ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

# Research on Text Sentiment Analysis of E-Commerce Comments Based on Deep Learning

Qiwei Chen[1, a], Xiaochao Liang[1]

[1]School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing, China 400065, China.

[a]565335478@qq.com

## Abstract

**For the huge amount of comment data, the high cost of emotions is analyzed manually, and the LSTM is used to analyze the emotions of business comment. In order to train the model, including text de-duplication, removal of stop words, Chinese word segmentation, and a series of data processing, the data are coded into word vectors. Finally, the crawling data is used to verify the accuracy of the method. The result shows that the accuracy of the method is 75%.**

## 1. Introduction

In recent years, with the popularity of the Internet, more and more people begin to use the Internet to shop. Domestic e-commerce platforms are also competing for success. Online shopping has become one of the main consumption channels. It is also an important step for users to express their feelings on the platform of electronic commerce after purchasing goods. For these huge amounts of commentary data, we want to manually analyze the emotions contained in them, and the cost is immeasurable.

Firstly, this paper analyses the main research directions and methods of text sentiment analysis at this stage. Based on this, the main research direction of this paper is emotion classification based on LSTM network model. In order to train the model, a series of data processing is carried out. These include text de-duplication, removal of stop words and Chinese word segmentation. After data processing, it is coded into word vectors to prepare for the next experiment.

Aiming at the application of neural network in affective analysis, it is found that in the task of text affective classification, cyclic neural network has better effect than convolutional neural network. But the problem of long-term dependence of cyclic neural networks seems to be somewhat inadequate. Therefore, this paper applies the LSTM model. LSTM networks can solve the problem of long-term dependence. Experiments show that the LSTM network model has a good effect in text emotion classification.

As an important part of the field of natural language processing, the commonly used machine learning methods for text sentiment partitioning include support vector machine and naive Bayes. Both belong to shallow learning methods. In the field of text analysis with huge sample space, the above-mentioned learning methods have insufficient expressive ability compared with the complex objective function, while in the large sample space, relatively sparse text samples have higher requirements for the fitting ability of machine learning model. For the above problems, we can use deep learning of the deep non-linear network structure to solve, its principle is to superimpose the simple neuron structure together, so as to achieve the merging calculation of complex functions. In recent years, in-depth learning has been applied to emotional analysis, and its effect is much better than the general model. Zhai Donghai et al. [1] proposed a parallel processing framework to solve the shortcomings of semi-supervised recursive self-coding text emotion classification model, which has been widely recognized by domestic scholars. Zhou Jingyi et al. [2] published an academic report on Bagging method,

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

which uses convolutional neural network model as a weak classifier, and made indelible contributions to improving the accuracy of emotional classification. Ji Chaojie et al. [3] put forward a view on emotional classification of joint modification relations in deep learning. This way relies on the construction of modification. It divides a sentence into a few small sentences according to the modification relations. Each small sentence marks a fixed hierarchical category by the rhetoric relations and corresponds to these small sentences. The semantic information, the types of grades and the modification and association between them are combined to get the emotional meaning and degree of this sentence, which provides a new way of thinking for the development of the field of text emotional analysis.

Nowadays, the main ways of affective classification are text affective analysis based on affective dictionary, text affective analysis based on machine learning and text affective classification based on deep learning.The key point of emotional dictionary is emotional dictionary and rules. We need to make this part. We need to divide and organize the text, analyze the grammar, calculate the emotional value, and finally rely on the emotional value to calculate the emotional tendencies contained in the text.

Emotional analysis of foreign texts began in the late 1990s. Hatzivassiloglou and McKeown et al. [4] adjusted and perfected the large sample, and they thought about the factors affecting emotional tendency among different parts of speech.

In recent years, in order to better solve the problem of tendency analysis of features, Miao et al. [5] tabled a report on the concept of quaternion extraction in commodity evaluation, and completed the classification of characteristics grade. Because different kinds of sentences have different ways of expressing emotions, Narayanan et al. [6] started the analysis of conditional sentences. Based on temporal information, they annotated sentences in different categories, synthesized the information presented by each feature, and obtained the classification methods of different sentence patterns, which achieved satisfactory results.

