A METHOD OF FINE-GRAINED SHORT TEXT SENTIMENT ANALYSIS BASED ON MACHINE LEARNING

G. Chang[*], *H. Huo*[†]

Abstract: Text sentiment analysis plays an important role in social network information mining. It is also the theoretical foundation and basis of personalized recommendation, circle of interest classification and public opinion analysis. In view of the existing algorithms for feature extraction and weight calculation, we find that they fail to fully take into account the influence of sentiment words. Therefore, this paper proposed a fine-grained short text sentiment analysis method based on machine learning. To improve the calculation method of feature selection and weighting and proposed a more suitable sentiment analysis algorithm for features extraction named N-CHI and weight calculation named W-TF-IDF, increasing the proportion and weight of sentiment words in the feature words Through experimental analysis and comparison, the classification accuracy of this method is obviously improved compared with other methods.

Key words: *fine-grained sentiment analysis, sentiment information classification, feature extraction, weight calculation*

Received: November 28, 2017 **DOI:** 10.14311/NNW.2018.28.019
Revised and accepted: June 20, 2018

1. Introduction

With the rapid development of the Internet, the function of the network is more and more comprehensive, and the use is more and more convenient. The rapid development of mobile Internet and the massive growth of mobile phone users, making the various functions of APP are constantly changing in a rapidly changing network environment. The development and replacement of the network security issues have also received more and more people's attention [1, 2]. Under the premise of ensuring network security,, social platforms such as Twitter, Facebook, micro-blog and WeChat, etc., have also rapidly emerged in the rapidly developing cyberspace and gradually developed from a single web-based terminal to a dual-platform based on mobile terminals.

[*]Guoqin Chang; Laboratory of Intelligent Computing and Application Technology for Big Data, Henan University of Science and Technology, Luoyang, Henan, 471023, China, E-mail: 15829715097@sina.cn

[†]Hua Huo – Corresponding author; School of Software, Henan University of Science and Technology, Luoyang, Henan, 471023, China, E-mail: pacific_huo@126.com, pacific_hua@126.com

Many of the comment information and real-time short texts contain the personal sentiment and tendencies of users, which are of great significance for users to personalize recommendations, interest circle division, network public opinion monitoring, and privacy protection [3]. How to use the computer technology to acquire and analyze these comments in the emotional information has attracted many experts and scholars to research. The information has attracted the competing research of many experts and scholars, which involves many fields such as artificial intelligence, natural language processing, data analysis and mining [4].

The main purpose of sentiment analysis is to process, extract, summarize and analyze the information in the text through different methods, so as to infer the emotion and viewpoint expressed by the author of the text, and divide the emotional tendency of the text through the subjective information contained in it. Text sentiment analysis can be divided into 3 tasks, including sentiment information extraction, sentiment information classification and sentiment information retrieval and induction [5]. Sentiment information classification is one of the important tasks of text sentiment analysis, and is also the focus of this paper.

Among them, Sina micro-blog has become the mainstream social platform with more than 300 million users. It is through the mutual attention between users, sharing real-time information, commenting on micro-blog content and other ways to socialize. But since micro-blog released the text words provisions shall not be more than 140 characters, the text length is short, the structure is different [6], the sentiment content in the text is relatively less. By analyzing the characteristics of this short text, this paper proposed a method of short text sentiment analysis in micro-blog bases on machine learning, which the proportion of emotional words and improves the accuracy of text sentiment analysis through the improvement of feature selection and weight calculation algorithm.

The specific contents of this paper are organized as follows: The second part mainly introduces the related work; the third part introduces the methods used in this paper and explains the proposed algorithm for feature extraction and weight calculation in detail; the fourth part gives the experimental results and analysis. Finally, the full text is summarized.

2. Related work

Natural language processing has extensive research scope, which includes sentiment analysis, machine translation, text classification, semantic analysis, etc. [7–9]. As the practicality of sentiment analysis has gradually become a research hotspot in recent years. Text sentiment analysis, also known as opinion mining, has become one of the hot topics in the field of natural language processing and data mining due to its practicality. Its main research methods can be divided into two categories, one method is based on the sentiment knowledge, another method is based on feature classification [10].

