Classification Algorithmof Chinese SentimentOrientation Based on Dictionary and LSTM

Ge Bin
Science and Technology on InformationSystems Engineering Key Laboratory , National University of Defense Technology,Changsha, Hunan, P.R. China, 410073
gebin@nudt.edu.cn

He Chunhui*
Science and Technology on InformationSystems Engineering Key Laboratory , National University of Defense Technology,Changsha, Hunan, P.R. China, 410073
xtuhch@163.com

Zhang Chong
Science and Technology on InformationSystems Engineering Key Laboratory , National University of Defense Technology,Changsha, Hunan, P.R. China, 410073
leocheung8286@qq.com

Hu Yanli
Science and Technology on InformationSystems Engineering Key Laboratory , National University of Defense Technology,Changsha, Hunan, P.R. China, 410073
huyanli@nudt.edu.cn

ABSTRACT

Chinese sentiment analysis is a hot research issue in information analysis, but the tagging corpus which can be used for machine learning algorithm training is poor. Machine learning algorithm is used for text sentiment classification, generally only categories are given while sentiment words can not be extracted. This paper proposed an automatic tagging strategy for training corpus and a classification algorithm for Chinese sentiment orientation based on dictionary and LSTM. It can label the training corpus automatically and accurately and efficiently, and also extract sentiment words. Experiment shows this method is effective and the accuracy of LSTM algorithm has reached 93.51% on the mixed data set of sentiment classification.

CCS Concepts

Applied computin →Document management and text processing → Document capture → Document analysis.

Keywords

Sentiment Analysis; Automatic Annotation; Long Short-Term Memory Neural Network

1. INTRODUCTION

With the rapid development of the Internet and social media and e-commerce platforms, a large number of users have generated a large amount of text data with sentimental tendencies in a short period of time. In recent years, using these text data to mine hidden negative or positive sentiment tendencies has become a very valuable research direction in the field of natural language processing, and a lot of research results have been obtained. Through the induction and analysis of relevant literature, it is found that the current mainstream text sentiment analysis methods mainly include sentiment analysis based on the sentiment database and template rule base, or statistically based methods using artificially labeled corpus to train machine learning algorithms. Then the trained algorithm or model is used to classify the sentiment tendencies of the text. In the process of sentiment analysis techniques and theoretical development, these two methods often infiltrate each other, prompting the sentiment analysis technology to continue to advance. Especially in the sentiment analysis of English, the researchers have put forward many efficient algorithms and mature tools. However, for Chinese sentiment analysis, the start is relatively late, and Chinese is still facing problems and challenges such as the lack of large-scale annotated data sets.

With the deep maturity of deep learning techniques and frameworks, some researchers have proposed to use deep neural network algorithms to deeply mine the sentiment tendencies in text.Although this method can greatly improve the performance of the algorithm under certain premise, it also has some shortcomings. The premise is that a large amount of labeled training data is needed as the input of the algorithm.Considering the fact that the high-quality labeling training corpus available in Chinese is particularly lacking, this is a major challenge for Chinese sentiment analysis;Second, such methods can only classify text sentiment tendencies.They don't give sentiment words that appear in the document, which is not friendly for many fine-grained Chinese sentiment analysis tasks.At the same time, they are often impossible to explain the classification results of sentiment orientation.

In order to better solve the above deficiencies, this paper proposes a classification algorithm of Chinese sentiment orientation based on dictionary and long-term and short-term memory neural network(LSTM).This algorithm combined with the sentiment dictionary and sentiment score calculation method can give the

ICBDR 2018, October 27–29, 2018, Weihai, China

978-1-4503-6476-8/18/10…$15.00
DOI: http://dx.doi.org/10.1145/3291801.3291835

sentiment words appearing in Chinese documents well, and it can also automatically construct high-quality training label corpus, and finally, using some manually labeled corpus and combining these corpus automatically annotated with the algorithm to form the training corpus of the deep learning algorithm.Thereby solving the problem of lack of training corpus and sentiment word extraction, and discriminating the sentiment tendency of documents.At the same time, the performance of the algorithm is verified by the combination of manual labeling and automatic labeling. The Accuracy index is used to verify the effectiveness of the method.

The structure of this paper is designed as follows: The second part briefly describes the research work related to this paper in the field of sentiment analysis; the third part gives a detailed overview of the proposed sentiment orientation classification algorithm; the fourth part will verify the performance of the algorithm through comparative experiments; the fifth part gives the summary and outlook.