Compared with foreign countries, domestic scholars are unwilling to lag behind. Xu Linhong and Lin Hongfei and other [7] sentence phrases and phrases are constructed to acquire nine semantic features that have an impact on sentence emotion. They use the form of manual joint automatic acquisition to create an emotional lexicon bank, which is a preliminary attempt in the field of emotional analysis. In recent years, Wenbin, He Tingting[8] and others have published reports on text emotion classification methods of semantic understanding, and put forward the idea of adding emotional causes to discriminating emotional tendencies. Wang Suge et al. [9] published a report on text emotional classification based on emotional dictionary. They used the intensity of feature tendency to define the weighted rough membership degree, and applied it in automobile evaluation, and obtained good classification performance.

Generally speaking, the advantages of using emotional dictionary in the analysis of text emotions are self-evident: sentence feature level, sentence level, fine granularity and high classification accuracy. But its shortcoming is also obvious, that is, it is easy to lose the important patterns hidden in the data set.

The problem of affective analysis in machine learning is text categorization. We can divide the target emotion into good and bad by judging whether the emotion is good or bad. This requires us to artificially annotate the training samples and then conduct supervised machine learning. In our learning and practice, we often see the classification models of machine learning: central vector method, naive Bayes, maximum entropy, K nearest neighbor classification and support vector machine.

Pang et al. [10,11] used machine learning to analyze the emotions of media public opinion. They used three models to conduct comparative experiments. Among them, there are maximum entropy, naive Bayesian and support vector machine. After the analysis and comparison of the experiments, an important conclusion is drawn: the effect of support vector machine is the best

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

in this kind of machine learning method, and the accuracy of classification reaches 80%, which caused a certain stir at that time. Whitelaw et al. [12] used the support vector machine model in their research. They acquired adjective words in the commentary data of film and television works, linked with the classical bag feature representation, and used vector space model to represent the text, to divide the positive and negative comment information. The accuracy increased to 90.2%. This is the case.

In China, Tang Huifeng et al. [13] used different parts of speech to represent different text features, and used document frequency, information gain, CHI statistics and mutual information to obtain different feature selection methods. They used comparative experiments, central vector method, K nearest neighbor, naive Bayesian and support vector machine models, etc. It is concluded that using n-Gram feature representation, information gain feature selection and support vector machine (SVM) classification method can achieve better emotional classification results when the training set is large enough and appropriate features are selected. Xia Huosong et al. [14] applied TF-IDF weighting algorithm to carry out some research. They applied support vector machine classifier about RBF kernel function to analyze customer evaluation of a website, and studied the influence of stop-use vocabulary on emotional classification.

In recent years, throughout the development of affective analysis, the key to affective analysis associated with machine learning methods is to obtain feature information efficiently. Its advantages are high accuracy, objectivity of knowledge acquisition, relatively long training cycle and high dependence on training corpus. Generally speaking, there is no obvious advantage between machine learning and affective dictionary. The development space of machine learning method is still very broad.

## 2. Preliminary

### 2.1. Text Preprocessing

In the field of natural language processing, Chinese processing needs more technology than English processing. Every sentence in English is made up of one word linked together. Each word is separated by a space, but the Chinese text is not like this. Our Chinese language is based on characters. These basic units can be linked together to form phrases. A sentence is made up of such many words or phrases. Most of the time, a word can not express a meaningful meaning very well. We can often identify a word composed of words as the smallest unit of meaning. This brings difficulties to our research. If there are problems, there are ways to solve them. This is why Chinese word segmentation technology comes into being.

Nowadays, the main methods used in word segmentation are statistical method, string matching method and understanding method. These three methods are the main word segmentation methods at present, each has its own advantages and disadvantages. The following is a detailed introduction:

Statistical Word Segmentation: General Chinese sentences are composed of many words, such as: I am a student. In this sentence, we can cut this sentence into: I, yes, one, student. Among them, "one" and "student" are two words, so we can count the number of adjacent words appearing at the same time, we can think that the more this number, the more likely it is to form a word. Therefore, the probability of co-occurrence of connected words can well reflect the credibility of words. But this method also has an obvious disadvantage, that is, it often extracts phrases with high co-occurrence frequency but not words, such as "yes", "mine", "this" and so on.