The method based on the sentiment knowledge is to use the existing sentiment knowledge, such as the sentiment lexicons, domain dictionary, etc. to classify the text. Turney et al. [11] put forward a method based on point mutual information for the co-occurrence of words in the corpus and use the method to make sentiment judgments on the words. Sentiment words are the basis of text sentiment

analysis [12], building sentiment lexicons was also very important research content. Baccianella et al. [13] used General Inquirer (GI) to construct an sentiment lexicons, Gyamfi et al. [14] used the MPQA corpus [15] to establish an sentiment seed lexicons and implemented an expansion of the sentiment lexicons in conjunction with WordNet. The construction of sentiment lexicon gradually transforms from relying completely on artificial addition to automatically adding new sentiment words through various methods [16–20].

The method based on feature classification is to select a large number of features that can represent the sentiment of the text and to classify the texts by means of statistics or machine learning methods such as Naive Bayes, K-proximity, support vector machine, etc. Nowadays, most of the paper are also used in this method [21–24], and these methods have also been successfully applied to different fields [25]. Pang et al. [26] classify the movie reviews, they used the similarity of sentences as a feature and training classification was conducted through NB and SVM that the experiments show classification based on reviews is more difficult than classification based on facts text. In the subjective and objective classification of English and French, Toprak et al. [27] used words, part of speech and lexical information of features, then used SVM as a classifier that the experiments proved that the recall of lexical information features to improve subjective and objective classification Better effect. Moraes et al. [28] used TF-IDF and GI as the feature extraction algorithm respectively and then classified used Pang et al.'s data sets with classifiers as NB, SVM and ANN, experiments proved the ANNs were found to be superior to unbalanced data sets good expressiveness. At the same time, sentiment analysis has also begun to gradually develop from coarse to fine-grained [29–31]. Fink et al. [32] conducted sentiment analysis using method of machine learning which subjective and objective classification using coarse-grained (sentence poles) and sentiment classification using fine-grained (clause or part-of-speech). Shi et al. [33] made use of the correlation of words and part-of-speech information to make sentiment analysis of hotel reviews and achieved good results. However, the method based on feature classification doesn't take into account the influence of sentiment words and there is still much room for improve the methods in feature extraction.

In addition, due to the continuous development of deep learning in various fields [34], the research of sentiment analysis gradually begins to tend to unsupervised classification [35–38]. Such methods are quite difficult and of great research significance also can save labor, but the method is not mature enough yet and the classified accuracy is relatively low which can't be used to the application now.

The development of the network has promoted the large-scale growth of data. The small language of foreign languages and the national language in Chinese have begun to occupy a place in the network. For the protection and development of various languages, the textual sentiment analysis of this small language has also attracted many eyes of scholars [39–42].

Due to the existing methods, there are still some defects [43]. Therefore, we combining the method based on sentiment knowledge and the method based on feature classification, and proposed a text sentiment analysis method based on machine learning, which applies the sentiment lexicon to feature extraction to achieve better classification results.

3. Fine-grained sentiment analysis method for short text in micro-blog

According to the disadvantages in existing methods for sentiment word has less proportion in feature and low weight, a new method of sentiment analysis for short text in Micro-blog is proposed, increases the proportion of sentiment words in feature extraction and weight calculation, the main process was shown in Fig. 1.