2. Related research work

The analysis of text sentiment orientation is mainly based on the content of the text to determine the sentiment tendency of the text. This technique has applications in many fields. The existing research results are mainly concentrated in the following aspects:

(1)Sentiment analysis based on dictionaries and rules.

(2)Sentiment analysis using statistical machine learning.

(3)Combine dictionaries, rules, and statistical machine learning methods to perform sentiment analysis together.

(4)Using a deep learning algorithm for sentiment analysis.

The dictionary-based and rule-based approach is mainly to collect and construct sentiment dictionary, combined with some specific rules to achieve text sentiment analysis.

Rao D [1] , Hatzivassiloglou V [2] and Wiebe [3] respectively explored a variety of sentiment word extraction methods; Turney [4] proposed a word classification algorithm based on point mutual information. The commonly used dictionary refers to HowNet and WordNet. In Chinese, Zhu Xi et al [5] based on HowNet combined with semantic similarity to determine the sentiment tendency of words; Zhao Q [6] uses sentiment vocabulary combined with weighted combination to classify sentences in sentiment polarity; The Chinese analysis system NLPIR developed by Kevin Zhang [7] has realized the calculation of text sentiment scores and the extraction of positive and negative sentiment words in the sentiment analysis module by constructing a weighted sentiment dictionary combined with machine learning algorithms.

The statistical machine learning method mainly uses the artificially labeled training data set to extract the features of the text through supervised learning to construct the sentiment classifier, and then uses the classifier to classify the analyzed text for sentiment orientation. Pang [8] combines the word bag and the N-gram model to distinguish the film reviews into positive and negative categories; Szalay [9] constructs an sentiment orientation classifier by selecting some important features and combining a variety of different machine learning algorithms.

Li T [10] and Melville P [11] proposed that combining the sentiment dictionary and some annotated corpus to train the classifier can effectively compensate for the deficiencies of the two alone; He Y [12] divides sentiment analysis into two steps: firstly use the sentiment dictionary to make initial judgment on the text sentiment tendency, then use this result to modify the new classifier, and finally use the modified classifier to classify the sentiment tendency of the text; Hot West Danmu Tuolhong Tai [13] used the idea of combining dictionary and machine learning methods to analyze the sentiment orientation of the text.

Deep learning algorithm is used to solve the problem of sentiment analysis. Since there is no need for manual feature selection, a large number of feature extraction can be reduced. Socher R [14] uses the recursive self-encoder (RAE) tree regression model and Chen T [15] uses BiLSTM-CRF and CNN to classify the sentimental tendencies of sentences in the text; Shin B [16] and Ouyang X [17] used CNN and Attention combined with different strategies to classify the sentimental tendencies of the text content.

Liu B [18] and Pang B [19] respectively proposed effective solutions in text view mining and sentiment analysis. By sorting out and summarizing the above research results, they got relevant inspiration: It is considered that it is feasible to classify text sentiment tendency by combining dictionary and deep learning algorithm. Therefore, the classification algorithm of Chinese sentiment orientation based on dictionary and LSTM is proposed.

3.Overview algorithm principle

The algorithm uses the separately constructed sentiment dictionary combined with the sentiment score calculation method of the sentiment word, which can extract the sentiment words appearing in the text well, and construct the annotation corpus for model training based on the sentiment score and the automatic labeling algorithm. Finally, by combining a small number of manually labeled corpora and a large-scale annotated corpus automatically constructed by the algorithm, the training and test corpus of the deep learning algorithm are formed together. Therefore, the problem of large-scale annotated corpus deficiency and sentiment word extraction and classification of document sentiment tendencies is better solved. The overall framework of the algorithm is shown in Figure 1.

As can be seen from Fig. 1, the algorithm is mainly divided into three stages: model training and sentiment dictionary construction and model testing: The model training parts mainly consists of several core steps: loading training data and sentiment dictionary, sentiment score calculation, automatic labeling training corpus, fusion manual corpus, deep learning model training and preservation; The sentiment dictionary construction stage is to complete the construction task of the new sentiment dictionary by collating and collecting the basic sentiment dictionary; The model testing parts mainly includes the steps of loading the classification model and the sentiment dictionary, the sentiment word extraction step, and the sentimentorientation classification step. The above steps are detailed in Sections 3.1 and 3.2.