Word segmentation method based on string matching: This method is sometimes called mechanical word segmentation. This method is well understood, that is, to match the Chinese text waiting for analysis with the words in a larger machine dictionary one by one according to

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

the rules we have formulated. If the matching is successful, a word will be identified. This method can be divided into forward matching and reverse matching.

Understanding-based Word Segmentation: This method allows computers to simulate human understanding of sentences, just like human beings. That is to say, grammatical analysis is carried out in the process of word segmentation, and grammatical information and semantic information are used to deal with ambiguous phenomena. Normally, it includes three parts: word segmentation subsystem, syntax and semantics subsystem and the most important control part. The responsibility of the general control part is to manage the sub-system of word segmentation to get the grammatical and semantic information of Related words and sentences to distinguish the ambiguity of words, that is to say, it will imitate the process of understanding the meaning of sentences. This method has certain requirements for technology. It is not yet mature and is still in the experimental stage.

In this paper, Python is used to complete relevant experiments. The common word segmentation libraries based on Python are jieba, Snow NLP, THULAC, NLPIR, NLTK, LTP, etc. This paper mainly uses Jieba participle, which supports three participle modes: exact mode, full mode and search engine mode. Accurate pattern is an attempt to segment sentences accurately, usually for text analysis. The whole pattern is to search in detail in sentences and find all the probabilistic words. Its efficiency is very high. However, it also has a disadvantage that it can not solve the problem of unclear meaning. Search engine mode relies on the precise mode mentioned above. It further divides the longer words and improves the recall rate. It is generally used in search engine word segmentation.

Jieba segmentation can be used not only in simplified Chinese but also in traditional Chinese, and also in custom dictionary. The method of word segmentation is statistical.

## 2.2. Remove Stop Words

Stop words refer to the automatic deletion of certain words or words before or after processing text data in order to improve the efficiency of traversal when we traverse data. These words or words are stop words. These stop words need to be input by ourselves, and then generate a stop word list. Of course, we can also use the stop word list provided to us by others. For this topic, the main objective is emotional analysis, so some words in the comments, such as "buy", "commodity", "brand name" and so on, can be used as stop words in this experiment.

## 2.3. Text Sentiment Analysis Based on Deep Learning

The key point of cyclic neural networks is to use historical information to help solve current decision-making problems. But at the same time, it also brings new technical problems: long-term dependence.In order to solve this problem, the LSTM network came into being, which also makes the cyclic neural network widely used.

LSTM uses a few "gates" to allow messages to determine the state of each moment according to certain conditions. The "gate" mentioned here is a method of using sigmoid neural network and multiplying by position. The combination of these two modes is a "gate structure". The forgetting gate and the input gate are the most important parts of LSTM. They are the key points to preserve the long-term memory of the circulating neural network. The function of "forgetting gate" is to let the circulating neural network delete the information that is not important at present. When the circular neural network deletes some of the previous information, it also needs to add new information to the current input, which is accomplished by the "input gate". According to $X_t$, $C_{t-1}$ and ht-1, the input gate will analyze and determine which information will be added to the current state $C_t$. LSTM structure can get the current state after some data processing, and then calculate the output of the output at the current time. This work is completed by the "output gate". The output gate determines the output $h_t$ of this time according to the latest state $C_t$, the output $h_{t-1}$ of the previous time and the current input $X_t$.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