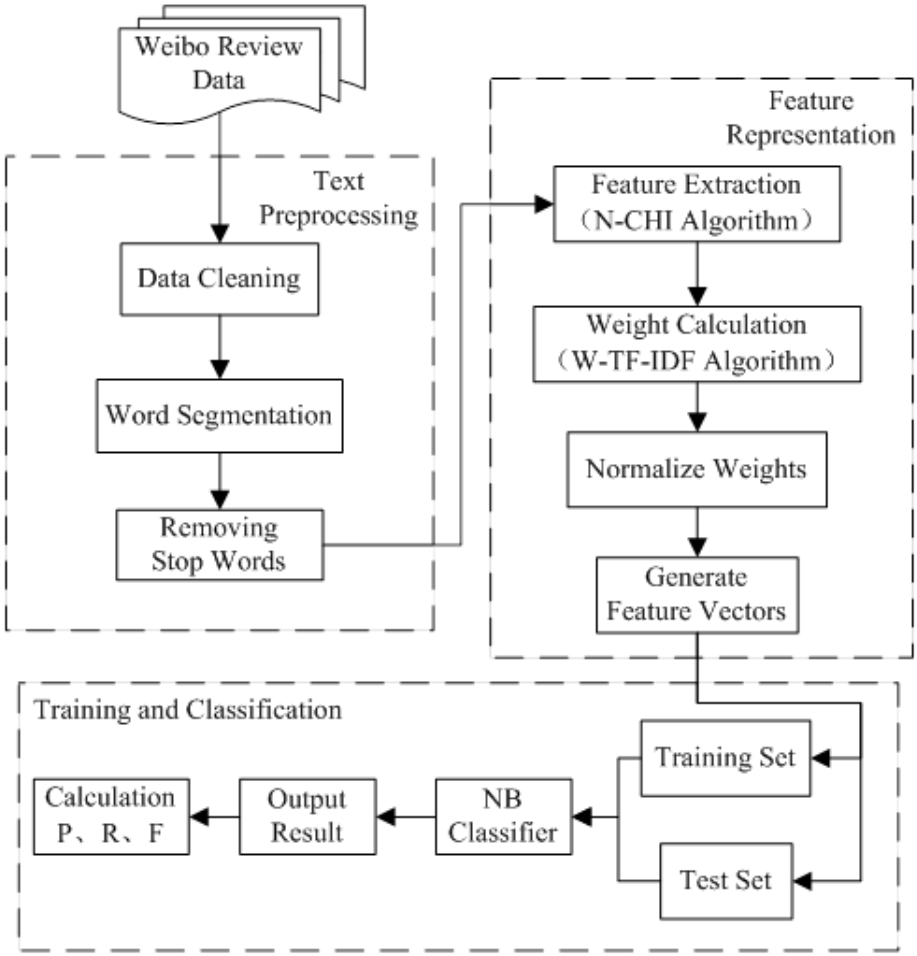

Fig. 1 *Sentiment analysis process.*

3.1 Text preprocessing

Since all the sentiment information was contained in the text, in order to extract features and analyze, first of all, the text should be preprocessed the text

to remove the excess interference information and retain useful information. Text preprocessing is the basic and important link in sentiment analysis, and the quality of preprocessing is very important for sentiment classification and information processing in the later stage. The contents of text preprocessing include data cleaning, word segmentation and removing stop word.

3.1.1 Data cleaning

As one of the most representative social networks nowadays, micro-blog has a huge amount of users and a large variety of comments on it. This is also due to the fact that this feature leads to the unique style of the texts in the Micro-blog which are often not in conformity with the changing forms Grammatical norms. However, these distinctive short texts contain a large amount of author's sentiment information. Due to the social nature of Micro-blog, these texts also include replies, references and sharing. Although these contents can express the user's emotions to some extent, many redundant and redundant information and noise are also involved in the extraction and influence to the theme and content. For example:

1、恭贺广州恒大淘宝队勇夺中超七连冠！#冠军不止 王朝不息##恒大七连冠#

Congratulations Guangzhou Hengda Taobao won seven consecutive Super League!# Champion not only the dynasty constant # # Hengda seven consecutive #

2、#恒大七连冠#，是顶点，更是恒大手球队衰败的转折点，一年不如一年!

 # Hengda seven consecutive #, is more vertices, is a turning point in the decline of the Hengda handball team, year after year!

The content between the two wells of the # Champion not only dynasty endless # and # Hengda seven consecutive # in the sentence belongs to Micro-blog's topic and the content expresses the joy of victory. The emotions expressed by users in comment 1 were consistent with the topics, while the contents of comment 2 were opposite to those expressed in the topics. Therefore, these topics do not fully represent the emotional tendency of the user, which play a role of interference and should be deleted.

