. Through a series of experiments conducted on three different datasets, we demonstrate the effectiveness of the sentiment classification model that combines RoBERTa with data augmentation methods.

Keywords: Sentiment analysis; RoBERTa; Data augmentation

1 Introduction

As a subfield of natural language processing, sentiment classification is the analysis and inference of the subjective emotional aspects of a text through text analysis[1]. For example, it involves analyzing comments on movies and TV shows based on personal experiences, evaluating online shopping products based on user experiences, and commenting on hot topics on social platforms like Weibo and WeChat. These online comments contain rich emotional colors, such as subjective, objective, positive and negative, etc. With the development of the information age, constructing efficient Chinese sentiment classification models using deep learning techniques has become a hot topic in sentiment classification research.

The task of sentiment classification is to identify and extract subjective information hidden in data. In the field of intelligent business, it can be used to understand users' interests and preferences, optimize products and services, and in the field of public opinion monitoring, it can provide timely insights into public sentiment and optimize government services. Sentiment classification of comments can be considered as text classification. Firstly, comment data is obtained through online platforms and saved in textual form. Based on the length differences of online comments, they can be classified into word-level, sentence-level, and discourse-level categories. Sentiment classification of online comments can also be categorized based on the number of classes, such as binary sentiment classification (positive and negative), ternary sentiment classification (positive, neutral and negative)[2], and multi-class sentiment classification (joy, anger, sadness, happiness, surprise, etc.). This study focuses on the binary sentiment classification of comment sentences using deep learning models.

With the maturity of deep learning theory and technology, pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) [3] have achieved better performance than traditional deep learning networks in many domains. Therefore, we introduce BERT and its variant RoBERTa model[4] into the Chinese sentiment classification task and conduct comparative experiments on multiple neural networks. The results show that the RoBERTa model is more suitable for Chinese comment text classification, which involves single-text classification. Moreover, considering the limited amount of data, we improve the experimental results by using data augmentation techniques.

2 Related work

Sentiment classification primarily involves analyzing, processing, summarizing, and inferring the subjective emotional aspects of a text. Traditional classification methods include recommendation methods based on CNN[5], RNN[6], and LSTM[7]. Kim[8] proposed a deep learning model for sentiment analysis using CNNs with different convolutional filter sizes. Wang et al.[9] applied an attention-based LSTMs model for aspect-level sentiment analysis. Zhang et al.[10] presented a microblog sentiment classification method that utilizes a dual attention mechanism. By applying dual attention, sentiment phrases and the original complete text are inputted together into BiLSTM. Liang Bin et al.[11] proposed a multi-attention convolutional neural network model for sentiment analysis of specific targets, which

2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS) | 979-8-3503-0442-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCIS59572.2023.10263002

allows for a more parallelized input of text information, resulting in improved accessibility.

Both RNN and CNN-based sentiment classification models require the conversion of raw text into word vectors for processing. The most popular word vector conversion model currently is Word2Vec[12], but it cannot handle the issue of polysemy. On the other hand, the Transformer-based encoder network structure used in BERT can easily address this problem. The self-attention mechanism used in the encoder allows sentences to learn the relationships between words, dynamically assigning different weights to each word based on the sentence context. This enables accurate and efficient extraction of sentence semantics in different contexts. Li et al.[13]used BERT's pre-trained language model to represent stock comments at the sentence level, and then fed the obtained feature vectors into a classifier for sentiment classification, achieving higher results compared to mainstream recurrent neural networks. Pota Marco et al.[14] pre-trained the BERT model using a corpus composed of Twitter data and applied it to sentiment classification in the Italian language.

Although current research has significantly improved sentiment classification accuracy, we have observed that BERT, especially the RoBERTa model, has not been extensively studied in the field of sentiment classification for Chinese comments. Therefore, this study proposes a new approach that combines RoBERTa with data augmentation for sentiment classification research.

The contributions of this paper are as follows:

1. We introduce the RoBERTa model into Chinese sentiment classification for comment text and remove the Next Sentence Prediction task. The results demonstrate that the RoBERTa model is more suitable for the single-text classification task of Chinese comments.

2. We construct a RoBERTa sentiment classification model with a classifier specifically designed for binary sentiment classification tasks in this study, achieving good results.

3. To address the issue of limited experimental data, we introduce data augmentation. Comparative experiments confirm the feasibility of data augmentation methods in Chinese sentiment classification tasks for comments.

