ISSN: 2664-9640

# Sentiment Analysis of User Generated Content Based on CNN Model

Zhiyun Wang, Lingling Wang*

School of Management Science and Engineering, Anhui University of Finance and Economic, Bengbu 233000, China.

* Corresponding Author

## Abstract

**Taking the Chinese hotel review dataset generated by users on the network platform as the research object, through the TensorFlow deep learning framework and the Keras artificial neural network library, a sentiment analysis model based on CNN convolutional neural network is constructed. From the perspective of user comment data mining on the network platform, the granular analysis of the emotional polarity of user comments is carried out to provide data support for hotel management to understand users' true emotional tendencies and improve their service quality. By analyzing the results of the verification data set crawled from the website, the accuracy of the CNN model can reach up to 0.88, but there is still a lot of improvement in the accuracy of sentiment judgment for some data sets. In future research, the model needs to be further optimized to obtain a stable and highly accurate deep learning model.**

## Keywords

**Hotel reviews, user-generated content, CNN, sentiment analysis.**

## 1. Introduction

With the development of science and technology, all aspects of people's lives have been covered by the network, and the current 5G era for people has also brought great convenience to people's lives. As of December 2021, the number of Internet netizens in our country has reached 1.032 billion, and the Internet penetration rate has reached 73.0%. As of June 2021, the number of online payment users in our country reached 872 million, accounting for 86.3% of the total netizens; the number of online shopping users in our country has reached 812 million, accounting for 80.3% of the total netizens; our country's online retail sales reached 6,113.3 billion yuan. At the same time, user shopping patterns are affected by live streaming, factory direct sales, etc., which have also contributed to the growth of online consumption.

As the Internet is more and more widely used in people's lives, the amount of text information data generated by users on the Internet increases, which contains a large number of comments, opinions, and ideas with emotional tendencies. How to extract users' emotions from complex and diverse text information, judge emotional tendencies, and explore their potential content has become an important research direction in the field of natural language processing. Up to now, sentiment analysis has made great exploration and progress. The methods commonly used by domestic and foreign scholars in the research process include text classification models based on sentiment dictionaries and rules, text classification models based on machine learning, and text classification based on deep learning model. The performance indicators of machine learning-based text classification techniques mainly depend on the labelling training in the corpus and the selection of effective features, while the deep learning-based text classification techniques largely depend on the training volume of the corpus [1].

ISSN: 2664-9640

User-generated Contents is the evaluation text with the subjective emotions of users. Sentiment analysis of user-generated content generally refers to mining, parsing, summarizing and inferring user-generated content based on machine learning, and obtaining subjective information expressing users' attitudes and opinions. According to the differences in the granularity of the text to be analyzed, it can be divided into three types: chapters, sentences and words [2]. Due to the complex semantics of text information, in the analysis process, it is easy to make mistakes in the judgment or loss or other issues of sentence sentiment analysis due to factors such as the analysis granularity is too coarse, the text information is difficult to vectorize, and the emotions expressed in different contexts are different [3].

In order to solve the above problems better, based on the text classification model based on neural network, this paper will use convolutional neural network to perform sentiment analysis on user-generated comments, mainly from the collection and preprocessing of experimental data, text vectorization, the construction of the convolutional neural network, the analysis and verification of the accuracy of the model are studied, to obtain the relevant indicators such as the accuracy of the convolutional neural network model in text comments. Finally, the sentences with known emotional tags are used for verification. The given comment is based on the model's accuracy ratio and other related information.

## 2. Related work at home and abroad

By looking for relevant papers on sentiment classification methods for user-generated content at home and abroad and conducting classification discussions, it can be found that, text sentiment analysis methods are divided into three types: methods based on sentiment dictionary, methods based on machine learning, and methods based on deep learning.

The method based on dictionary and rules is to match the preprocessed words with the words in the sentiment dictionary, and judge the sentiment polarity of the text through dictionary analysis, sentence pattern analysis and other methods. Emotional dictionary refers to a collection of various words with fixed emotional tendencies. Song Guanyu et al. [4] developed words by constructing three dictionaries: a dictionary of emotional polarity (positive/negative), a dictionary of degree adverbs, and a dictionary of negative words. The sentiment judgment method is adopted, and the text is segmented, stop words removed, and sentiment weighted, so as to obtain the overall sentiment score of the text. Hou Jinfei et al. [5] calculated and analyzed the sentiment value of Su Shi's surviving online open poems by adopting the method of a single dictionary database and a compound dictionary database, and realized the sentiment analysis of online open poems. Li Jidong et al. [6] improved the performance of Chinese microblog sentiment analysis by analyzing the inter-sentence rules and sentence-pattern rules through the extended sentiment dictionary and semantic rules. Zhang s X et al. [7] first constructed and expanded the sentiment dictionary, and then obtained the sentiment value of the microblog text by calculating the weight, and realized the sentiment classification of the microblog text.