For interference items similar to the hot topic existing in the data set, they should be cleaned up in order to lay a good foundation for the extraction of the later emotional information, and delete the contents as shown in Tab. I.

3.1.2 Word segmentation and removing stop word

After data cleaning, the text in the data set be preserved are strong availability, but in order to achieve better classification effect and need to have text processing in further, that is word segmentation and removing stop words. The objective was to completely remove the redundant information in the data set as much as possible and to retain the most valuable content.

Word segmentation: Both Chinese and English are very important parts of natural language. There are some similarities between the two languages but there are still many differences. English words are separated by a space, but there is no such structure in Chinese. We need word-level fine-grained sentiment analysis and

Delete Content	Before Cleaning	After Cleaning
Delete Topic	文章内容。这是荣耀手机#极限少女养成营#，她们去鸟巢滑行	这是荣耀手机，她们去鸟巢滑行
	(This is Rongyao mobile phone # limit girls develop camp #,They go to the bird's nest to slide)	(This is Rongyao mobile phone, They go to the bird's nest to slide)
Delete Mentioned Person	@三木摄影全球旅拍用镜头诠释西藏的神秘；	用镜头诠释西藏的神秘；
	(@ Miki global photography trip Use the lens to interpret the mystery of Tibet;)	(Use the lens to interpret the mystery of Tibet;)
Delete Reply	倍感欣慰// 蒙牛是很差劲	倍感欣慰
	(Feel happy //Mengniu is very bad)	(Feel happy)
Delete Source	湖北大学的研究生廖可富，1年申请16项发明专利（武汉晚报）。	湖北大学的研究生廖可富，1年申请16项发明专利。
	(Liao Kefu, a graduate student at Hubei University, applied for 16 invention patents in one year(Wuhan Evening News).)	(Liao Kefu, a graduate student at Hubei University, applied for 16 invention patents in one year.)
Delete Link	今天，给大家支两招	今天，给大家支两招
	https://mp.weixin.qq.com (Today, give you two strokes https://mp.weixin.qq.com)	(Today, give you two strokes)

Tab. I *Data cleaning content.*

one of the very important step is to carry out Chinese word segmentation. Chinese word segmentation involves grammar, statistics and other fields of knowledge and there are some difficulties, but now have a lot of more mature word segmentation system. Through the screening and comparison, this paper uses NLPIR Chinese word segmentation system (also known as ICTCLAS2016), this system has higher accuracy in Chinese word segmentation and has better recognition effect on colloquial vocabulary and network new words recognition. The segmentation effect is as follows:

原句：我一直都很爱喝蒙牛的纯牛奶 一直，很爱。

分词结果：我/ 一直/ 都/ 很/ 爱/ 喝/ 蒙牛/ 的/ 纯/ 牛奶/ 一直/，/ 很/ 爱/

Original sentence: I've always love to drink Mengniu's pure milk, and very love it.

Word segmentation result: I/ Have/ Always/ Love/ To/ Drink/ Mengniu/ Pure/ Milk/ ,/ and/ Very/ Love/ it/

Removing stop words: In Chinese there are some functional words that make the text fluent, such as function words, prepositions, quantifiers and so on. These words usually neither have practical meaning nor emotional information, but exist in a large amount in the text, at the same time, there are also some extra information such as Roman symbols, mathematical characters and punctuation, which we collectively refer to as stop words. There are many different types of stop word list, we integrate the existing stop words by screening and contrast, and retain some adverbs which have some influence on the emotion of the text. The complete segmentation data set to remove stop words, data preprocessing part is completed.

3.2 Feature extraction

In the process of sentiment analysis, feature extraction is also a very important link. If all the words appear directly as the characteristics of sentiment analysis, the data size is huge and the classification time is long. Therefore, we need to select words with strong emotional tendency as the features and calculate the appropriate feature dimensions, which can effectively avoid the over-fitting caused by excessive feature dimensions and improve the efficiency and speed of the classification.