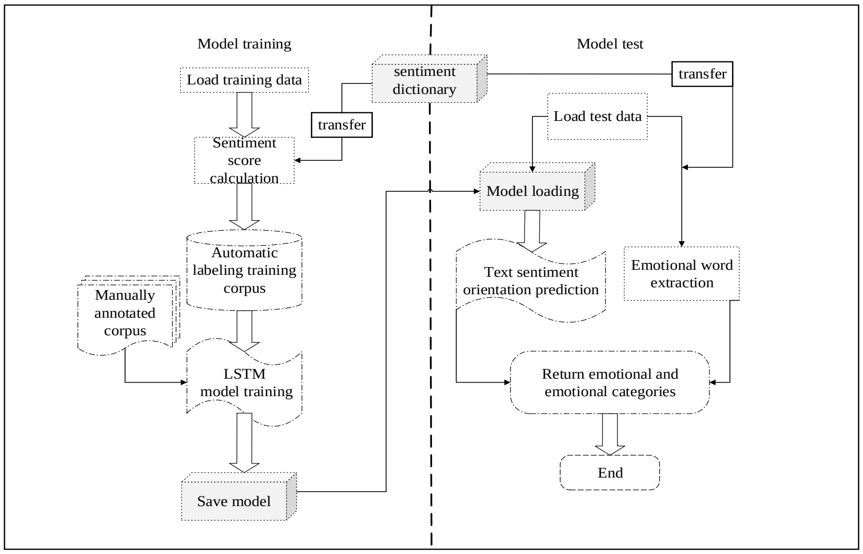

Figure 1. Structure diagram of Chinese sentiment orientation classification algorithm based on dictionary and LSTM

3.1 Model training

The model training part generally consists of four core modules, which are data preprocessing module, sentiment score calculation module, training data set automatic labeling module, model training and saving module. Details of each module are described in section 3.1.1 and 3.1.5.

3.1.1 Data collection and preprocessing

There are two cores type of data sets used in the experiment:

The first category: the data sets that researchers have collated and published for Chinese sentiment classification research, mainly including Chinese shopping comment data, hotel comment data, and SemEval-2016 Chinese evaluation data set. These data are labeled with positive and negative text data. After filtering, formatting and deduplication, the data set DataSetA with positive and negative categories is obtained.

The second category: text data collected by Internet crawlers. For the 600,000 original texts collected by the reptile, after pre-processing steps such as clause, word segmentation, stop, de-weight, de-drying and sentiment score calculation, Then, according to the set sentiment score threshold, automatically screening the documents that meet the requirements to construct a data set DataSetB with category labels;

Automatic labeling of text data collected by reptiles is an innovative point of this paper. The specific implementation process is described in section 3.1.4. Finally, by deduplication and merging the two types of data sets DataSetA and DataSetB, a hybrid approximation data set DataSetC is obtained to verify the performance of the algorithm. The basic situation of the sample size contained in the above data set is shown in Table 1.

Table 1. Dataset basic situation table

Dataset name	Number of positive documents	Number of negative documents
DataSetA	8712	8053
DataSetB	20245	19423
DataSetC	26573	26608

It can be seen from Table 1 that the manually constructed data set DataSetA is a small-scale data set, and DataSetB, which is completely annotated by the algorithm, and the fused hybrid dataset DataSetC are a large-scale mixed annotated dataset.

3.1.2Construction of sentiment dictionary

In order to extract sentiment words, a new sentiment dictionary is constructed by tidying up the basic sentiment words.

The basic sentiment words are mainly from five public sentiment dictionaries, namely Chinese HowNet, Dalian University of Technology's sentiment vocabulary ontology library, Boson's sentiment dictionary, Harbin Institute of Technology open synonym forest, Tsinghua University Li Jun's dictionary of derogatory meaning [20].

Based on the above dictionary, after deduplication, weight check and manual screening steps, a dictionary with a weight of about 33,310 words is constructed. The words weight are 1 or -1; the positive sentiment weight is 1, and the negative weight equals -1.

