Sentiment Analysis and Fine-Grained Feature Mining of Hotel Reviews Based on BERT-BiLSTM

Chenyu Wang[1*]

[1]School of Future Technology, South China University of Technology, 511400, Guangzhou, China

Abstract. With the rapid development of online tourism platform, hotel reviews published by users on the platform have become important data resources that affect consumer decision-making and improve service quality. For the sentiment analysis of Chinese hotel reviews, this paper proposes a hybrid neural network model combining BERT and BiLSTM, and verifies it on the hotel review data set. At the same time, combining LDA topic modeling and TF-IDF keyword extraction technology, fine-grained feature mining is carried out for negative comments. The experimental results show that the BERT-BiLSTM hybrid model achieves 93.10% accuracy and 93.19% F1 score on the test set, which is significantly better than the single BiLSTM and BERT model. LDA revealed that the negative comments mainly focused on the two core themes of front desk service and check-in process, hardware facilities and comprehensive experience. TF-IDF keywords further quantified the user's focus. This study provides a complete analysis framework from emotion classification to problem positioning for the hotel industry, which has important theoretical value and practical significance.

1 Introduction

With the popularity of online travel platforms, hotel reviews published by users on the platform have become important data resources that affect consumer decisions and improve service quality. For consumers, these comments are important references for making occupancy decisions; For hotel managers, these comments are an effective way to find problems and improve service quality.

Sentiment analysis aims to automatically identify and extract the emotional tendencies contained in the text. In recent years, the mainstream methods of text sentiment analysis include machine learning, deep learning and ensemble learning [1]. Machine learning methods such as support vector machine (SVM) and Naive Bayes are widely used in emotion classification tasks. The deep learning model represented by long short-term memory (LSTM) has achieved remarkable results with its powerful sequence modeling ability. In order to further overcome the limitations of LSTM in dealing with long-distance dependence, bi-directional long short-term memory (BiLSTM) introduces a bidirectional structure, which can better capture context information.

Raw text data are not directly interpretable by machine learning algorithms and must first be transformed into numeric feature vectors. One widely adopted approach for this transformation is word embedding, which maps words or phrases to dense vector representations. Cahyani et al. combined word embedding model word2vec with convolutional neural network (CNN) and LSTM for sentiment analysis of Indonesian language, indicating the effectiveness of the combination of word2vec and deep neural network [2]. With the rapid development of deep learning technology, the transformer model-based BERT proposed by Devlin et al. can learn rich context semantic representation, and has become the mainstream technical solution of sentiment analysis [3]. Singh et al. used the BERT model to analyze the emotion of Twitter [4].

On the basis of BERT, researchers began to explore its combination with CNN and LSTM. Abas et al. proposed a new model of BERT-CNN, which uses BERT to dynamically generate semantic vectors and input them into CNN for output prediction, proving the effectiveness of the combination of BERT and CNN [5]. Si and Wei proposed the BERT-LSTM method, which shows the effectiveness of LSTM in extracting context semantic relations [6].

In terms of fine-grained feature mining, the Latent Dirichlet Assignment (LDA) topic model can automatically summarize the core topics that users pay attention to. Chen and Wei classified public attitudes towards urban regeneration based on word cloud and LDA classification [7]. As a classic keyword weight calculation method, Term Frequency - Inverse Document Frequency (TF-IDF) can reveal the important words in a specific document set from a statistical perspective, and provide micro level data support for fine-grained analysis.

Although some progress has been made in the field of sentiment analysis of hotel reviews, there are still shortcomings. First, most studies only focus on improving the accuracy of emotion classification, and

* Corresponding author: 202364870242@mail.scut.edu.cn

lack of mining and analyzing the underlying causes of negative comments; Second, the combination of BERT and BiLSTM is relatively simple, which fails to make full use of the rich semantic information of the multi-layer output of BERT; Third, the research on fine-grained feature mining in the context of Chinese hotel reviews is not sufficient.

To solve the above problems, this paper constructs a BERT-BiLSTM hybrid model, which effectively integrates multi-level semantic features by splicing the last four layers of hidden states of BERT as input. Secondly, on the basis of emotion classification, combined with LDA topic modeling and TF-IDF keyword extraction technology, fine-grained feature mining is carried out for negative comments, and two core topics of user complaints are summarized. Finally, this study carried out experiments based on real Chinese hotel review data, verified the effectiveness of the proposed method, and provided a complete analysis framework from emotion recognition to problem location for the hotel industry.