3 Approach

This section will present the method for Chinese sentiment classification of comments. The specific process is illustrated in the Figure 1. Firstly, the dataset is obtained and preprocessed and the preprocessed text is tokenized and vectorized to obtain input tokens. The RoBERTa model is utilized to extract semantic features from the input tokens. Subsequently, sigmoid classification is performed, and the sentiment polarity is outputted.

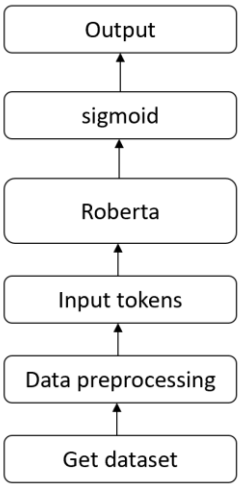

Figure 1 Overall flowchart

3.1 Data Preprocessing

Due to the nature of neural networks, it is necessary to preprocess the text to ensure that the input text vectors have the same dimensions. To preserve the text information as much as possible and meet the requirements of the model, we set the length of each text to 25 Chinese characters (counting an English word as one Chinese character). Any part of the text exceeding 25 characters is discarded, and any text with a length shorter than 25 characters is padded with 0.

3.2 BERT Model

BERT is an improvement based on the Transformer model, and it utilizes the encoder module of the bidirectional Transformer to build the model. The core idea is consistent with the Transformer model, which combines the relationships between each word in the text and other words, removes the limitation of distance, explicitly represents the dependencies between the current word and the rest of the words in the sentence, fully integrates the contextual information of the sentence, and better captures the semantic information of the sentence. It also allows for parallel processing. The network structure is illustrated in Figure 2.

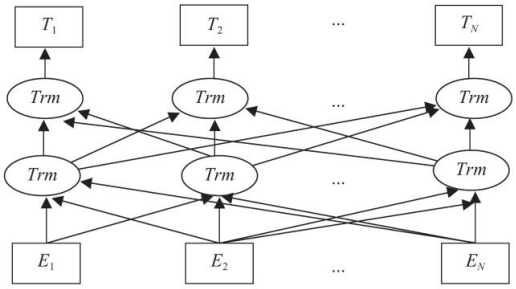

Figure 2 Structure diagram of BERT model

The bidirectional nature of the BERT allows each word in the model to capture the meaning of other words in the sentence. However, this creates a problem for prediction tasks as the model already has access to the entire sentence. To address this issue, two methods are

proposed from the input sequence perspective: (1) Masked Language Model: In the training set, approximately 15% of the words are randomly masked. Specifically, 80% of the time, a word is replaced with the "[MASK]" token, 10% of the time it is replaced with another random word, and 10% of the time it is kept as the original token. (2) Next Sentence Prediction: By randomly replacing some sentence pairs, the model is trained to predict whether two segments of text are consecutive. This helps the model learn the relationships between sentences. Both tasks are simultaneously performed, and the overall loss value is computed. These methods ensure that the model is not overly reliant on the surrounding context and can effectively capture the meaning of individual words and their relationships within the sentence.

3.3 Sentiment Classification Model

The RoBERTa model used in this paper is an optimized version of the BERT model. While the overall structure is the same as BERT, there are differences in training conditions and some details. RoBERTa features a more comprehensive MASK mechanism, removes the next sentence prediction task present in BERT, increases the amount of training data, improves computational power, and extends training duration, resulting in stronger generalization capabilities. The RoBERTa model is constructed by stacking multiple layers of the Transformer model as its main structure. The multi-head attention mechanism allows for more comprehensive and efficient semantic representation. The incorporation of residual connections and normalization operations ensures continuous optimization of model parameters without the risk of vanishing gradients. Finally, an additional classifier is added, the model outputs the corresponding class through a sigmoid classifier.

4 Experiments

In this section, we introduce the details of our experiments, including datasets, results as well as analyses.

4.1 Datasets

The experiments were conducted using the Chinese microblog datasets Nlpcc2013 and Nlpcc2014, which are publicly available from the International Conference on Natural Language Processing and Chinese Computing (NLPCC). In terms of label processing, the data with the "none" label were removed, and the remaining six labels were transformed. The positive labels "like," "surprise," and "happiness" were converted to label "0," while the negative labels "fear," "anger," and "sadness" were converted to label "1." The text data underwent denoising after label processing to obtain the final experimental data. Nlpcc2013 consists of approximately 5,000 data points, while Nlpcc2014 consists of around 18,000 data points. The specific data breakdown is as follows:

Table I Overview of the dataset

Dataset	Total Samples	Positive Samples	Negative Samples
Nlpcc2013	4937	2675	2262
Nlpcc2014	17704	9014	8690

4.2 Setting

This experiment utilized the TensorFlow deep learning framework and was implemented using the Python programming language. To achieve better experimental results, the BERT and RoBERTa models were pre-trained models specifically designed for Chinese processing tasks, as developed by Cui et al.[15]. These models were combined with classifiers suitable for sentiment classification tasks in this study. The BERT Chinese model has a size of 364M, while the RoBERTa Chinese model has a size of 1.1G.