Based on traditional machine learning methods, the text is generally encoded using the bag-of-words model, and then the selected features are used to represent the entire text. Shang Yongmin et al. [8] used Naive Bayes, Support Vector Machine and Snow NLP methods to classify the sentiment of data, and obtained the optimal classification method for online comment sentiment classification. Hu Mengya et al. [9] used the idea of control variables to determine the combination of Bernoulli NB, Multinomial NB, Logistic Regression, SVC, Linear SVC, and Nu SVC to build the classifier with the highest accuracy.

Features obtained by deep learning-based methods can be directly used to predict probabilities, or can be classified using shallow classifiers such as support vector machines [10]. Xu Kangting et al. [11] used a deep learning model to extract deep-level features from the original text, key

ISSN: 2664-9640

emotional segments and emotional sets respectively, then weighted them for fusion, and finally used a classifier to judge the emotional polarity. Xue Yu [12] used the CNN-Softmax model to replace the original binary tree structure in the model, which has significantly improved the performance of English sentiment analysis. Li Yang et al. [13] combined the convolutional neural network (CNN) and the bidirectional long short-term memory network (Bi-LSTM) to give full play to the ability of CNN to extract local features and the ability of bidirectional long short-term memory network to extract text sequence features, which improved the text sentiment analysis performance.

Lexicon-based methods rely on the construction of sentiment lexicons, and the effect of sentiment analysis is proportional to the quality of the lexicon, and is usually less effective. This method needs to manually set the sentiment polarity value for each sentiment word and semantic rule in advance, and sharing a common dictionary for sentiment tasks in different fields will inevitably bring artificial errors, and the sentiment polarity labeling of sentiment words and rules requires consumes a lot of manpower. The traditional machine learning method relies on artificial feature construction, the effect is easily affected by feature extraction, and the applicability is not strong. This method is generally based on the bag-of-words model, which ignores contextual semantics and requires feature engineering. The method based on deep learning can automatically learn deep features from a large number of texts, has good sentiment analysis effect and strong model adaptability, and automatically realizes the end-to-end learning and reasoning process. In view of this, this method has become a research hotspot in recent years.

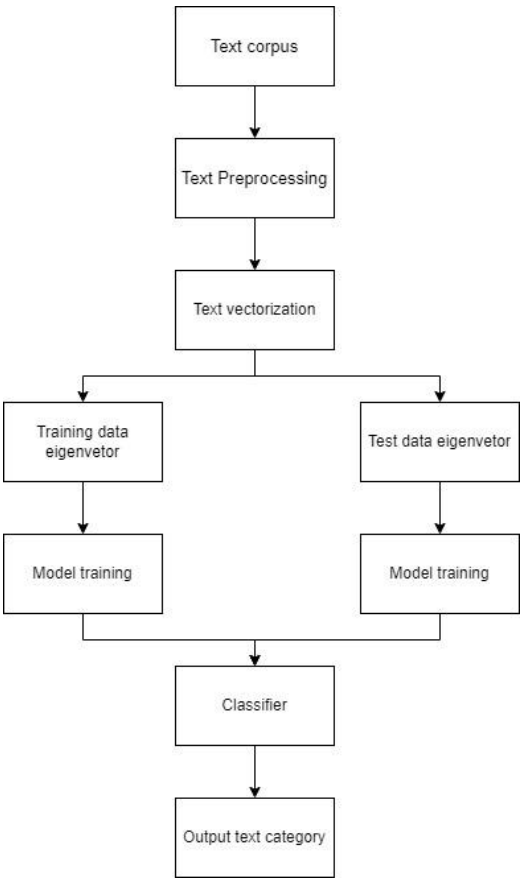


Figure 1: Flowchart of text classification

## 3. Related technical analysis

Sentiment classification of user-generated content is the process of automatically classifying user-generated text content into preset label attributes. This process mainly includes text

ISSN: 2664-9640

preprocessing, text vectorization, training data model training, and classifiers. Its general process is shown in Figure 1.