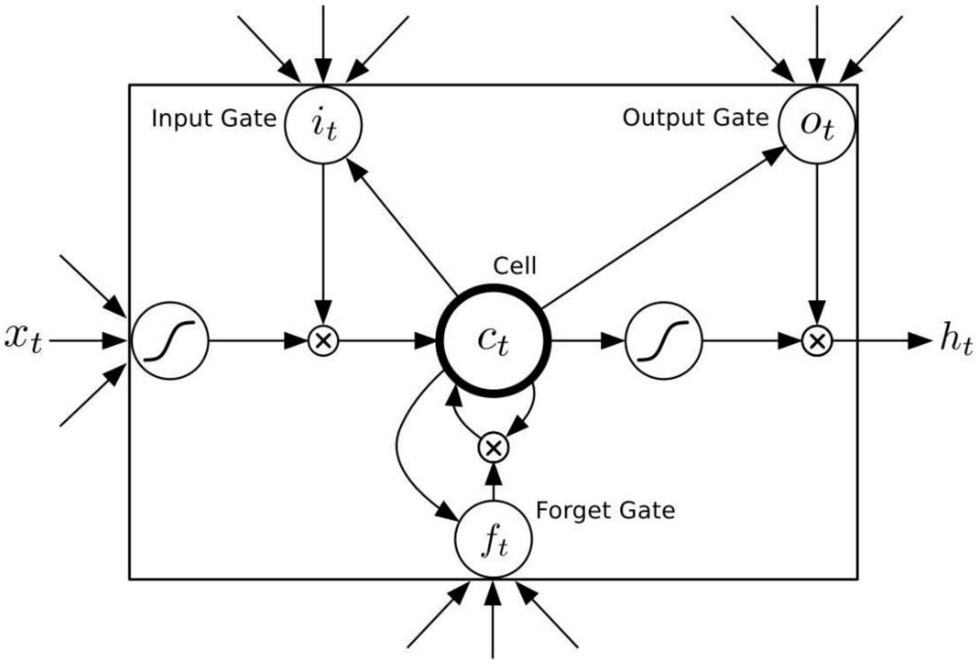


**Figure 1.** LSTM model structure

## 3. Method

### 3.1. Input Processing

Before the model starts, we need to process the input data. We represent the text data as word vectors. In this way, the input data can be represented as vectors, and then we can build a data model on the vectors. Considering the computational complexity and the relationship between words and expressions, we adopt the word vector representation method of Distributed Representation, which was published by Hinton in 1986. It is a low-dimensional real vector, which maps words to another address field, and then uses multi-dimensional continuous real vector to express its meaning. This low-dimensional spatial representation solves the problem that dimension is too high to calculate, and the problem of weak association attribute between words. So this method improves the accuracy of vector semantics to a certain extent.

### 3.2. Model

In order to achieve the purpose of emotional analysis of user reviews, this paper applies the network model of LSTM to achieve emotional analysis of text. The general steps of the model are as follows: firstly, the input comment data is encoded by word vector model, and then the text is transformed into word vector representation. Then, the LSTM network model structure is adopted. Finally, the emotional classification of text is completed through the full connection layer. The LSTM network can solve the problem of gradient disappearance or gradient explosion in RNN model. Its structure includes three special structures: input gate, output gate and forgetting gate. The input and output gates are responsible for the input and output of data, forgetting the selectivity of the gate and some of the previous data, in order to solve the problem of Gradient going to extremes.

## 4. Experiment

### 4.1. Experimental Setup

The experimental data in this paper come from user reviews of commodities on e-commerce platform. A crawler project is designed with Python language to capture the data, and some

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

tools of Python are used to clean and organize the data. The final result is saved as text, and the data format is shown in the figure.