To ensure better comparability of the experiments, we trained a word vector model using the Word2Vec toolkit with a 1.5G Chinese Wikipedia corpus. The final word vector model obtained a size of 620M, which caters to the requirements of traditional deep learning networks for word vectors during input.

4.3 Evaluation metrics

In the experiment, we used commonly adopted evaluation metrics to assess the results: Accuracy (A), Precision (P), Recall (R), and F1-score. Accuracy represents the proportion of correctly predicted samples to the total number of samples. Precision measures the ratio of correctly retrieved instances to the total number of instances retrieved. Recall represents the ratio of correctly retrieved instances to the total number of instances that should have been retrieved. The F1-score is a harmonic mean that combines both precision and recall, as shown in Formula 1:

$$F1 = 2 * P * R / (P + R) \tag{1}$$

4.4 Results

In this experiment, a comparison was made with several baseline deep learning models. The baseline models used in the experiment included TextCNN[16], a model for processing basic text; BiLSTM[17], BiGRU[18], BERT and RoBERTa models, which were retrained specifically for Chinese tasks. The test set and training set were extracted from the original data in a 2:8 ratio. The specific experimental results are shown in Tables II and III.

From the tables above, we can observe that bidirectional recurrent networks perform better overall than unidirectional networks and also outperform the baseline TextCNN model. However, the BERT model shows significant improvements across all metrics. Additionally, the RoBERTa model demonstrates slight improvements over the BERT model in overall performance. In the

Nlpcc2013 dataset, RoBERTa achieved a 0.5% increase in accuracy and a 0.2% increase in recall compared to the BERT model. In the Nlpcc2014 dataset, RoBERTa achieved an improvement of 0.08% in accuracy and 0.31% in recall. These results demonstrate the successful application of the RoBERTa model in Chinese sentiment analysis for comment classification.

Table II Experimental results of nlpcc2013 dataset

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
TextCNN	79.25	81.50	80.44	80.97
LSTM	78.44	78.51	83.58	80.97
BiLSTM	79.25	81.15	81.00	81.07
BiGRU	80.36	82.83	81.00	81.90
BERT	86.34	87.90	87.08	87.49
RoBERTa	86.84	88.58	87.27	87.92

Table III Experimental results of nlpcc2014 dataset

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
TextCNN	83.23	79.72	87.86	83.59
LSTM	83.82	80.35	88.33	84.15
BiLSTM	83.90	79.00	91.11	84.63
BiGRU	85.65	86.22	83.91	85.05
BERT	89.95	90.32	88.85	89.58
RoBERTa	90.03	88.74	91.06	89.88

4.5 Ablation study

In this section, we analyze the impact of training set size on model performance and propose methods to increase the training dataset size for validation.

4.5.1 Impact of training set size

We used the RoBERTa model for experiments on the Nlpcc2013 and Nlpcc2014 datasets. For these different datasets, we randomly selected 20% of the data as the test set. The remaining data was divided into four different proportions: 1/80, 1/8, 1/2, and 1/1, and used as the training set for experiments. The specific experimental results and corresponding metrics are shown in Tables IV to V .

Based on the experimental results above, we can conclude that as the number of training set samples increases, there is an improvement in all the metrics across the datasets. This demonstrates that increasing the size of the training dataset leads to better model performance.

Table IVI Nlpcc2013 dataset

Number of Train Set Samples	Accuracy (%)	Precision (%)	F1 (%)
49	54.96	54.91	70.90
493	80.57	76.20	84.13
1974	86.34	85.51	87.89
3949	86.84	88.58	87.92

Table VII Nlpcc2014 dataset

Number of Train Set Samples	Accuracy (%)	Precision (%)	F1 (%)
177	50.97	49.80	66.42
1770	87.57	86.67	87.32
7081	89.78	88.68	89.60
14463	90.03	88.74	89.88

4.5.2 Data augmentation

Data Augmentation is a method of generating more data by applying certain transformations under the constraint of limited data. EDA (Easy Data Augmentation) includes simple augmentation techniques such as synonym replacement, random insertion, random deletion, and random swapping. Back-translation involves expanding the data by translating the original text language into another language (such as English) and then translating it back to the original language. In our experiments, we combined EDA and back-translation with the RoBERTa model. EDA expanded the dataset by a quarter for each augmentation technique, while back-translation involved expanding the data by translating from Chinese to English and back to Chinese. The specific expansion details are shown in Table.