2 Models and methods

This paper uses BiLSTM, BERT, BERT-BiLSTM models to classify emotions, and uses LDA topic model and TF-IDF keyword extraction method to mine the details of negative comments. The overall model takes the Chinese hotel comment text as the input, outputs the positive and negative emotional classification results, and further carries out fine-grained analysis on the negative comments.

2.1 BiLSTM emotion classification model

Among sequence-oriented neural architectures, BiLSTM remains a widely adopted choice due to its ability to mitigate the vanishing gradient issue commonly observed in traditional RNNs when processing long texts. The information is filtered and updated through forgetting gate, input gate and output gate in LSTM. BiLSTM is a combination of forward LSTM and backward LSTM, which can model the historical and future information of text at the same time, and more comprehensively capture the context semantic dependency. In this study, BiLSTM splices forward and reverse hidden states to obtain a complete context representation of each location. In order to get the fixed dimensional representation of the whole sentence, average pooling is applied to the output hidden states across all time steps, and then the risk of over fitting is reduced through the dropout layer. Finally, the emotion classification is completed through the full connection layer and Softmax function.

2.2 BERT emotion classification model

BERT is a pre-training language model based on transformer encoder. It learns the universal deep bidirectional semantic representation through the mask language model and the next sentence prediction task. It has significant advantages in Chinese text understanding tasks.

In the emotion classification task, BERT receives the text sequence after word segmentation, adds word embedding, position embedding and segment embedding, and inputs it into the multi-layer transformer encoder. The [cls] position vector output by the model contains the global semantic information of the whole text, which can be directly used for classification tasks. In this paper, the [cls] vector is sent to the linear classification layer after dropout regularization, and the positive and negative emotional probability distributions are output through the Softmax function. The whole model makes end-to-end fine-tuning through back-propagation in the training process, so that the pre-training knowledge can better adapt to the target field.

2.3 BERT-BiLSTM hybrid model

In order to integrate the deep semantic representation ability of BERT and the sequence modeling advantages of BiLSTM, a hybrid model of BERT-BiLSTM based on multi-layer feature stitching is proposed in this paper. First, the input text is encoded by BERT, and the hidden state of all layers is obtained. In order to integrate the features of different levels of abstraction, the hidden states of the last four levels are spliced on the last dimension to obtain the enhanced representation of each token. This operation can integrate the low-level syntactic features and high-level semantic features. Then, the spliced vector sequence is used as the input of double-layer BiLSTM. The local dependencies between adjacent tokens are further captured by LSTM. To get a fixed length sentence representation, the BiLSTM output is averaged and pooled, and then sent to the full connection layer after passing through the dropout layer. The classification probability is output through the Softmax function.

2.4 Fine-grained analysis method

2.4.1 LDA topic model

LDA is a generative probabilistic model for collections of discrete data. It assumes that each document arises from a mixture of several latent topics, where every topic corresponds to a probability distribution over words. To identify the most suitable number of topics for the corpus, the perplexity and u_mass consistency index under different number of topics (2-10) were calculated, and the points with low perplexity and high consistency were selected. The calculation formula of perplexity is:

$$\text{Perplexity} = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (1)$$

The perplexity metric is calculated using Eq. (1), where the numerator accumulates the log-likelihood of each document $p(w_d)$ as estimated by the model, and the denominator sums the word counts N_d across all M documents. A lower perplexity value indicates stronger

generalization performance of the topic model on unseen data.

u_mass consistency is an internal evaluation index based on document co-occurrence statistics, which is used to measure the semantic consistency of words within a topic. For a topic, select the T words with the highest probability, and the calculation formula of u_mass value is:

$$\text{u\_mass} = \frac{2}{T(T-1)} \sum_{i=2}^{T} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_i) D(w_j)} \quad (2)$$

In Eq. (2), $D(w_i)$ counts how many documents contain the word w_i, while $D(w_i, w_j)$ counts documents where both w_i and w_j appear together; ϵ is a small smoothing constant. A higher u_mass score reflects stronger within-topic co-occurrence patterns among the top words, indicating better semantic coherence. This study integrates perplexity and consistency with u_mass, finally selected the number of topics $K = 2$, estimated the parameters through Gibbs sampling, and each topic was described by the 10 words with the highest probability.