## 3.1. Text Preprocessing Technology

Text can be divided into Chinese text and English text, because spaces can be used to divide meaning between English texts, while Chinese texts cannot be divided according to spaces or single characters because they are all connected together. Therefore, compared with the word segmentation of English text, Chinese text preprocessing is more cumbersome, including but not limited to word segmentation, stop word removal, part-of-speech tagging, etc. For Chinese text, the quality of word segmentation results will also be to a certain extent. It affects the final results obtained by using the verification text data. Therefore, Chinese word segmentation technology is a key technology for Chinese text classification, and it is also the content that needs to be studied in the field of natural language processing. The current Chinese word segmentation techniques are as follows:

Word segmentation method based on string matching. This method is also called dictionary-based word segmentation method, which is one of the most widely used word segmentation techniques. If the phrase is in the dictionary, it indicates that the word segmentation is successful, otherwise, it needs to continue to divide until it successfully matches the existing dictionary content. Although this word segmentation technology is fast, there is a problem that it cannot eliminate ambiguity. That is, when the same word has completely different meanings in two sentences, the word segmentation technology will mechanically divide it into one word, which is easy to cause ambiguity.

Statistical word segmentation method. The method is to count the probability of a single word appearing together according to the text information of the corpus. The more commonly used words, the greater the probability value. The technology will set a preset probability value, and when the probability value of the combination of adjacent words is greater than the preset value, it will be divided into a word. Compared with the word segmentation method based on string matching, this technology saves the time of artificially constructing a dictionary, and the effect of this technology is better in large-scale expected word segmentation, but the corpus is an important factor that determines the word segmentation effect of this technology. Therefore, in order to use this technology to achieve better results, a large-scale training sample is required.

## 3.2. Text vectorization

Since the Chinese text content cannot be recognized by the computer, it is necessary to convert the Chinese text into a vector form with digital features that can be recognized by the computer while retaining the original semantics of the Chinese text as much as possible. This process is called text vectorization.

The text vectorization technology used in this paper is word2vec, which converts each word in the text into a vector with a unified meaning and a unified dimension, and simplifies the text vector from a high-dimensional space to a low-dimensional space. Differently, two text vectorization algorithms, CBOW and Skip-gram, are formed [14].

The CBOW algorithm predicts the target word from the context word, that is, the word in the context in which a word is located is used as input, and itself is used as output. After training in a large corpus, the vector value of the context is calculated in the projection layer and summed, and the information of the target word is output.

The Skip-gram algorithm uses the words of the context in which the target word is located as the output, and itself as the input, and uses the current word to estimate the information of the context word.

ISSN: 2664-9640

## 3.3. Classification model

According to the vectorized features of the extracted text data, these features need to be classified to solve the problem of text classification better. The existing text classification models can be roughly divided into: text classification models based on machine learning and text classification models based on deep learning.

### 3.3.1. Text Classification Model Based on Machine Learning

(1) K nearest neighbor model

In the training sample set, each data has a label. After inputting new data, select the k most similar data, and the classification with the most occurrences in the k most similar data is used as the classification of the new data. For the K-nearest neighbor model, using fewer neighbors corresponds to higher model complexity, while using more neighbors corresponds to lower model. If the extreme case is considered, that is, the number of neighbors is equal to the number of all data points in the training set, then the neighbors of each test point are exactly the same (that is, all training points), and all prediction results are also the same (that is, the most frequent occurrence in the training set category).

(2) Naive Bayes Model

Naive Bayesian models are very similar to linear models, but train faster and generalize slightly worse than linear classifiers. Naive Bayesian models are very efficient because they learn parameters by looking at each feature individually, and need to collect simple class statistics from each feature. The model works well for high-dimensional sparse data and is relatively robust to parameters. Due to the fast-training speed of Naive Bayesian models, they are often used on very large datasets where even linear models can take a significant amount of time to train.

(3) Decision tree model

A decision tree model is a widely used model for classification and regression tasks, and in essence, it learns and draws conclusions from the "if, else" problems at each layer. To construct the decision tree, the algorithm searches all possible tests to find the one that is most informative for the target variable. This recursive process produces a binary decision tree where each node contains a test. The data is divided repeatedly until each divided region (i.e., each leaf node) contains only a single target value.

### 3.3.2. Text Classification Model Based on Deep Learning

One of the main advantages of deep learning text classification models is the ability to capture the information contained in large amounts of data and use this information to build incredibly complex models. For deep learning models, as long as sufficient computing time and data are given and parameters are adjusted, deep learning models can often surpass other machine learning models. Since deep learning models perform best on homogeneous data, and homogeneous data, i.e., all features of the data, have similar meanings, so deep learning models require careful parameter tuning to obtain the best results.

## 3.4. Convolutional Neural Networks

Convolutional Neural Networks, or CNN, were initially mainly used in the field of image recognition, and has subsequently achieved good results after the introduction of text analysis. The convolutional neural network consists of an input layer, a convolutional layer, a pooling layer, a fully connected layer and an output layer. The convolutional layer is the core of the convolutional neural network. The convolution kernel flips and translates in the text vectorized matrix with a certain step size, and a convolution operation is performed each time, thereby obtaining a new matrix with feature information. The purpose of the convolution kernel is to learn the text features. Different text features will result in different convolution kernels. The feature data obtained after convolution processing is sampled by the pooling layer, and the

ISSN: 2664-9640

obtained results are transmitted to the full connection floor. The convolutional neural network has fewer training parameters and consumes a relatively short time, and can realize functions such as weight sharing and local perception.

