Fine grained sentiment analysis based on Bert

Chao Chen¹, Xiaoli Hu²*, Huibing Zhang¹, and Zhaoyu Shou³

¹Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
²Teaching Practice Department, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
³School of Information and Communication, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China

Abstract. The fine-grained sentiment mining based on online review data is helpful to analyze user pain points and improve user experience and business marketing. A fine-grained sentiment analysis model based on Bert is proposed. Commodity feature words and emotion words are extracted by dependency parsing; the polarity of emotion was analyzed by the combination of Bert and emotional words; through the corresponding emotional polarity of feature words, we can get the emotional tendency of commodity features. The experimental results in movie review data set and film review data set show that, the accuracy and F1 value of Bert with feature words and emotional words were 94.67% and 94.55%, which were better than other models.

1. Introduction

When consumers purchase goods on the e-commerce platform, they will generate corresponding comments. These comments contain users' feelings about the different attributes of the product. It is a hot topic to analyse and mine these unstructured comments. In the past, the research objects of affective analysis were mostly long texts or sentences to judge the overall emotional tendency of the text or sentence. However, in the actual comment text, a comment may contain different opinions corresponding to different emotional polarity. For example, the emotional polarity of the two attributes of "photo taking" and "price" in "mobile phone photo taking is very good, but the price is too expensive". Therefore, fine-grained sentiment analysis can mine more detailed emotions of users, get users' satisfaction with different attributes of products, and analyse users the pain points, so as to help consumers provide decision support and help businesses with accurate marketing.

Fan et al. [1] used large-scale review data to analyse user pain points, the emotional polarity is analysed by searching dictionaries and counting word frequency. Although the user pain points can be obtained, the feature extraction and sentiment analysis methods used are too simple, and the accuracy is low. Our studying is based on this improvement, in order to provide a more accurate and efficient method for user pain point analysis. Zhao et al. [2] used statistical review objects and emotional words to co-exist, combined with the dependence of the two, emotion graph is constructed to extract the comment objects and emotional words. Wang et al. [3] constructs chunking rules with feature and opinion words to extract feature point pairs. Saru et al. [4] extracted target words and Opinion Words Based on graph collaborative sorting algorithm. However, the above methods do not consider the
extraction of implicit objects, which leads to incomplete extraction. In the processing stage, the sentence is segmented to realize the unification of sentence structure. Feature words and emotion words are extracted by pairs of rules, and the process includes the processing of default objects.

Pang et al. [5] used naive Bayes, maximum entropy classification and SVM to analyse the emotion of long text, and summarized the problems in emotion classification task. Liu et al. [6] used naive Bayes classifier to classify emotions on big data sets, and designed big data analysis system to analyse millions of movie reviews, and achieved results. The above machine learning method has achieved good results, but it needs a lot of manual rules. Mikolov et al. [7] used paragraph vectors. This learning method is unsupervised. It can express texts of different lengths with vectors. Vector is represented by learning the relationship between text contexts. However, the traditional machine learning methods have the problems of high dimension and sparse features.

With the development of deep learning and attention mechanism, better classification results have been achieved in affective analysis tasks. Kim et al. [8] proposed TextCNN algorithm to deal with natural language problems using convolutional neural network, and found that convolutional neural network has good experimental effect. Hocheriter et al. [9] used LSTM, a special structure of RNN, to build a model, which can overcome the problems of RNN gradient disappearance. Because LSTM can consider the relationship between text contexts. Singh et al. [10] used forward LSTM and backward LSTM to form Bi LSTM, which can better capture bidirectional semantic dependency. Thanks to the development of pre training models on large-scale corpus, the pre training language model Bert [11] learns text semantics by masking prediction, good results have been achieved on 11 datasets of multiple tasks. In this paper, we use Bert as the prototype to extract the feature words and emotional words which contain the main content of the sentence into the model, so as to improve the learning ability of the model features and better realize the sentiment polarity discrimination of the comment text.

2. Model Introduction

![Model Flow Chart](image)

Figure 1 shows the main steps of fine-grained sentiment mining. For the convenience of description, some terms are introduced below. Suppose that all the comment sets for the product are marked \( D = \{d_1, d_2, ..., d_m \} \), \( d_i \) is the \( i \)-th comment data, \( m \) is the total number of comments. The set of characteristic words is \( T = \{t_1, t_2, ..., t_n\} \), \( t_i \) is the \( i \)-th feature word, and \( n \) is the total number of feature words; Typical features \( E = \{E_1, E_2, ..., E_i, ..., E_j\} \), \( E_i \) is the \( i \)-th feature; The set of emotional words is marked as \( Q = \{q_1, q_2, ..., q_n\} \). Emotional polarity set \( P = \{p_1, p_2, ..., p_n\} \) (\( p_i \in \{\pm 1\} \)) corresponding to features. The typical feature \( T_i \) corresponds to the emotional score \( S_i \).