2.4.2 TF-IDF keyword extraction

TF-IDF is a widely used statistical measure for estimating how important a word is to a specific document within a larger collection. It combines two components: term frequency (TF), which counts how many times a word appears in a given document, and inverse document frequency (IDF), which discounts words that occur in many documents across the corpus. A word receives a high TF-IDF weight when it appears frequently in a particular document but rarely elsewhere——making it both locally significant and globally distinctive. In this study, this measure was applied to the set of negative comments identified by the classification model. By extracting the 20 words with the highest TF-IDF weights, the specific aspects that users complain about most often were pinpointed.

Through the combination of the above two methods, this paper further realizes the fine-grained feature mining of negative comments on the basis of emotion classification, which provides data support for subsequent analysis.

3 Experiment

3.1 Data source and pretreatment

This study uses a subset of hotel reviews from the public Chinese sentiment analysis corpus [8]. This part of the data consists of 5000 positive comments and 5000 negative comments. The positive and negative samples are balanced, with a total of 10000. Clean the original text, remove isolated numbers and noise words such as "supplementary comments". It is randomly divided into training set, verification set and test set according to the ratio of 6:2:2, and stratified sampling is used to ensure that the proportion of positive and negative samples in each set is consistent. Use the bert-base-chinese word breaker to encode the text into input_ids and attention_mask, and uniformly fill or truncate to the maximum length of 128.

3.2 Results and analysis

3.2.1 Model performance comparison

To assess the effectiveness of each model, this study adopted two standard metrics: accuracy and F1 score. On the test set, the experimental results of the three models (BiLSTM, BERT, BERT-BiLSTM) are shown in Table 1.

Table 1. Performance of different models on test set.

Model	Accuracy（%）	F1 score（%）
BiLSTM	89.25	88.83
BERT	92.25	92.16
BERT-BiLSTM	93.10	93.19

Table 1 clearly indicates a substantial performance advantage of BERT over BiLSTM, with improvements of 3.00% in accuracy and 3.33% in F1 score, verifying the strong semantic understanding ability of the pre-training language model in Chinese emotion analysis. Secondly, the BERT-BiLSTM hybrid model achieved the best performance in all indicators. Compared with BERT, the accuracy increased by 0.85%, and the F1 score increased by 1.03%, which proved the effectiveness of introducing BiLSTM to further strengthen the sequence characteristics.

3.2.2 Fine-grained analysis results of negative comments

In order to explore the causes of user dissatisfaction, the experiment further carried out topic modeling and keyword extraction for negative comments.

Based on 5000 negative comments from real tags, the optimal number of topics is determined by calculating the perplexity and consistency of the number of topics from 2 to 10. When the number of topics $K = 2$, the perplexity is the lowest and the consistency is the highest. With the increase of the number of topics, the perplexity gradually increases and the consistency fluctuation decreases. Therefore,$K = 2$ is selected for LDA clustering, and two core complaint themes are summarized. Their keywords and percentages are shown in Table 2.

Table 2. LDA clustering topics and keywords.

Topic serial number	Topic name	Proportion（%）	Keyword
1	Front desk service and check-in process	40.78	hotel, room, check-in, front-desk, Ctrip, waiter, no, just, let, yuan
2	Hardware facilities and comprehensive experience	59.22	room, hotel, poor, service, facilities, general, bad, breakfast,

			airport, convenience

The results showed that more than 59% of the negative comments focused on the hardware facilities and comprehensive experience. Users generally complained about the old room facilities, poor service quality, poor breakfast and inconvenient geographical location; The front desk service and check-in process accounted for 40.78%, mainly involving the front desk attitude, check-in handling efficiency, booking platform (Ctrip) experience, price (yuan) and other specific details. These two themes provide a clear direction for hotel improvement.

In addition, this study also extracted the 20 keywords with the highest weight of TF-IDF and their weights, as shown in Table 3. This is consistent with the analysis results of LDA. High frequency words such as hotel, room, service and facilities confirm the core position of hardware facilities and comprehensive experience, while words such as front desk, check-in, waiter and Ctrip echo the problems of front desk service and check-in process.

Table 3. The 20 keywords with the highest TF-IDF weight and their weights

Keyword	Weight	Keyword	Weight	Keyword	Weight
Hotel	0.0769	Room	0.0640	Service	0.0404
Commonly	0.0348	Front-desk	0.0289	Facilities	0.0284
Check-in	0.0275	Not good	0.0255	Airport	0.0248
Breakfast	0.0235	Waiter	0.0214	Ctrip	0.0200
Price	0.0193	Feeling	0.0193	Convenience	0.0188
Unable	0.0178	Environment	0.0167	Very bad	0.0166
Night	0.0164	Service attitude	0.0156		

4 Discussion

Through model comparison, this study proved the advantages of the BERT-BiLSTM hybrid model in emotion classification task. This model not only uses the dynamic semantic representation ability of BERT, but also uses the sensitivity of BiLSTM to sequence information to enhance the extraction of context features. At the level of fine-grained analysis, TF-IDF keywords quantify the user's focus from the micro level, and confirm each other with LDA topics, forming a complete analysis link from macro topics to micro keywords.