3.1.3Sentiment score calculation

The sentiment score calculation module mainly calculates the positive and negative sentiment scores of the document and is divided into two stages [21]:

The first stage is to process the documents in a sentence and calculate the sentiment score of each sentence. For the specific calculation, see formula (1):

$$\mathrm{Orientation}(W_{S_k}) = \begin{cases} 0,\text{Sentence } S_k \text{ does not contain emotional words} \\ \sum_{i=1}^{n}(P_i * Q_i),\text{other} \end{cases} \quad (1)$$

Where $\mathrm{Orientation}(W_{S_k})$ represents the sentiment score of the k-th sentence in the i-th document, S_k refers to the k-th sentence, n refers to the total number of all words in the k-th sentence, and P_i and Q_i represent the word i negative prefix coefficient and sentiment word weight, The negative prefix coefficient P_i is defined in formula (2):

$$P_i = \begin{cases} 1, & \text{The word i is an sentiment word} \\ & \text{and the prefix has no negative words.} \\ 0, & \text{Word i is not an sentiment word.} \\ -1.5, & \text{The word i is an sentiment word} \\ & \text{and the prefix has a negative word.} \end{cases} \quad (2)$$

Studies [22] have shown that the effect of a prefixed sentiment lexicon on a sentence's comprehensive sentiment score is larger than a single sentiment word without a prefix, so the weight is set -1.5. The definition of sentiment word weight Q_i is shown in formula (3).

$$Q_i = \begin{cases} 1, & \text{Word i belongs to positive sentiment word.} \\ 0, & \text{Word i is not an sentiment word.} \\ -1, & \text{Word i belongs to negative sentiment word.} \end{cases} \quad (3)$$

Using formula(1)-(3), it is convenient to calculate the sentiment scores of all sentences in a document. On the basis of some commonly used prefix negative words compiled by Li Aiping [22], some negative words are added to construct a new prefix negative vocabulary. The final negated words with negative meaning are shown in Table 2.

Table 2. Negated words with negative meaning

Frequently used negative words
不、不想、没、没有、没能、非、不是、不要、不够、不可、不便、不宜、不许、不能、不曾、不准、无、别、莫、勿、毋、未、未曾、未必、未能、难以、从不、从未、从没、无法、绝非、绝不

The second stage is to calculate the sentiment scores of all the sentences in a document. For the specific calculation method, see formula (4):

$$\text{Orientation}(D_i) = \sum_{k=1}^{N} \text{Orientation}(W_{S_k}) \quad (4)$$

Where D_i refers to the i-th document, $\text{Orientation}(D_i)$ refers to the sentiment score of the i-th document, N represents the total number of sentences in the i-th document, and k represents the k-th sentence in the i-th document.

3.1.4 Method of automatically labeling corpus

The sentiment score for each document can be obtained according to the formula (4) in Section 3.1.3.Considering that the calculation method of sentiment scores is relatively simple, in order to verify the quality of the automatically constructed training corpus, 1200 papers were randomly selected from the reptiles for sentiment score experiment and comparative analysis, and the experimental results shown in Fig. 2 and Fig. 3.

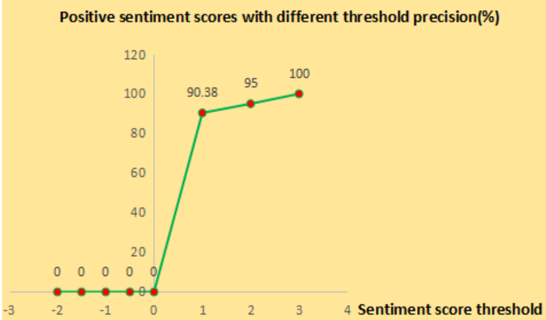

Figure 2. Positive sentiment score threshold precision result

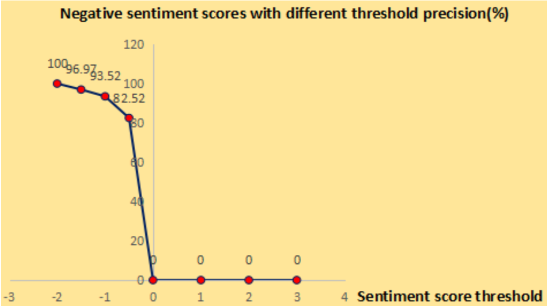

Figure 3. Negative sentiment score threshold precision result

According to the results of Fig. 2 and Fig. 3, when the positive sentiment score threshold is higher than 2 points, the positive classification can achieve more than 95%, and the negative sentiment score threshold is lower than -1.5 points, the negative classification can achieve more than 95%. Therefore, in the data set used in this paper, the document for positive classification in the automatic labeling process is threshold by 2, and the document with the sentiment score of not less than 2scores is automatically labeled as positive; For negatively classified documents with the -1.5 threshold, documents with an sentiment score of no more than -1.5 are automatically labeled as negative. All documents with an sentiment score between -1.5 and 2 are labeled as unknow, and finally those documents labeled as unknow are filtered out. This automatic labeling algorithm is one of the innovative point in this paper. The specific implementation steps of the automatic labeling algorithm are given below.

