# Sentiment analysis based on CNN - LSTM hotel reviews

# Sentiment analysis based on CNN-LSTM hotel reviews

**Lisha Yao**[1]

[1]School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui, China

jsjyaolisha@163.com

**Abstract.** In order to obtain the implied semantic information of hotel reviews for emotional analysis, the correlation between discontinuous words is ignored in the traditional convolutional neural network (CNN) emotional analysis. Therefore, a novel sentiment analysis method based on CNN-LSTM model is proposed. In this method, CNN is used to extract semantic features from hotel review texts, and LSTM is used to add sentence structure features to enhance deep semantic learning. This model improves the accuracy and F1 value on the CHNsenticorp-HTOL-BA-6000 hotel review data set, and can better solve the task of text sentiment analysis and discover the emotional orientation of text information.

## 1. Introduction

Text sentiment analysis is a kind of text mining. Currently, most commonly text sentiment analysis methods are based on traditional machine learning methods, including support vector machine (SVM) [1], Naive Bayes [2], K-nearest Neighbor (KNN) [3], maximum entropy method [4], etc. The process of modeling using traditional machine learning methods includes: 1) text preprocessing; 2) Text feature representation; 3) Training classifier; 4) Get classification results. Machine learning methods need to rely on a large number of manual annotation, time-consuming and laborious. In recent years, deep learning methods are common, such as CNN and LSTM. Therefore, combining the advantages of CNN and LSTM models, this paper proposes a sentiment analysis method for hotel reviews based on CNN-LSTM model.

## 2. Related work

Deep learning can autonomously learn features from a large amount of data with the help of large-scale corpora, effectively making up for the shortage of manual feature extraction. Zhu[5] proposed the convolutional neural network based comment text sentiment analysis method and achieved good results. Zhai[6] proposed a model combining hybrid neural network and conditional random field (CCF), which uses the CCF model as a classifier and fully considers the context information to enrich the features learned. SARDAR[7] proposed a wor2vec_TFIDF fusion feature representation and CNN text classification method. CNN has been widely used in sentiment classification. However, CNN can only mine local information of text and lacks the ability of long-distance dependent information extraction. The recurrent neural network can make up for this shortcoming.

Tang [8] modeled the text-level text and proposed a hierarchical RNN model. Wu[9] proposed an emotion analysis algorithm for Chinese short texts combining Selfattention and BiLSTM. Al-smadi [10] used two kinds of LSTM neural networks to achieve sentiment analysis of Arab hotel reviews. ZHANG[11] proposed an ON-LSTM-SA model based ON on-LSTM and self-attention mechanism, which achieved good results in the emotion data set.

Considering that CNN may lose a lot of information when extracting the features of hotel reviews in the pooling layer. The ordering information of the text is considered, but the semantic structure features of the text is ignored. CNN ignores the correlation between discontinuous words. This paper proposes a method of sentiment analysis of hotel reviews based on CNN-LSTM model. Experimental comparison on chNsenticorp-Htol-BA-6000 hotel review data set proves the validity of the proposed model.

## 3. Hotel review sentiment analysis based on CNN-LSTM

### *3.1. Model structure*

The model to encode the input text sentence first, coding way is to use the word vector model is used to encode, after converted into word vector said, after dealing with the LSTM model, through the convolutional neural network after get the relevant features of the sentence, finally after all connections using the classifier to finish the work of text sentiment analysis.

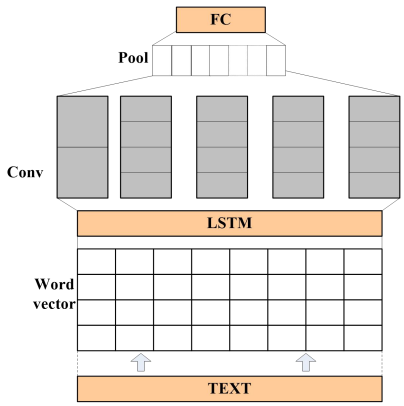


**Figure 1.** CNN-LSTM model diagram.

The sentiment analysis process of hotel reviews designed in this paper is as follows:

(1) Read the text, count the word frequency for each clause's segmentation, and limit the sentence length to 50;

(2) Word2Vec word vector model is used to transform corpus into word vector, and the sentence input embedding layer is used to form 32-dimensional word vector for each word, and the matrix of 50×32 is output.