However, this study also has some limitations. First, in terms of model complexity and deployment efficiency, although the BERT-BiLSTM hybrid model has advantages in classification performance, it has a large number of parameters, and a long forward reasoning time. The TinyBERT framework proposed by Jiao et al. provides an effective way for the lightweight of such models [9]. Secondly, the interpretability and automation of topic modeling need to be improved. In this study, LDA is used for topic mining. Topic naming and semantic interpretation still rely on manual judgment, which is subjective. In the future, it is possible to introduce a new generation of topic models based on embedded clustering, such as BERTopic, and automatically generate topic tags with more semantic consistency by combining word vectors and document embedding [10]. Finally, the granularity and dimension of emotion analysis can be further refined. The current model only supports positive and negative classification, and it is difficult to distinguish emotional tendencies with different intensities or fine granularity. In the future, it is possible to build a multi category model or introduce aspect level emotion analysis to locate users' emotional attitudes towards specific aspects such as the front-desk, breakfast and facilities through the attention mechanism, so as to provide more specific improvement directions for hotel managers.

5 Conclusion

This paper constructs and compares BiLSTM, BERT and BERT-BiLSTM models for the emotional analysis task of Chinese hotel reviews. As demonstrated by the experimental results, the hybrid model of BERT-BiLSTM achieves 93.10% accuracy and 93.19% F1 score by integrating the ability of deep semantic understanding and sequence modeling, which is significantly better than the single model. Through the analysis of perplexity and consistency, the optimal number of topics was determined to be 2. LDA topic modeling revealed that negative comments mainly focused on the two core topics of front-desk service and check-in process and hardware facilities and comprehensive experience. TF-IDF keywords further quantified the user's focus. This study provides data-driven decision support for the improvement of hotel service quality, and has important practical value.

Although this research has achieved good results in sentiment classification and fine-grained mining, there is still room for improvement in the efficiency of model deployment, the interpretability of topic modeling and the granularity of sentiment analysis. In the future, it is possible to explore lightweight models to reduce deployment costs, introduce advanced theme models such as bertopic to improve interpretability, and expand to aspect level emotional analysis to provide more refined improvement suggestions for hotel managers.

References

1. Q. Y. Wang, S. C. Shi, H. J. Wang. A survey of text sentiment analysis. Software Guide, **24**(01): 193-202. (2025)
2. D. E. Cahyani, A. P. Wibawa, D. D. Prasetya, L. Gumilar, F. Akhbar, E. R. Triyulinar. Text-based emotion detection using CNN-BILSTM. 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS): 1-5. (2022)
3. J. Devlin, M. W. Chang, K. Lee, K. Toutanova. Bert: Pre-training of deep bidirectional transformers for

language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers): 4171-4186. (2019)

4. M. Singh, A. K. Jakhar, S. Pandey. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Social Network Analysis and Mining, **11**(1): 33. (2021)
5. A. R. Abas, I. Elhenawy, M. Zidan, M. Othman. Bert-cnn: A deep learning model for detecting emotions from text. Computers, Materials, & Continua, **71**(2): 2943. (2022)
6. H. Y. Si, X. Y. Wei. Sentiment analysis of social network comment text based on LSTM and bert. Journal of Circuits, Systems and Computers, **32**(17): 2350292. (2023)
7. K. H. Chen, G. Y. Wei. Public sentiment analysis on urban regeneration: A massive data study based on sentiment knowledge enhanced pre-training and latent Dirichlet allocation. Plos one, **18**(4): e0285175. (2023)
8. mlln-cn. Corpus of Chinese sentiment analysis. 2018. https://mlln.cn/2018/10/11/, 2026-03-01. (2018)
9. X. Q. Jiao, Y. C. Yin, L. F. Shang, X. Jiang, X. Chen, L. L. Li, F. Wang, Q. Liu. Tinybert: Distilling bert for natural language understanding. Findings of the association for computational linguistics: EMNLP 2020: 4163-4174. (2020)
10. M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. (2022)