2.1. Data Preprocessing

Firstly, the comment data is cleaned to remove useless comments, and punctuation and space are standardized. Then, the NLP processing tool [12-13] is used to segment, segment, tag part of speech, analyse dependency syntax, and remove stop words. The Processed comment data set \( L = \)
\{l_1, l_2, ..., l_j, ..., l_n\} is obtained. Finally, the word vector after segmentation is trained by word2vec [14], and the vector representation of each word is obtained.

2.2. Extraction of Feature Words and Emotional Words

The NLP tool was used to count the word frequency and TF-IDF value of nouns, gerunds and noun groups in all comment data. When the noun, gerund and noun group in the sentence meet the corresponding threshold, the word is added to the feature word set as a feature word. Starting from the extracted high-frequency feature words, the corresponding modifiers are found and added to the emotional word set through the modification relationship. The dependency relationship and part of speech collocation rules are used to extract feature words and emotional words, as shown in Table 1.

| Dependency                  | Part of speech collocation rules |
|-----------------------------|----------------------------------|
| s-v relation (SBV)          | noun + adjective, noun + verb    |
| verb-object relation (VOB)  | verb + noun, verb + verb         |
| State-middle relation (ADV) | adjective + verb, adverb + adjective |
| Dynamic complement relation (CMP) | verb + adjective, verb + classifier |

From the above collocation rules, we can know the corresponding position of emotional words in the sentence. The nouns with modifying relation in sentence structure are extracted as feature words and added into set T. Then, the emotional words in the sentence are extracted as the candidate emotional words by searching the emotion dictionary, and then the qualified candidate emotional words are added to the emotion word set Q by combining the dependency relationship. The feature words and emotion words are extracted, Add word pair set \(P = [T, Q]\) according to sentence position.

Because of the randomness of the comment text, many feature words and emotional words are missing. Such as "expensive" and "hot" in mobile phone reviews, expensive is generally to modify the price, hot hand means that the mobile phone battery heating problem. In these cases, feature words can not be extracted directly. For the word pairs that lack feature words or emotional words, the co-occurrence frequency of word pairs is used to improve and supplement.

2.3. Sentiment Classification of Bert

The classification of text emotion by Bert mainly relies on the prediction task of cover words. When the model conceals different words, the prediction effect of the model is different. Compared with the covered prepositions and conjunctions, emotional words have a greater impact on classification accuracy. To solve this problem, we add feature words and emotional words to improve the attention mechanism to emotion. The extracted feature words and emotional words are added to the end of the sentence and input into the model for pre training. When the downstream task is emotion classification task, more attention should be paid to emotional words in the pre training stage, which is conducive to better obtaining the emotional features of sentences. Suppose a sentence sequence \(s\) is masked by the word \(W_{\text{word}}\), and the corresponding representation \(X_{\text{word}}\) is extracted by Transformer, for masking word prediction task, the loss function is defined as shown in Formula 1.

\[
loss_i = \sum_i L(x^i_{\text{mask}})F(x^i_{\text{mask}})
\]  

(1)

Where \(L(x)\) is the loss function of the i-th masked word X in the masked prediction task, \(F(x)\) is the weight of the word X. The higher the weight, the greater the influence on the masking prediction task. With the addition of feature words and emotional words, the emotional words which have a greater impact on the sentence are more likely to be masked, which improves the attention of emotional words.

The next sentence prediction (NSP) task in bert is to randomly select the next sentence from the text corpus and input it into the model to predict the next sentence with a 50% probability, 50% probability
is used to predict the next sentence of the original text. When two sentences have the same or similar semantics, it can better help the model to complete the prediction task. A short sentence with feature words and emotional words is added to the original sentence to help the model predict the next sentence more accurately. Therefore, it can reduce the predicted loss function value in NSP tasks and further improve the classification effect. The combination loss function of MLM task and NSP task is the minimum, and the model reaches the optimal value. Therefore, adding feature words and emotional words can improve the ability of model learning semantics from two aspects. A classification layer is added to the transformer output of the model for sentiment analysis. The result of the analysis is used as the emotional polarity $P = \{p_1, p_2, ..., p_n\} (p_i \in \pm 1)$ of the corresponding feature word of the sentence.