## 4. Sentiment analysis based on CNN model

### 4.1. Experimental environment

The experiments in this paper were completed in the experimental environment shown in Table 1.

Table 1: Experimental environment and configuration

| lab environment | operating system | CPU | RAM | Programming Tools | word segmentation tool | Deep Learning Framework | word embedding training tool |
|---|---|---|---|---|---|---|---|
| Environment configuration | Windows10 | Intel(R) Core(TM) i5-9300H 2.40GHz | 8GB | Python 3.6 | jieba 0.42 | TensorFlow 1.13、Hard 2.6.0 | Word2vec |

### 4.2. Data collection and processing

Datasets: This paper uses two datasets to conduct experiments on CNN-based text classification methods. One dataset is based on the hotel review corpus compiled by Tan Songbo. There are 7766 hotel reviews in this dataset, including 5322 positive reviews and 2444 negative reviews. This paper expands on the basis of this data set, and selects a total of 10,000 review samples, including 5,000 positive samples and 5,000 negative samples, which are respectively used for positive and negative evaluations of hotels. Examples are shown in Table 2. In this paper, the review samples selected from the datasets are collected together and divided into training set and test set with a ratio of 4:1. The data of another dataset is user reviews crawled from Ctrip.com for verify the accuracy of the convolutional neural network' prediction results.

Table 2: ChnSentiCorp dataset example

| Positive evaluation | negative evaluation |
|---|---|
| A very good five-star hotel, the rooms are large and the facilities are very new. Importantly, its location is in the financial center, which is very convenient to go anywhere. I will consider staying again in the future. | Depressed!!! Angry!! I don't understand that the optical fiber is actually slower than the Shanghai Jinjiang Inn. If you want a fast internet speed at night, don't go to this place!!! |
| The room is clean, the facilities are ok, the furniture is a little old. The floor reception of the business room is good, and the price of his home is relatively low in 4 stars. | The room has never been arranged to have a frontal lake view, especially the reception level at the front desk is really poor, there are resentful women, and there are expressionless faces. |
| The hotel is very clean, the waiter will recommend me to the ladies' non-smoking floor, the facilities are also good, and the small snacks in the restaurant are also good. | To a certain extent, it is not as good as a good two-star or no-star hotel. |

Data preprocessing:

ISSN: 2664-9640

(1) Retain Chinese texts with more semantic information;

(2) The text format in the original data is GB2312. In order to facilitate the computer's identification, the text document in this format needs to be converted into a UTF-8 encoded file;

(3) Adjacent words in English sentences are connected by spaces, but adjacent words in Chinese text cannot be distinguished by spaces, so analysis and processing are required here, and the final experimental effect has a certain relationship with the quality of word segmentation;

(4) Use the Skipgram model proposed by Mikolov et al. to train word embedding, which has been implemented by the word2vec tool and can be used directly. It should be noted that in order to obtain high-quality word embedding, the size of the corpus used for training cannot be too small.

## 4.3. Experimental evaluation indicators

The experimental evaluation indicators used in this paper include precision rate, recall rate, F1 value and accuracy. The text information with positive emotions is regarded as positive text, and the text information with negative emotions is regarded as negative text, and the classification result matrix is shown in Table 3:

Table 3: Classification result matrix

| text message emotion | result | |
|---|---|---|
| | positive (positive) | Negative (negative) |
| Positive text | TP (true positive example) | FN (false negative example) |
| Negative text | FP (false positive positive) | TN (true negative example) |

Among them, TP refers to the number of positive texts that are correctly classified as positive; FN refers to the number of positive texts that are incorrectly classified as negative; FP refers to the number of negative texts that are incorrectly classified as positive; TN refers to the number of negative texts that are correctly classified as negative.

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{TN}{TN+FN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{TN}{TN+FP} \quad (2)$$

$$F1 = \frac{2*P*R}{P+R} \quad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

## 4.4. Model parameter settings

In this paper, the open-source word vector model of Baidu Encyclopedia is used to train the text information into a 300-dimensionalword vector, and the parameters of the CNN part of the convolution layer are as follows: the maximum number of words is set to 300; the height of the convolution kernel is 2, 3, and 4, respectively; The number of convolution kernels is 64; the regularization parameter is 0.6; the batch size is 20; As shown in Table 4:

Table 4: Model parameter settings

| dimension | Maximum number of words | Convolution kernel height | Number of convolution kernels | Regularization parameter | Algorithm work | batch size |
|---|---|---|---|---|---|---|
| 300 | 300 | [2,3,4] | 64 | 0.6 | 50 | 20 |

ISSN: 2664-9640

## 4.5. Analysis of experimental results

### 4.5.1. Analysis of model training results

5,000 pieces of positive emotional and negative emotional text information were analyzed, and 10,000 pieces of text information were divided into a training set and a test set according to 4:1, that is, a total of 8,000 text information in the training set, including positive emotional text information and negative emotional text information. There are 4,000 pieces of positive emotional text information, and 2,000 pieces of text information in the test set, including 1,000 pieces of positive emotional text information and 1,000 negative emotional text information. The results obtained after processing by the CNN model are as follows. As shown in Table 5 and Table 6.