Input: a Chinese document

Output: the document and its corresponding category

Step (1): Input a document

Step (2): Calculate the sentiment score of the document using formula (1) - formula (4)

Step (3): Classification labeling based on sentiment score

```
If (Sentiment Score>=2):
    category=positive;
else if (SentimentScore <=-1.5):
    category=negative;
else:
    category=unknow
```

Step (4):Return the document and the category

Finally, using the automatic labeling algorithm given above, the 600,000 documents collected by the reptile were preprocessed, and the sentiment scores were calculated. After setting the sentiment score threshold and filtering the specified categories, approximately 39,668 documents that meet the requirements and with positive and negative categories are automatically selected to form the data set DataSetB of the automatic label category.

3.1.5 LSTM model training and preservation

The experiment is based on Google's open source CPU version of the deep learning framework TensorFlow as the back end. The upper layer is matched with the advanced API of Keras package to realize the network construction of the standard single hidden layer long-term and short-term memory neural network LSTM algorithm, and finally the trained model is used as the classifier of Chinese sentiment orientation. During the experiment, the data set adopt80% training and 20% test. In the course of the experiment, the jieba packagein python is used for Chinese word segmentation and the customized stop word table is used to filter the stop words. To prevent over-fitting of the model, the strategy of using Dropout as 0.5 was used in the experiment. In the training and testing process of the LSTM algorithm, some of thehyperparameters used in the experiment are shown in Table 3 after many adjustments and optimization.

Table 3. The value of partial hyperparameters in LSTM algorithm

name	value	name	value
Number_cell	128	Dropout	0.5
Dense	1	Activation	sigmoid
Loss	binary_crossentropy	optimizer	adam
Batch_size	64	number_epoch	10
Sentence_Maxlen	60	class_mode	binary

After the algorithm training in Table 3 is completed, the value of the Accuracy indicator is given, and the trained model is persisted to the hard disk to facilitate subsequent calls.

3.2 Model test

The main task of the model test is to realize the classification of sentiment categories and the extraction of sentiment words. Therefore, it can be divided into two parts: document sentiment tendency classification and sentiment word extraction. For details, please refer to Section 3.2.1 and 3.2.2.

3.2.1 sentiment word extraction

Extracting positive and negative sentiment words from the document is a very meaningful work, but at present many tasks based on machine learning algorithms can only give the categories of test samples, and can not accurately give positive and negative sentiment words. In order to solve this problem, this paper uses the constructed positive and negative sentiment dictionary to extract the positive and negative sentiment words of the document, and extracts the sentiment words according to the following rules:

(1) If the current word belong to the sentiment dictionary and the weight is 1, it is marked as a positive word and a positive sentiment word list is added;

(2) If the weight is -1, it is marked as a negative word, and a negative sentiment word list is added;

(3) If the word is not in the sentiment dictionary, skip it and go to the next word until all words have been drawn.

3.2.2 Document sentiment classification

According to Fig. 1, the classification of document sentiment tendency is divided into two steps. Firstly, the trained LSTM model is used to classify the positive and negative sentiment categories of the test text; and then the correction result is combined with the extraction result of the sentiment words. The specific way is as follows: First, the classification result of the LSTM algorithm is assigned to the sentiment category DLable of the document. At this time, the DLable has only two categories: positive or negative; Then determine whether the element in the positive and negative sentiment word extraction list of the document is empty. If it is not empty, directly return the value in the DLable value and all the sentiment word lists, for special cases: If all the elements in the list of positive and negative sentiment words are empty, it means that the document does not contain sentiment words, and therefore the original document does not have positive and negative sentiment colors. At this time, the current value of DLable is corrected to neutral and DLable is returned. In this way, the final text will be positive or negative or neutral, so the algorithm can accurately give the text sentiment tendency category.