(3) After the matrix is processed by LSTM model, 256 convolution kernels are input, and the convolution window is 3×3 convolution layer to extract features. The size of the feature matrix after convolution is $50 \times 256$. After the simplification of the maximum pooling layer, the output feature matrix is 17×256.

(4) After 32 convolution kernels with convolution Windows of 3×3, the feature matrix of $17 \times 32$ is output, and a one-dimensional vector with length of 544 is obtained after Flatten operation;

(5) Through Dropout and BN layer to prevent data overfitting;

(6) Finally, the probability of emotion classification is obtained through softmax excitation function to complete the emotion classification.

Table 1 is the model parameter Settings.

**Table 1.** Model parameter setting.

| Model design structure |
|---|
| Input: 50×32 |
| LSTM: 50×32 |
| C1: 3×3 conv, 256 |
| S1: 3×3, max-pooling |
| C2: 3×3 conv, 32 |
| Flatten |
| Dropout: 0.3 |
| BN |
| FC: 256 |
| Dropout: 0.2 |
| Softmax |

### *3.2. Data preprocessing*

(1) Data loading: call OS method to extract local data files according to different data sets or load specified data sets according to the load method of Keras.

(2) Word segmentation. If the input sequence is the whole sentence, word segmentation is generally required. You can call Jieba word segmentation tool or use keras platform's own tool.

(3) Stop words are removed, and input sequences of different lengths are converted into uniform fixed length sequences. Fixed vectors are realized for splicing with insufficient length, and words with low frequency are removed for excessively long sequences.

### *3.3. Word vector training*

The word vector is trained by using Google's open source Word2vec tool, and data set is preprocessed by word segmentation, and then the processed data set is used as the model input. In the Skpp-gram model of Google's open source Word2vec tool used in this paper, I set the size of the context window as 5, the dimension of the word vector as 50, and the sampling value as 1e-3.

## 4. Analysis of experimental results

### *4.1. Dataset*

This paper uses the ChnSentiCorp Hotel Review dataset, which contains 10,000 pieces of data. The experiment adopts 5 fold cross validation, and the following experimental data is the average of 5 tests.

### *4.2. Experimental configuration*

Table 2 is the experimental configuration of this paper.

**Table 2.** Experimental configuration.

| Experiment Tools | Configuration |
|---|---|
| Operating System | Window 10 |
| Memory | 8G |
| Programming Language | Python 3.7 |
| Deep Learning Framework | Karsa 2.4 |

### *4.3. Analysis of experimental results*

Figure 2 and 3 are the change curves of accuracy and loss values of the experimental training set and verification set of CNN-LSTM for sentiment analysis of hotel reviews.

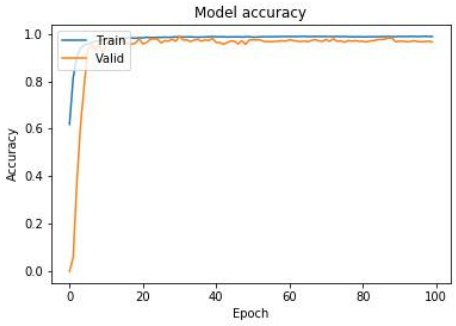


**Figure 2.** CNN-LSTM accuracy curve.

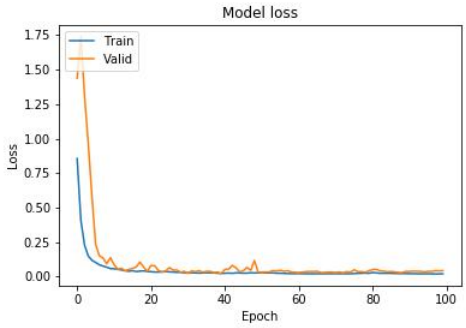


**Figure 3.** CNN-LSTM loss value change curve.

In order to further verify and compare the effects of the proposed model, the CNN model and LSTM model with the same structure were used in the experiment. Table 3 is the results of accuracy rate and F1 value.