Table 5: Classification result matrix after CNN model processing

| text message emotion | result | | |
|---|---|---|---|
| | Positive | Negative | support |
| Positive text | TP (true positive example): 1268 | FN (false negative example): 232 | support_pos |
| Negative text | FP (false positive positives): 729 | TN (true negative example): 771 | support_neg |

Table 6: CNN model processing results

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| POS | 0.64 | 0.32 | 0.43 | 1000 |
| NEG | 0.55 | 0.82 | 0.65 | 1000 |
| accuracy | | | 0.57 | 2000 |
| macro avg | 0.59 | 0.57 | 0.54 | 2000 |
| weighted avg | 0.59 | 0.57 | 0.54 | 2000 |

in:

1. Macro avg: Sum and average the precision, recall, and F1 values of each category.

2.Weighted avg: An improvement to macro average, considering the proportion of the number of samples in each category in the total sample.

### 4.5.2. Validation dataset result analysis

Using crawler technology to crawl 1500 pieces of positive and negative text information, divide the 3000 pieces of data into 30 groups of data, and analyze the accuracy, recall, F1 value and accuracy of these 30 groups of data. The graphs generated from the 30 sets of data are as follows. As shown in Figure 2,Figure 3,Figure 4 and Figure 5.

According to the analysis of the above pictures, it can be found that: when the accuracy rate of text information is high, the recall rate is low; the accuracy rate of positive text can reach 0.94 at the highest, and the lowest value is 0.72; the accuracy rate of negative text can reach the highest 0.92 , the lowest value is 0.2; the recall rate of positive text can reach up to 0.913, and the lowest value is 0.5349; the recall rate of negative text can reach up to 0.8929, and the lowest value is 0.5758; the F1 value of positive text can be up to 0.8750, the smallest The value is 0.6154; the maximum F1 value of negative text is 0.8846, and the minimum value is 0.3125; the accuracy of the CNN model can reach up to 0.88.

Comprehensive analysis of the CNN model shows that there is still a lot of room for improvement in the accuracy of the CNN model, and further improvement is needed to achieve more accurate sentiment prediction.