4. Experimental

In order to verify the performance of the algorithm and the quality of the automatically annotated data, experiments and comparative analysis were carried out from different dimensions using support vector machine SVM algorithm and long- and short-term memory neural network LSTM algorithm and IRNLP tool on different data sets. How many dimensions in a single document feature vector to perform the classification task will be the best? There is generally no standard answer to this question, because the feature dimension is generally closely related to the corpus used. In order to optimize the performance of the algorithm in the experiment, the SVM algorithm is used to compare the feature dimensions on the DataSetA with different feature dimensions and whether to use the stop words. The test results are shown in Fig. 4 and Fig. 5.

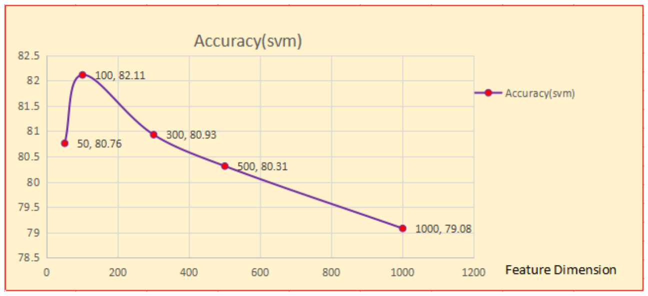

Figure 4. different features dimension& no stop words result

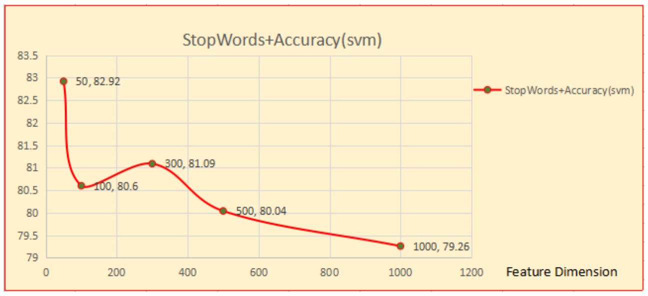

Figure 5. different featuresdimension &add stop words result

According to Fig. 4 and Fig. 5, the performance of the algorithm is verified from the selection of different feature numbers and whether to filter the two dimensions of the stop words. Figure 4 does not use stop word filtering, select five different dimensions of 50, 100, 300, 500, 1000 as the feature dimension of the document, and use SVM algorithm to test on DataSetA. When the dimension is found to be 100, the algorithm's Accuracy is the highest, reaching 82.11%; Figure 5 is combined with the stop word table to filter the stop words and then tested. The rest of the settings are exactly the same as the experiment in Figure 4, but the experimental results are different from those in Figure 4. When the stop word is added, the algorithm achieves 82.92% in the lower dimension of the 50-dimensional Accuracy. The above experimental results show that by selecting the appropriate stop words in the text classification task, not only can the dimension of the feature space be effectively reduced, but also the classification performance of the algorithm can be improved to some extent.

According to the above experimental results, it is found that the SVM algorithm has the highest classification performance when combined with the stop words and feature dimension equals 50. Therefore, subsequent comparison experiments on different data sets are set in this way so that comparisons can be made more objectively. Experiments and analysis were performed on the different datasets using the SVM algorithm , LSTM algorithm and the IRNLP[7] sentiment analysis tool. The experimental data set partition situation and different algorithm experimental accuracy are shown in Fig. 6 and Fig. 7.