There are many existing feature selection methods, such as point mutual information, information gain, chi-square test and so on. Comparing the advantages and disadvantages of various methods, point mutual information without considering the influence of redundancy features and in the information gain all categories can only use the same features, but the chi-square test is to avoid these shortcomings, so the final choice is improved and the improved algorithm is used to select features of the chi square test.

Chi-square test, also known as CHI statistics, reflects the degree of association between feature words and categories. The formula is as follows:

$$CHI(t,c) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}. \tag{1}$$

Among them t represents the characteristic term, c represents the category, N represents the total number of texts, A, B, C, and D represent as shown in Tab. II.

	Belongs to Category c	Does not Belong to Category c
Inclusion Feature t	A	B
Does not Include Feature t	C	D

Tab. II *Formula representation.*

Since $N, (A+C), (B+D)$ are all constant in Eq. 1, then Eq. 1 can be simplified as follows:

$$\mathrm{CHI}(t,c) = \frac{(AD-BC)^2}{(A+B)(C+D)}.$$

However, since the chi-square test does not calculate the frequency of the feature words in the statistical process, it is easy to select the words with the lower word

frequency and in the original method the influence of the sentiment words is not taken into consideration. For these problems, we join the word frequency and sentiment words into the algorithm that improved chi-square test as follows:

$$\text{N-CHI}(t,c) = \text{CHI}(t,c)f(t).$$

The improved algorithm was named N-CHI and its scores was used as the basis for selecting the feature words. $\text{CHI}(t,c)$ is the basis scores of the feature words, and $f(t)$ is the weight function of the feature words. The main contents of the $f(t)$ are as follows:

$$f(t) = \begin{cases} 2 & t = \text{SentimentWord} \quad \& \quad f > 1 \\ 1.5 & t = \text{SentimentWord} \quad \& \quad f = 1 \\ 1 & t \neq \text{SentimentWord} \quad \& \quad f > 5 \\ 0.5 & t \neq \text{SentimentWord} \quad \& \quad f \leq 5 \end{cases}$$

where f is the frequency of the feature word. The algorithm adds some weight to the original CHI algorithm, increases the score of the sentiment word as much as possible, makes the sentiment word more easily become the feature word and reduces the probability of the CHI algorithm selecting the low word frequency so as to improve the classification accuracy rate of the later. The improvement effect as shown Fig. 2.