**Table 3.** Accuracy and F1-score comparison results.

| | CNN | LSTM | CNN－LSTM |
|---|---|---|---|
| Accuracy | 0.963578 | 0.974058 | 0.979681 |
| F1-score | 0.963523 | 0.974029 | 0.979676 |

By analyzing and comparing the above experimental results, it is found that the training results of LSTM model are better than those of CNN model, because LSTM model is used to overcome the problem of long dependence. The result of CNN+LSTM training is better than that of LSTM, because after adding convolutional neural network, the whole network model not only overcomes the problem of long dependence, but also learns features. After combining features obtained by convolutional neural network with features obtained by LSTM model, better results can be obtained.

## 5. Conclusion

In view of the traditional deep learning algorithm sentiment analysis does not fully consider the text feature and input optimization problem, combining CNN and LSTM, a fusion model of CNN and LSTM, namely CNN-LSTM model, is proposed, based on CNN-LSTM sentiment analysis method of hotel review. In this method, word2VEc is used to train word vectors, and the correlation between discontinuous words is ignored when the sentiment analysis of CNN hotel reviews is completed. Therefore, LSTM is combined to extract semantic structure features of sentences. Experimental results show that the method is feasible and effective. Compared with the basic model, this method can further improve the accuracy and stability, and can better solve the task of text sentiment analysis and find the emotional orientation of text information.

## Acknowledgements

Special thanks to the following funds for their support: Key Research Project of Natural Science in Universities of Anhui Province(No.KJ2020A0782); University-level Quality Engineering Demonstration Experiment and Training Center "Big Data Comprehensive Experiment and Training Center" (No. 2020 sysxx01).

**References**

[1] LIU Jinshuo,LI Zhe,YE Xin,CHEN Jiamin,DENG Juan. Sentiment Orientation of Text: bfsmPMI-SVM [J]. Journal of Wuhan University(Natural Science Edition), 2017(63):264.
[2] XU Jun,DING Yu-xin,WANG Xiao-long. Sentiment Classification for Chinese News Using Machine Learning Methods [J]. Journal of Chinese Information Processing, 2007, 21(6):95-95.
[3] ZHANG Xiao-Fei,HUANG He-Yan. An Improved KNN Text Categorization Algorithm by Adopting Cluster Technology [J].Pattern Recognition and Artificial Intelligence, 2009, 22(006):936-940.
[4] ZHANG Lei,LI Shan,PENG Jian,CHEN Li,LI Hong-you. Feature-Opinion Pairs Classification Based on Dependency Relations and Maximum Entropy Model [J].Journal of University of Electronic Science and Technology of China, 2014(03):420-425.
[5] ZHU Ye;CHEN Shi-ping. Commentary Text Sentiment Analysis Combining Convolution Neural Network and Attention[J]. Journal of Chinese Computer Systems, 2020, 41(3):551-557.
[6] ZHAI Xueming, WEI Wei. Text sentiment analysis combining hybrid neural network and conditional random field[J]. CAAI Transactions on Intelligent Systems, 2021, 16(2): 202-209.
[7] SARDAR Parhat, MIJIT Ablimit, ASKAR Hamdulla. Kazakh Short Text Classification Based on Stem Unit and Convolutional Neural Network[J]. Journal of Chinese Computer Systems, 2020, 41(8):1627-1633.
[8] TANG Duyu, QIN Bing, LIU Ting. Document modeling with gated recurrent neural network for sentiment classification[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1422−1432.
[9] WU Xiaohua;CHEN Li;WEI Tiantian;FAN Tingling. Sentiment Analysis of Chinese Short Text Based on Self-Attention and Bi-LSTM[J]. Journal of Chinese Information Processing, 2019, 33(6):100-107.
[10] Alsmadi M,Talafha B, Alayyoub M, et al. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(8):2163-2175.
[11] ZHANG Zhong-lin,Li LIN-chuan, ZHU Xiang-qi, MA Hai-yun.Aspect Sentiment Analysis Combining ON-LSTM and Self-attention Mechanism[J]. Journal of Chinese Computer Systems, 2020, 41(9): 1839-1844.