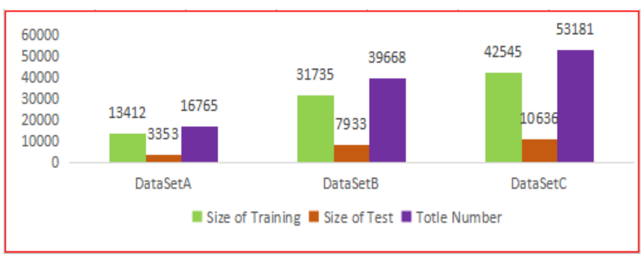

Figure 6. experimental data sets partition situation

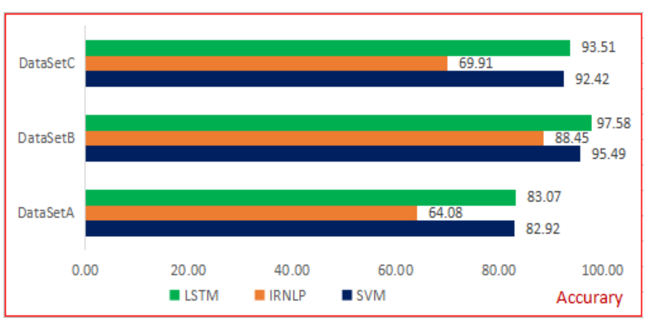

Figure 7.SVM, IRNLPtool and LSTM experimental accuracy

According to the data sets partition situation and experimental results withFig. 6 and Fig.7, in the manually labeled DataSetA, the Accuracy of the SVM reached 82.92%, and the Accuracy of the LSTM reached 83.07%, which may be due to the under-fitting of the algorithm. The IRNLP tool is used to calculate the sentiment score of the 20% sample of DataSetA, and then the Accuracy obtained by dividing the category into positive and negative according to the score greater than 0 and less than 0 is 64.08%. Similarly in the data set DataSetB automatically labeled by the algorithm, each algorithm performs well, the Accuracy of SVM reaches 95.49%, the Accuracy obtained by IRNLP tool is 88.45%, and the Accuracy of LSTM is 97.58%. The results show that the quality of the automatic labeling data is higher. Finally, in the mixed dataset DataSetC, SVM and LSTM generally performed better. The Accuracy of SVM was 92.42%, and the Accuracy of LSTM was 93.51%. However, the sentiment classification effect of IRNLP tools was relatively general, and Accuracy was only 69.91%. From the comparison, LSTM accuracy is better than SVM and IRNLP tools in each data set used, which fully verifies the performance of the algorithm. The general reason for the IRNLP effect may be that the selected test corpus contains a large amount of short text to cause such a result.

The following is the LSTM algorithm trained using the largest mixed data set DataSetC in this experiment, combined with the constructed sentiment dictionary, a small amount of comment data and BBS text obtained from major e-commerce platforms and post bars on the Internet can be used as test data to test the classification accuracy and generalization ability of the algorithm. The sentiment word extraction and classification results of some samples are given below. The specific result are shown in Table 4.

Table 4. Classification and sentiment word extraction result

Text content to be tested	sentiment word extraction	judgement result	Real category
The location of the hotel is remote, and the price is not substantial.	remote /substantial	Negative	Negative
This mobile phone is beautiful and suitable for birthday gift.	beautiful/ suitable	Positive	Positive
I can't get through this day. I haven't had enough time to eat.	can't	Negative	Negative
This lipstick is of good quality and worth buying.	good quality/ worth	Positive	Positive
After dinner, then go shopping.		Neutral	Neutral
Genuine and fast logistics!	genuine /fast	Positive	Positive
The shoes are a little distorted. I wonder how long they can wear.	distorted	Negative	Negative
Experts point out that artificial intelligence will drive economic development in the future.	Drive/de-velopment	Positive	Positive

Hotel environment is good, affordable, well received, next time to continue to stay in.	good / well received	Positive	Positive
You are so boring!	boring	Negative	Negative

According to the experimental results in Table 4, the LSTM model trained by the large-scale mixed data set DataSetC can highly accurately classify these "unknown" new data, and can fully verify the classification performance and generalization ability of the algorithm.

5. Conclusion

Combining the dictionary and machine learning algorithms to analyze the sentiment of the sentiment, it can extract the sentiment words and ensure the quality of the sentiment classification, which provides a better solution for the sentiment analysis task. Based on the comprehensive analysis of the experimental results, the following conclusions are drawn:

(1) LSTM algorithm is better than SVM algorithm in document sentiment classification task;(2) In the text classification, there is a certain relationship between the best feature dimension when the document is vectorized and the corpus and stop words used;(3) According to the experimental results of the algorithm on the fully automatic labeling DataSetB and the mixed DataSetC, it can be known that the classification of fully automatic label data sets is better, which verifies the feasibility of the automatic labeling strategy, and also provides a way to solve the current situation of the lack of data in Chinese sentiment analysis.(4) sentiment classification combined with dictionary and LSTM is better than using dictionary or statistical learning algorithm alone.

Inadequacies: The method of calculating sentiment scores based on sentiment lexicon is relatively simple, and does not consider the influence of factors such as degree adverbs; In the experiment, only SVM and LSTM and the IRNLP tool were used for comparison, and other relevant algorithms were not used for further verification.

The current LSTM algorithm analyzes the sentiment tendency of the chapter level, and attempts to add attention mechanism in the future to further optimize the classification performance of the algorithm.