2.4. Analysis of Typical Characteristics of Commodities

The vectors corresponding to feature words are extracted to represent $w_i = [r_{i1}, r_{i2}, ..., r_{ik}]$ for clustering. The cosine similarity of typical features and feature word vectors is calculated, and the formula is as follows (2). Typical features are used as the initial category to start clustering. The feature words with the largest similarity with the typical features are merged. The combination needs to meet the constraint conditions and repeat until the cluster termination conditions are met. For example, the typical features of mobile phones such as "screen" and "resolution", "battery" and "battery life" have the same meaning.

$$\text{sim}(w_x, w_y) = \frac{\sum_i (r_{xi} \ast r_{yi})}{\sqrt{\sum_i r_{xi}^2} \ast \sqrt{\sum_i r_{yi}^2}}$$ (2)

The text $L = \{l_1, l_2, ..., l_i, ..., l_n\}$ input model corresponding to feature words is used to classify emotion. The emotional polarity $p_i (p_i \in \pm 1)$ corresponding to the feature words is obtained. The characteristic fraction $S_i$ corresponding to the typical feature is calculated as formula (3).

$$S_i = \frac{\sum_i p_i}{n}$$ (3)

3. Analysis of Experimental Results

In order to verify the performance of the fine-tuning Bert model in emotion classification tasks, we selected two data sets of typical emotion classification domain to conduct experiments. Select one of the goods to analyse the typical characteristics.

3.1. Experimental Data

Table 2 shows information about the two datasets. Among them, the Chinese film review data set (CMD) contains 50000 comments from different movie users, which is a two category review data with obvious tendency. The Chinese product review data set (CGD) contains the review data of 5W products. The data label was divided into two categories (0 / 1), with 50% positive and negative comments. The model training was divided into 7:2:1 ratio.

| parameter name | CMD     | CGD     |
|----------------|---------|---------|
| field          | movie   | goods   |
| classification | 2       | 2       |
| Train          | 35000   | 35000   |
| Dev            | 10000   | 10000   |
| Test           | 5000    | 5000    |
| Len(<80)       | 80%     | 85%     |
3.2. Experimental Design and Evaluation Index
Parameters of the computer used in the experiment GPU: 16g-teslav100, CPU:8Coers, RAM: 32GB, Disk: 100GB. In order to verify the effectiveness of the pre training model. Three contrast models are selected for the text, which are TextCNN, LSTM and Bi-LSTM. The comparison model and the fine tuning Bert model proposed in this paper are tested on two datasets in Table 2. Precision and F1 were used as evaluation indexes, such as formula (4) - (6), TP is the number of true positive cases in the classification results, FP is false positive cases, FN is false counterexamples.

\[
P = \frac{TP}{TP + FP} \tag{4}
\]

\[
R = \frac{TP}{TP + FN} \tag{5}
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \tag{6}
\]

The hidden vector dimension of TextCNN and LSTM models is set to 256, the size of batch is 150, the dropout rate is set to 0.5, the maximum input length of data set is set to 75, and the learning rate is 0.005. Table 3 shows the settings of the Bert parameter.

| Table 3. BERT parameters |
|--------------------------|
| name | value |
| optimizer | Adam |
| Learning rate | 5e-5 |
| Weight decay | 0.01 |
| warmup | 0.1 |
| epoch | 20 |
| batch | 16 |
| eval | 500 |

3.3. Results and Analysis
Table 4 and table 5 show the accuracy and F2 values of each model on the two datasets.

| Table 4. experimental results on CMD dataset |
|---------------------------------------------|
| model | Precision(%) | F1(%) |
| TextCNN | 83.61 | 83.34 |
| LSTM | 78.27 | 78.21 |
| Bi-LSTM | 80.61 | 80.59 |
| BERT | 89.31 | 89.05 |
| BERT-2 | 90.10 | 89.25 |

From the experimental results in Table 4, it can be seen that the best fine-tuning model with feature words and emotional words achieves the maximum accuracy and F1 value on the CMD data set.

| Table 5. experimental results on CGD dataset |
|---------------------------------------------|
| model | Precision(%) | F1(%) |
| TextCNN | 79.57 | 79.77 |
| LSTM | 74.47 | 75.39 |
| Bi-LSTM | 78.81 | 78.52 |
| BERT | 93.78 | 93.64 |
| BERT-2 | 94.67 | 94.55 |

The experimental results in Table 5 show that the accuracy and F1 value of Bert and its fine tuning model are far ahead in CGD dataset. It shows that the model can learn the semantics of the text in different contexts. The results show that the classification effect of Bert is worse than that of CGD
data set. By observing the characteristics of the data set, we can find that the movie reviews in the CMD contain the information of the person's name, place name and so on. The results show that better is not good at covering up long entity names, and it is not good at dealing with long text, which leads to different accuracy and F1 value of Bert in two datasets. TextCNN, LSTM and Bi LSTM are less effective than bert because they introduce a lot of prior knowledge in the pre training stage, and only need to fine tune the specific data set during training. At the same time, the powerful attention mechanism transformer can extract text features more effectively and help model learn text semantics. The accuracy and F1 value of the fine-tuning model with feature words and emotional words are improved by 0.89% and 0.91%, which verifies that emotional words and feature words can improve the classification effect.