ISSN: 2664-9640

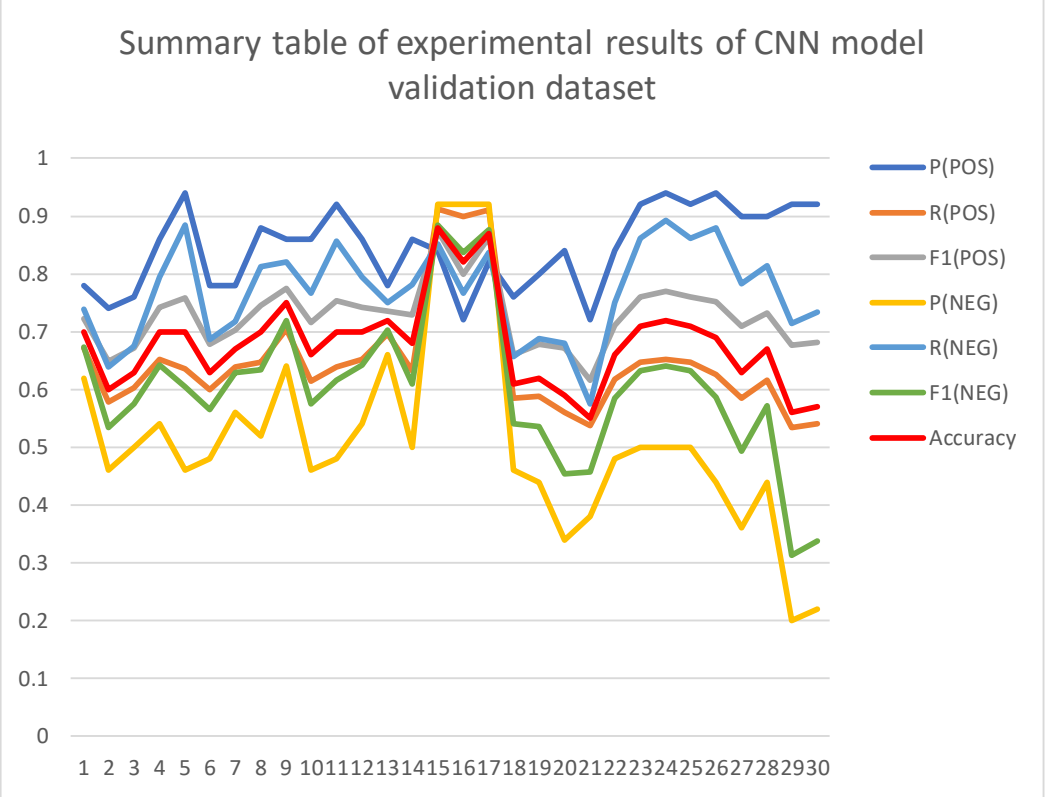


Figure 2: Summary table of experimental results of CNN model validation dataset

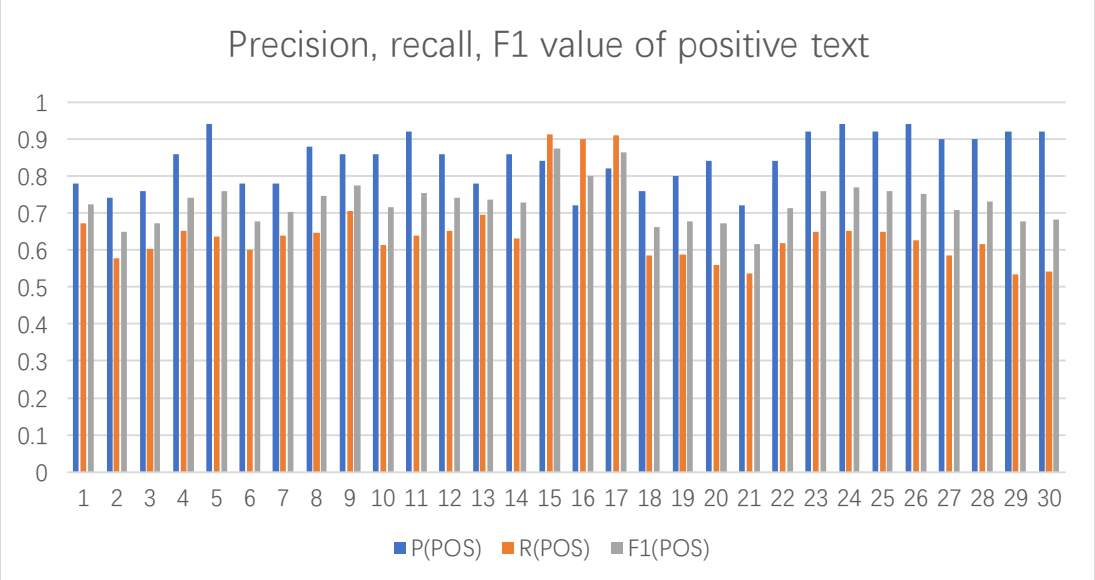


Figure 3: Precision, recall, F1 value of positive text

ISSN: 2664-9640

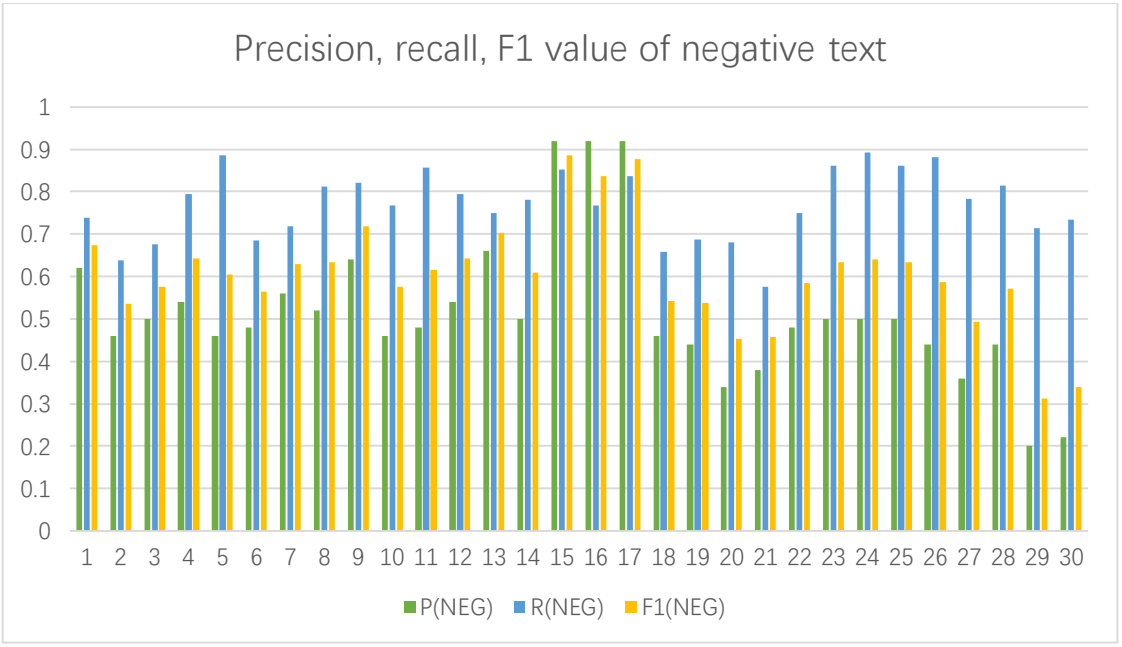


Figure 4: Precision, recall, F1 value of negative text

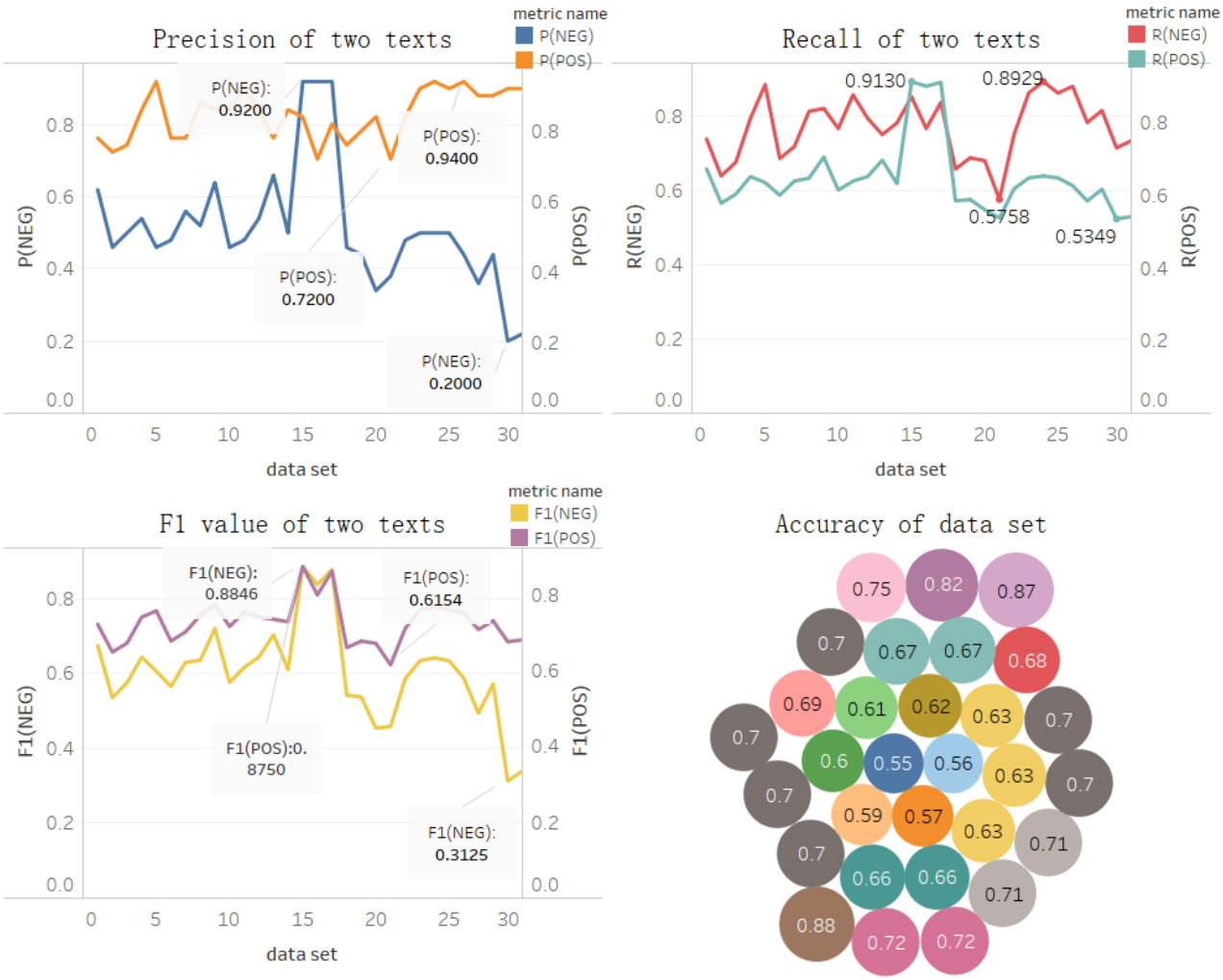


Figure 5: Precision, recall, F1 value and data set accuracy of two texts

ISSN: 2664-9640

## 5. Conclusion

This paper uses the CNN deep learning model to analyze the sentiment of hotel reviews, and uses the data set downloaded from the Internet for training and the data set collected by the crawler to verify. The results show that the CNN deep learning model is classified accurately to a certain extent. The rate can reach 0.88, but it needs to be further improved. The purpose of the CNN model implemented in this paper is to judge the emotional tendencies of the hotel reviews generated by the users of the network platform, and to conduct sentiment analysis on the reviews about the hotel on the network platform, so as to help the hotel management to evaluate the keywords of users in the process of serving customers. The extraction and judgment of emotional polarity provide an analytical method, which provides a scientific basis for hotels to understand the real evaluation and needs of customers and improve service quality in a targeted manner. Sentiment analysis of user-generated content can effectively find out whether users have a sense of identity with a certain hotel, observe users' love for the hotel, help hotel management to discover their own strengths and weaknesses, improve service levels, and enhance user satisfaction [3].