6. ACKNOWLEDGMENTS

This research is supported by the NNSF of China under Grant No. 61302144 and Hunan NSF under Grant No.2018JJ2475.

7. REFERENCES

[1] Rao D, Ravichandran D. 2009. Semi-Supervised Polarity Lexicon Induction. In *Proceedings of the European Chapter of the Association for Computational Linguistics* (Athens, Greece, March 30 – April 03, 2009), 675-682.

[2] Hatzivassiloglou V, Mckeown KR. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics* (1997), 174-181.

[3] Wiebe. Learning Subjective Adjectives fromCorpora. 2000.In *Proceedings of the 17th National Conference on Artificial Intelligence and 20th Conference on Innovative Applications of Artificial Intelligence*. AAAI Press, 735-740.

[4] Turney P D, Littman M L. 2003. Measuring praise and criticism: Inference of Semantic Orientation from Association. *ACM Trans on Information Systems.* 21, 4, 315-346. DIO= http://doi.acm.org/10.1145/944012.944013.

[5] Zhu Wei, Yan Jin, Zhou Yaqian. 2006. Semantic Orientation Computing Based on HowNet. *Chinese Journal of Information Science*. 20, 1 (2006), 14-20.

[6] Zhao Q, Sun J, Yu C, et al. 2009. A Paralleled Large-Scale Astronomical Cross-Matching Function. In *Proceedings ofAlgorithms and Architectures for Parallel Processing International Conference* (Taipei, Taiwan, June 08-11, 2009). ICA3PP 2009, Proceedings. DBLP, 604-614.

[7] Zhang Huaping, Gao Kai, Huang Heyan, etc. 2014. *Big Data Search and Mining*. Science Press, Beijing, China.

[8] Pang T B, Pang B, Lee L. 2002. Thumbs up?: Sentiment Classification Using Machine Learning. In *Proceedings of Emnlp* (2002), 79-86.

[9] Szalay A S, Brunner R J. 1998. Astronomical Archives of the Future: A Virtual Observatory. *Future Generation Computer Systems*. 16, 1, 63-72.

[10] Li T, Zhang Y, Sindhwani V. 2009. A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In *Proceedings of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP: Association for Computational Linguistics*(2009), 244-252.

[11] Melville P, Gryc W, Lawrence R D. 2009. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France), 1275-1284. DIO= http://doi.acm.org/10.1145/1557019.1557156.

[12] He Y, Zhou D.2011. Self-training from Labeled Features for Sentiment Analysis. *Information Processing & Management*. 47, 4, 606-616.

[13] Rexidanmu Tuerhongtai,Wushour Silamu,Yierxiati Tuergong. 2017. Uyghur Text Sentiment Analysis by Combining Lexical Knowledge with Machine Learning Methods. *Chinese Journal of Information*. 31, 1, 177-183.

[14] Socher R, Pennington J, Huang E H, et al. 2011. Semi-supervised Recursive AutoEncoders for Predicting Sentiment Distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), 151-161.

[15] Chen T, Xu R, He Y, et al. 2016. Improving Sentiment Analysis Via Sentence Type Classification using BiLSTM-CRF and CNN. *Expert Systems with Applications.* 72, 10 (2016), 221-230.

[16] Shin B, Lee T, Choi J D. 2016. Lexicon Integrated CNN Models with Attention for Sentiment Analysis. In *Proceedings of the EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2016), 1-10.

[17] Ouyang X, Zhou P, Li C H, et al. 2015. Sentiment Analysis Using Convolutional Neural Network. In *Proceedings of the IEEE International Conference on Computer and Information Technology*(2015), 2359-2364.

[18] Liu, B. 2016. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies.* 30, 1(2016), 152-153.

[19] Pang B, Lee L. 2018. Opinion Mining and Sentiment Analysis. *Foundations & Trends in Information Retrieval*. 2, 1(2008), 459-526.

[20] Li J, Sun M. 2007. Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques. In *Proceedings of the2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering*(2007), 393-400.

[21] Nian Mei, Liu Ruolan, Marhaba Esiti, et al. 2016. Analysis of the Sentence Tendency in Uighur Language. *Journal of Computer Systems*. 25, 7(2016), 171-175.

[22] LI Aiping, ZHAI Peng, DUAN Liguo. 2015. Document Sentiment Orientation Analysis Based on Sentence Weighted Algorithm. *Journal of Chinese Computer Systems*. 36, 10(2015), 2252-2256.