3.4. Analysis of Typical Characteristics of Commodities
Using the review data of a brand mobile phone to analyse the typical characteristics of the product. The extracted feature words are clustered and scored to get the final score of the feature. Several features with high user attention are selected for analysis, as shown in Table 6.

| features    | fraction |
|-------------|----------|
| screen      | -0.0344  |
| battery     | -0.0215  |
| camera      | 0.0918   |
| price       | -0.0022  |
| appearance  | 0.0129   |

The mobile phone is a high-end flagship mobile phone focusing on taking photos. It can be seen from the table that the scores of screen, battery and price are negative, while the scores of camera and appearance are positive. It shows that the users are dissatisfied with the screen, battery and price of the products, but they are fond of the camera and appearance. It is in line with the characteristics of good photo taking, beautiful appearance, poor battery life, unsightly hole screen and expensive price. The analysis of typical features is in line with the actual situation, which shows that our method is effective.

4. Conclusion
This paper explores the application of pre training language model Bert in emotion classification task. It is proposed to add emotional words and feature words to improve the ability of Bert to learn text semantics. Combining the classification results with the characteristics of goods, the analysis of typical characteristics is helpful to analyse user pain points, improve user experience and business marketing. In the next step, we will improve the analysis method, and apply the method to the research of user pain point analysis, in order to get a perfect user pain point analysis process.

Acknowledgments
This research was funded by the National Natural Science Foundation (61662013, 61967005) and Guangxi Innovation-Driven Development Project (Science and Technology Major Project, AA17202024).

References
[1] Fan W H and Xu J. (2018) Analysis of user pain points based on emotional computing of online user reviews - Taking mobile reviews as an example. Intelligence theory and practice. 288: 94-99.
[2] Zhao Q, Wang H , Lv P, Zhang C.(2014)A Bootstrapping Based Refinement Framework for Mining Opinion Words and Targets. In:the 23rd ACM International Conference. Shang Hai pp.1995-1998.
[3] Wang S G and WU S H.(2012)Feature-Opinion Extraction in Science Spots Reviews Based on Dependency Relation. Journal of Chinese Information Processing, 26:116-121.

[4] SARU and KETKI B M.(2017) A new approach towards co-extracting opinion-targets and opinion words from online reviews. In: International Conference on Computational Intelligence & Communication Technology.Ghaziabad.pp.1-4.

[5] Pang B, Lee L, Vaithyanathan S.(2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Philadelphia PA.pp.79-86.

[6] Liu B, Blasch E, Chen Y, Shen D, Chen G.(2013) Scalable sentiment classification for big data analysis using naive Bayes classifier. In: IEEE international conference on big data.Silicon Valley.pp.99-104.

[7] Le Q and Mikolov T.(2014) Distributed representations of sentences and documents. In: International conference on machine learning.New York.pp.1188-1196.

[8] Tymoshenko, K., Bonadiman, D., & Moschitti, A.(2016) Convolution Kernels: Feature Engineering for Answer Sentence Reranking. In: north american chapter of the association for computational linguistics.San Diego.pp.1268-1278.

[9] Hochreiter, S., & Schmidhuber.(1997). Long short-term memory. Neural Computation, 9: 1735-1780.

[10] Singh B, Marks T K, Jones M, Tuzel O, Shao M.(2016) A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. In: 016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas.pp.1063-6919.

[11] Devlin J, Chang M W, Lee K, Toutanova K.(2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.arXiv: Computation and Language.

[12] Ting L, Wanxiang C, Zhenghua L I.(2011) Language Technology Platform. Journal of Chinese Information Processing.

[13] Krishnan P, and Jawahar C V.(2018) HWNet v2: An Efficient Word Image Representation for Handwritten Document. Document Analysis & Recognition.

[14] Mikolov, T., Chen, K., Corrado, G. S., Dean, (2013). Efficient Estimation of Word Representations in Vector Space. international conference on learning representations.