The data set used in this experiment is relatively small, there is no effective analysis for expressions, non-semantic symbols, etc., the model training time is too long, and the accuracy of the validation set is very different. In future research, the utilization of multi-parameters, the analysis of emoji information, and the optimization of models will be the next research directions. In future research, the optimized CNN model will also be compared with other neural network deep learning models, so as to obtain a deep learning model with continuously enhanced text information recognition and generalization capabilities.

## Acknowledgements

Fund project: Supported by Anhui University of Finance and Economics Undergraduate Research and Innovation Fund Project (No.: XSKY2120ZD), And the Science Research Project of Anhui University of Finance and Economics under (No.: ACKYC20085).

The author wish to thank Xu Yong, Li Feng. This work was supported in part by a grant from Li Feng.

## References

[1] Y. He, H. C. Yang, Y. Pan, S.Q. Xu: Text sentiment analysis based on improved CNN, Journal of Ping Ding Shan University, Vol. 36 (2021) No.5, p.59-62.

[2] X. J. Huang, J. Zhao: Sentiment analysis for Chinese text, Proceedings of the Communications of CCF, Vol. 4(2008) No.2.

[3] B. Li, H. L. Li, Q. Guan, Y. Liu: Fine-grained sentiment analysis of university library social network platform based on CNN-BiLSTM-HAN hybrid neural network, Journal of Agricultural Library and Information.

[4] G. Y. Song, D. Cheng, S. Zhang, W. Liu, X. W. Ding: Text sentiment score calculation model based on sentiment dictionary, Information and Computer (Theoretical Edition), Vol.33(2021) No.22, p.56-58+62.