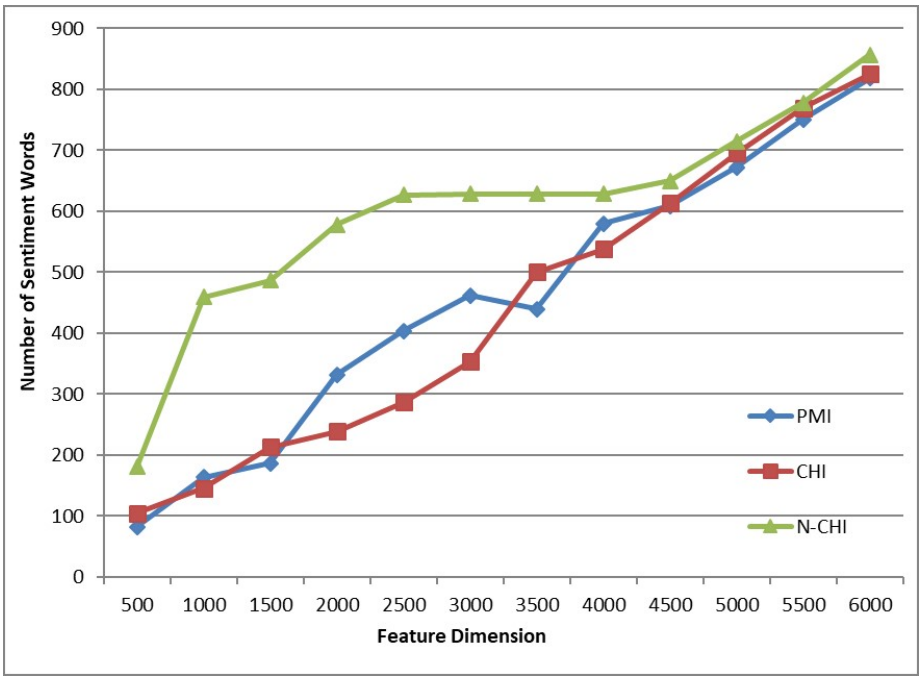

Fig. 2 *N-CHI improvement effect.*

It can be seen from the figure that the number of sentiment words contained in the improved algorithm is significantly higher than that of the no improved

algorithm in different dimensions. We choose the words with higher scores as the features and the choice of their feature dimensions is also very important which number is too much or too little will affect the accuracy rate. The specific content of the feature data selection will be discussed in the fourth section.

3.3 Weight calculation

The weight refers to the value of the feature. This value shows the correlation between this feature and all kinds of types, the commonly used weight calculation methods are frequency, TF-IDF and so on. Because the TF-IDF algorithm is easy to implement, not only the word frequency but also the influence of the anti-document frequency are taken into account. Therefore, the TF-IDF algorithm is improved and the improved algorithm is taken as the weight. For a document D consisting of K categories, the formula for TF-IDF is as follows:

$$W_{ik} = tf_{ik} \cdot idf_{ik} = tf_{ik} \cdot \log\left(\frac{D_k}{d_{ik} + x} + x\right).$$

Among them D_k represents the total number of documents in the k-th category, tf_{ik} represents the number of times that the feature word t_i appears in document D_k, it means the word frequency, d_{ik} represents the number of documents in which D_k contains the feature word t_i , and a constant x is usually set in idf_{ik} to smooth the function curve and prevent the case where the frequency is 0 in one category. However, since the original algorithm still does not take into account the influence of emotion words, modifiers and so on, this paper improves the TF-IDF algorithm and names it as W-TF-IDF. Its purpose is to make it more suitable for the task of emotional classification, which is distinguished from common text classification by the following formula:

$$W_{ik} = (tf_{ik} \cdot (1 + f_1(t_i)) + f_2(t_i)) \cdot \log\left(\frac{D_k}{d_{ik} + 0.5} + 0.5\right).$$

$f_1(t_i)$ represents the feature words t_i whether a sentiment words, if t_i is sentiment words the value is 1, not the value is 0; $f_2(t_i)$ represents the feature words t_i whether a modifier, if t_i is the modifier the value is 1, not a modifier value is 0, and the smooth curve constant x is set to 0.5. In this way, the weight of the sentiment words can be sufficiently improved and the influence of the modifiers on the sentiment is also taken into consideration. The improvement effect is shown in the following Tab. III.

	TF-IDF(pos)	W-TF-IDF(pos)	Sentiment Words	Modifier
喜欢 (Like)	246.92	507.84	YES	NO
今天 (Today)	142.39	142.39	NO	NO
极其 (Extremely)	6.21	12.99	NO	YES

Tab. III *W-TF-IDF improvement effect*

It can be seen from the table that the improved algorithm for sentiment words and modifiers has a significant increase in its weight. In the meantime, in order to make the calculation more convenient and rapid in the later training classification, we also normalized the expression of the weights.

Regardless of the improvement of the feature selection algorithm or the weight calculation algorithm, the complexity of the algorithm is increased to some extent compared with the original algorithm, but there is not much difference in the running time. The main purpose is to improve the text sentiment classification. The final correct rate.

3.4 Feature vector and constructing classifier

Feature vector: The selected features and calculated weights are not directly used as input into the classifier, so we store them in the feature vector with an n-dimensional vector of text d as follows:

$$d = [(t_1, w_1, (t_2, w_2, (t_3, w_3, \ldots, (t_n, w_n)].$$

t_i represents the n-th feature of the text, w_i represents the weight of the n-th feature, which is often used to represent features in the text classification because it is simple and easy to implement and does not consider the correlation and order of the features.