generalcomments.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
5089273,包装比较一般 怎么说几千的东西8又是玻璃易碎的 居然盒子里没个气泡纸什么的。开机就70左右的电了打了两把王者又下载了一点软件之后也没怎么玩 到
5089273,不多给了，就给2颗星吧，第一次在京东写评论，好紧张啊。我这手机买的也是出人意料的惊心动魄。2018.1.2晚上为了赶元旦活动，急急忙忙下单，2018.
5089273,你们的也这样吗 吓人啊有点 充电不充电都这样,3,2
5089273,一次买了两个，白颜色的外包装松松垮垮感觉被拆过，外盒子包装也不一样，白色的系统也一会会不合适，官网上也没有查到序列号的服务期限，总感觉有
5089273,我现在觉得买手机不要在网上买了，说好的发票也不给劳资发，还说什么电子发票自己可以去打印，而且我买的这部手机对于强迫症来说那就是会爆炸的啊
5089273,本来买着为了让妈妈开心的，她舍不得买这么贵的东西，舍不得花钱，结果你们居然在过节的时候发一个有瑕疵的，侧面好多蹭了的痕迹，有的还能擦掉有
5089273,还没用顺手，但是感觉有时候屏幕总点不动，感觉特奇怪。苹果应该都很顺畅啊，反正我是第一次用苹果手机,3,2
5089273,机器拿到后认真看了一下屏幕和四周。没有划痕和掉漆。查了一下序列号，没有激活。激活使用后目前没有发现异常。但这机器明显是人家退货的。大家看
5089273,机子看着很不错，为什么我查了一下它的激活时间是在5月4号，难道我买到的是翻新机么？,3,2
5089273,第二次在京东上买苹果了，物流肯定是没得说。只是这次发过来的果8屏幕上有划痕，还是出去贴膜的时候店主撕开膜看见告诉我的，不过不认真看完全看
5089273,收到时，包装盒都坏了，幸好手机是完好的吧！也确实挺好看的，只是手机确实不咋滴啊！首先充电器不是原装的吧！反正我充不进去电，就找了客服问了
5089273,这个包装未免太简陋了吧，什么东西都没有垫，就直接一个纸盒一套就寄过来了手机盒都摔凹了 手机本身使用后还不错就是这个包装太让人不满意了&hel
5089273,之前用的苹果7 感觉手机太小打游戏不舒服打算换个5.5的 因为是学生没什么钱 看来看去打算在京东自营上买台二手7p 128g四千块钱 想着虽然是二手没
5089273,中评吧！才拿出来还蛮惊艳的，用了以后发现不太行，通话声音时大时小，才一个月手机就发烫，手机有时候会卡拉不动，都是小问题都不是很严重，但是
5089273,手机的左上角的位置有一个明显凹痕，本来激动万分的心情顿时飞落千丈?一直凉凉送给我自己。这算是京东店大欺人吗！这样的产品竟然还拿出来卖，真
5089273,手机不卡，但是相机感觉有问题，然后就是快递拿回来就是这个样子，不是很上心,3,2
5089273,手机其他都还好，也是未激活的，就是前置玻璃面板居然一边是刚合适的，一边有缝！！实在是做工太差，或者是被开启翻修过的吧？？！！,3,2
5089273,这手机跟期望的还是有差距、说实话我感觉这根本不像苹果的地新机，尤其是刚开始激活的时候显示内蒙古，我也是醉了，我根本在北京也没有选择内蒙古
5089273,帮别人买的，除了外包装太次了，听说手机外观糙点儿，接缝处有胶、背板左下角有进尘、底部插口上方还掉了一块,3,2
5089273,包装差，电池差，信号差，网络差，总结完毕。,2,2
5089273,京东速度快，用了一段时间来评价的，一般般，我觉得很多地方不如华为好用。,3,2
5089273,其他都挺好的，就是用着有反应有的慢，开机要反应两秒，有时候开其他应用还会卡住，感觉和我用了2年的6s差不多的感受～有点不舒服～虽然是64G的但
5089273,老婆的生日礼物，买前问她还有一个月就出i9了，要不要等等，说不等?，可买了不到半个月，降价便宜了1000多块，找谁说理啊！！！买时一定要银色，
5089273,整体来说还行！我觉得包装应该放点海绵进去好点，还有就是放卡的那个小洞插花了，应该是放卡进去的了,3,2
5089273,手机很好看，但是有点担心不是正品，用没多久发热就很严重；不知道怎么回事,3,2
5089273,这手机看起来好看，但是并没有那么好。没有人家说的说那么反应快。还有就是。机子表面看着是很好，但是你斜着看反光看 全是划伤。像被刀子划过一
5089273,耗电问题有点严重，对比之前也在京东买的一台ios10.3的6sp 现在的8p居然耗电远远快于6sp 客服也给不出一个所以然的答案！希望是ios11.41的问题
5089273,充电线应该不是原装的吧，充电极不稳定，之前用的插头怎么放都可以，这次这充电线只有一面可以冲，另一面插进去就接触不良。这才10来天的时间，但
5089273,手机总体来说还可以，但是外观有点小瑕疵，背后玻璃缝隙上有一边做得很不好，有缝隙。而且变框怎么会是粉色，不是金色，很奇怪，序列号是正品的。
5089273,买了五千多的手机，背面的左边明显缝隙，客服说一定会有公差，但是难免受不了，只能给个两星“良心”了反正你们有那么多的好评！正版是
5089273,买了不到一个月，使用感受：我是从6换的8p.嗯，反应速度倒是很快，唯一不满意是屏幕有点黄，和我的小6相比黄的不止一点点。按照客服说的调了一下
5089273,用了十天了，没啥问题，就是有点失望，真的是裸机，以前京东送膜送手机套子，这次啥都没有送，后来发现****要便宜好几百，顿时无语，希望京东能改