[5] J. F. Hou, Y. D. Liang: Research on sentiment analysis of online open poetry based on dictionary database, Computer Programming Skills and Maintenance, (2022) No.2, p.42-44.

[6] J. D. Li, Y. Z. Wang: Chinese Microblog Sentiment Analysis Based on Extended Dictionary and Semantic Rules, Computer and Modernization, (2018) No.2, p.89-95.

[7] S. X. Zhang, Z. L. Wei, Y. Wang, et al: Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary, Future Generation Computer Systems, (2018) No.81, p.395-403.

ISSN: 2664-9640

[8] Y. M. Shang, Y.Q. Zhao: Online review sentiment analysis and implementation based on machine learning, Journal of Dali University, Vol.6 (2021) No.12, p.80-86.

[9] M. Y. Hu, C.J. Fan, Y. Zhu: Sentiment Analysis of Weibo Comments Based on Machine Learning, Information and Computer (Theoretical Edition), Vol.32 (2020) No.12, p.71-73.

[10] D. J. Liu, Y. H. Wang, W. F. Ling, Y. Peng, W. Z. Kong: Emotion recognition based on brain-computer collaborative intelligence, Journal of Intelligence Science and Technology, Vol.3 (2021) No.1, p.65-75.

[11] K. T. XU, W. Song. Chinese Text Sentiment Analysis Method Combining Language Knowledge and Deep Learning, Big Data, (2022), p.1-16.

[12] Y. Xue: Research on English sentiment analysis method based on NLP and deep learning methods, Electronic Design Engineering, Vol.29(2021) No.13, p.95-99.

[13] Y. Li, H. B. Dong: Text sentiment analysis based on CNN and BiLSTM network feature fusion, Computer Applications, Vol.38(2018) No.11, p.3075-3080.

[14] M. L. Xu: Research on text sentiment analysis combining sentiment dictionary and neural network, Jiangxi University of Science and Technology, (2020).