Construct classifiers: There are many ways to construct classifiers, such as method based on the statistical, methods based on machine learning, method based on deep learning, which method based on the machine learning are also included in the MaxEnt (Maximum Entropy), KNN(K-Nearest Neighbor), NB and SVM(support vector machine) etc. NB is a classic probabilistic model algorithm that determines the classification result by calculating the probability that text d belongs to class c, $p(d \mid c)$ is the conditional probability that text d belongs to class c and $p(c)$ is the prior probability of that class. This method has simple logic, classification algorithm, mature and stable characteristics, combined with the characteristics of text classification using NB constructing classifier and achieved better classification results, and compared with other more mature algorithms, the formula is as follows:

$$p(d \mid c) = \frac{p(d \mid c) \cdot p(c)}{p(d)}.$$

4. Experimental design and results analysis

In this part, we mainly introduce the process and result of the experiment in detail and give a brief analysis of the experimental results. The data set and training were processed in python on a desktop computer with a 3.3 GHz Intel Core i3 CPU and 2 GB RAM.

4.1 Experimental design

4.1.1 Experimental data

We used the COAE 2014 (Sixth corpus of Chinese tendentious analysis and uation) task 3 and task 4 corpus as the original data, the data set contains Micro-blog comments that have been labeled positive and negative two sentiment Which contains 1138 positive reviews and 1193 negative reviews. After the heavy reprocessing, have 1048 positive reviews and 993 negative reviews including 6222 words are included. We set 90 % of them as training set and 10 % as verification set.

For the sentiment lexicon, we have generated a new sentiment lexicon through selections, additions and subtractions between the sentiment words in Hownet and NTUD's sentiment lexicon in Taiwan University. The new sentiment lexicon include 7164 positive lexicons and 12,060 passive lexicons.

4.1.2 Evaluation method

Often used in machine learning, natural language processing, information retrieval and other fields of evaluation indicators are the following categories: accuracy, precision, recall and F-measure. The accuracy rate is defined as the ratio of the number of correctly classified samplers to the total number of samples in a given test set classifier. Although it can determine the effect of classifiers to a certain degree, it is not the most objective and effective. Therefore, In the experiment, we use the precision (P), the recall rate (R) and the F-measure (F) as the evaluation criteria. The calculation method is as shown in Eqs. 2, 3 and 4, and the meanings are shown in Tab. IV.

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{4}$$

	Positive	Negative
Be Judged as Positive t	TP	FP
Be Judged as Negative t	FN	TN

Tab. IV *Evaluation index.*

4.1.3 Selection the method of weight normalization

There are many weight normalization methods. The purpose of normalization is to apply the weights to the classifier better. The correct choice of normalization method has great influence on the efficiency and accuracy of data processing. Species normalization method for experimental comparison, the method is as follows:

Method 1: Linear function normalization is the linear transformation of the original data to the range of $[0, 1]$, the method to achieve the original data is scaled, where X^* is the normalized data, X is the original data, $X_{\max}$ and $X_{\min}$, respectively the maximum and minimum values of the original data set, the normalized formula is as follows:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}}.$$

Method 2: The weight of a feature in a class is divided by the sum of the weights of the feature in each class to normalize the feature weight to [0,1], taking into account the effect of low word frequency. When the feature weight is greater than the feature item the average value of all categories will be normalized weight retention, otherwise 0, the normalized formula is as follows:

$$X_i^* = \begin{cases} \frac{X_i}{\sum_{k=1}^{k} X_i} & X_i \geq \overline{X} \\ 0 & X_i < \overline{X} \end{cases}.$$

Method 3: Since the final purpose is to judge the textual tendencies, the weight is directly reduced to $[-1, 0, 1]$ according to the propensity. X_1 is the weight of the feature t in the positive class, X_2 is the weight of the feature t in the negative class, The method is as follows:

$$X^* = \begin{cases} -1 & X_1 = 0 \cup \frac{X_1}{X_2} < 0.6 \\ 0 & 0.6 \leq \frac{X_1}{X_2} \leq 1.5 \\ 1 & X_2 = 0 \cup \frac{X_1}{X_2} > 1.5 \end{cases}.$$

Since we do not know the optimal dimension of the feature temporarily, we choose 6000 as the feature dimension and then use the three normalization methods to process the feature weight respectively. The obtained results are shown in the following Tab. V.