The format of the content of the document is: commodity ID, comment content, number of stars (1, 2, 3, 4, 5) and comment level (1, 2, 3). The comment level 1 represents bad reviews, 2 represents medium reviews and 3 represents good reviews. The experimental data are collected online.

### 4.2. Experiment Result

Accuracy is used to evaluate the accuracy of the model, which means the proportion of the correctly predicted samples to the total sample. However, it has a disadvantage. If the training set of the sample is different from the data of the three classified samples of the test set, the result may be affected.

We use 200,000 commentary data as training set and test set respectively. The ratio of training set and test set is 1:1, and the final accuracy is about 75%.

## 5. Conclusion

The application of cyclic neural network in emotional classification is studied and experimented. The text emotional classification model of cyclic neural network is established. The model is trained with the processed training data set and tested after training. Experiments show that the model has a good accuracy of text emotion classification.

## References


[1] Zhai Donghai, Hou Guilin, Liu Yue. Parallel Algorithms for Text Sentiment Analysis Based on Deep Learning [J/OL]. Journal of Southwest Jiaotong University:1-9[2019-04-19].

[2] Zhou Minyi, Guo Yan, Ding Youdong. Sentiment analysis of Chinese movie reviews based on deep learning [J]. Journal of Shanghai University(Natural Science Edition),2018,24(05):703-712.

[3] Ji Chaojie. Sentiment Analysis Based on Deep Learning and Rhetorical Relation [D]. Nanchang University,2018.

[4] Hatzivassiglouv, McKeownK R. Predicting the semantic orientation of adjectives. in:Proceedings of the 35th annual meeting of the European Chapter of the ACL.Morristown, NJ,USA: ACL, 1997. 174-181.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202004_3(4).0024

[5] Qingliang Miao ,Qiudan Li, Ruwei Dai. AMAZING:A sentiment mining and retrieval system. Expert Systems with Applications: An International Journal,2009,36(3): 7192-7198.

[6] Ramanathan Narayanan, Bing Liu, Alok Choudhary. Sentiment Analysis of Conditional Sentences. in: Proceedings of the 2009 Conference on EMNLP.Morristown, USA: ACL, 2009. 180-189.

[7] Xu Linhong, Lin Hongfei. Discourse Affective Computing Based on Semantic Features and Ontology. JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT, 2007,44(3):356-360.

[8] WEN Bin, HE Ting-ting, LUO Le, SONG Le, WANG Qian. Text Sentiment Classification Research Based on Semantic Comprehension. COMPUTER SCIENCE,2010,37(6):261-264.

[9] Wang Suge, Li Deyu, Wei Yingjie. A Method of Text Sentiment Classification Based on Weighted Rough Membership. JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT,2011,48(5):855-861.

[10] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia,USA: 2002. 79-86.

[11] Bo Pang, Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. in: ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown,USA:2005.115-124.

[12] Casey Whitelaw, Navendu Garg, Shlomo Argamon. Using appraisal groups forsentiment analysis. in: Proceedings of the 14thACM international conference on Information and knowledge management. New York,USA: 2005. 625-631.

[13] TANG Hui-feng TAN Song-bo CHENG Xue-qi. Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques. JOURNAL OF CHINESE INFORMATION PROCESSING,2007,21(6):88-94.

[14] Xia Huosong, Tao Min, Wang Yi, Wei Xiang. The Influence of Stop Word Removal on the Chinese Text Sentiment Classification Based on SVM Technology. JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC ANDTECHNICAL INFORMATION,2011,30(4):347-352.