Table VIII Training set expansion table

Approach	Models	Original	Expanded
EDA	Nlpcc2013	3949	7898
Back-translation	Nlpcc2013	3949	7898
EDA	Nlpcc2014	14163	28326
Back-translation	Nlpcc2014	14163	28326

The specific experimental results on the nlpcc2013 dataset and nlpcc2014 dataset are shown in Table VII and Table VIII, respectively.

Table VII Comparison of nlpcc2013 dataset

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
RoBERTa	86.84	88.58	87.27	87.92
EDA+ RoBERTa	87.04	88.76	87.45	88.10
Back-translation+ RoBERTa	87.35	88.40	88.56	88.48

Table VIIIX Comparison of nlpcc2014 dataset

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
RoBERTa	90.03	88.74	91.06	89.88
EDA+ RoBERTa	90.96	90.75	90.65	90.70
Back-translation+ RoBERTa	90.99	89.79	91.93	90.85

According to the experimental results, we can see that both data augmentation methods, when combined with the RoBERTa model, outperform the baseline RoBERTa model in overall performance. This demonstrates the effectiveness of these two data augmentation methods for the RoBERTa model in Chinese microblog sentiment classification tasks. Furthermore, under the experimental data settings in this paper, the back-translation method performs slightly better than EDA in this task.

5 Conclusions

This paper proposes a sentiment classification method based on the RoBERTa model and data augmentation. We build a sentiment classification model by leveraging the semantic extraction part of the RoBERTa model as the foundation and constructing a classifier suitable for the sentiment binary classification task. To address the issue of insufficient experimental data, we introduced two data augmentation methods, EDA and back-translation. Through a series of experiments, the effectiveness of the sentiment classification method combining RoBERTa and data augmentation is demonstrated. In the future, we will explore more data augmentation methods and delve into effective approaches for Chinese sentiment ternary classification and multi-class classification tasks.

Acknowledgements

This work is supported by the General project of the 14th Five Year Scientific Research Plan of the State Language Commission (YB145-16), China Post doctoral Science Foundation (2022M722231), National Natural Science Foundation of China (62076167) and National Key Research and Development Program of China (2020AAA0109700).

References

[1] Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends[J]. Knowledge-Based Systems, 2021, 226: 107134.

[2] Hussein D M E D M. A survey on sentiment analysis challenges[J]. Journal of King Saud University-Engineering Sciences, 2018, 30(4): 330-338.

[3] Kenton J D M W C, Toutanova L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of naacL-HLT. 2019, 1: 2

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[5] Liao S, Wang J, Yu R, et al. CNN for situations understanding based on sentiment analysis of twitter data[J]. Procedia computer science, 2017, 111: 376-381.

[6] Wang X, Jiang W, Luo Z. Combination of convolutional and recurrent neural network for sentiment analysis of short texts[C]//Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers. 2016: 2428-2437.

[7] Liang Jun, Chai Yumei, Yuan Huibin, et al. Sentiment Analysis Based on Polarity Shift and LSTM Recursive Network.,2015,29(05):152-159.

[8] Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. ACL, Doha, Qatar (2014).

[9] Wang, Y., Huang, M., Zhao, L., Zhu, X.: Attention-based LSTM for Aspect-level Sentiment Classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing 2016, pp. 606–615. ACL, Austin, Texas (2016).

[10] Zhang Yangsen, Zheng Jia, Huang Gaijuan, et al. Micro-blog emotion analysis method based on Dual attention model [J]. Journal of Tsinghua University (Science and Technology),2018,58(02):122-130.

[11] Liang Bin, Liu Quan, Xu Jin, et al. Target Specific Emotion Analysis Based on Multi-Attention Convolutional Neural Networks [J]. Journal of Computer Research andDevelopment,2017,54(08):1724-1735..

[12] Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[13] Mingzheng Li, Lei Chen, Jing Zhao, et al. Sentiment analysis of Chinese stock reviews based on BERT model[J]. Applied Intelligence, 2021(prepublish).

[14] Pota Marco, Ventura Mirko, Catelli Rosario, et al. An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian.[J]. Sensors (Basel, Switzerland),2020,21(1).

[15] Cui Y , Che W , Liu T , et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[J]. 2020.

[16] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In EMNLP 2014. 1746–1751.

[17] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.

[18] Cho K , Merrienboer B V , Gulcehre C , et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.