	P	R	F
Method One	0.91	0.90	0.90
Method Two	0.89	0.88	0.88
Method Three	0.89	0.89	0.89

Tab. V *Comparison of normalization methods.*

As shown in the above table, the first method has higher P, R and F. The method retains the advantage of the sentiment word in the weight, while the second method ignores the function of the sentiment word. When dealing with the low frequency words, the range is too large, resulting in more weight is 0, although the third method is more intuitive to make the weight close to the classification, but the method ignores the weight of feature words size, and range selection has some limitations, scalability is not strong. So this paper choice the method one as a method for weight normalization.

4.2 Experimental results

In this paper we improved the algorithm of feature selection and weight calculation. In order to verify its improvement effect we designed four sets of experiments: The first group is choice the feature dimension; The second group is the improved verification of N-CHI algorithm; The third group is the improved verification of the W-TF-IDF algorithm; The fourth group is the effect of the two algorithms on the experimental results. The main results and analysis of specific experiments described in detail below.

Experiment 1: choice the feature dimension In the process of training classification the selection of the feature dimension always has been a very important part. When the number of feature words is small, these feature can't fully express the sentiment expressed in the text, resulting in a low correctness rate. If the feature is too much, in the training process easily lead to over fitting. Through many experiments, it is found that when the feature dimension between 5800-5900 the precision is the highest and the best classification result can be achieved. Therefore, we choose 5950 as the dimension of the feature, and the experimental result is shown in Fig. 3.

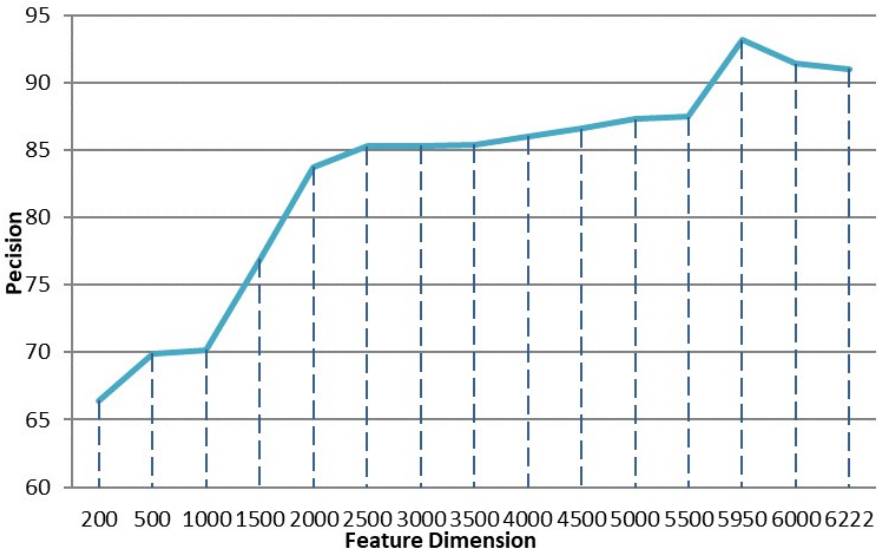

Fig. 3 *Choice the feature dimension.*

Experiment 2: improved verification of N-CHI algorithm In order to verify the effect of improved W-TF-IDF algorithm, the three contrast test design in this experiment, three group experiments were performed using the TF-IDF as the weight calculation: The first group using mutual information as the feature selection method, named MI; The second groups using CHI as a feature selection

method, named CHI; The third groups using improved N-CHI as feature selection method, named N-CHI. With the increase of dimensions, three methods of P, R, F-measure experimental results are shown as follows in Fig. 4, 5, 6.

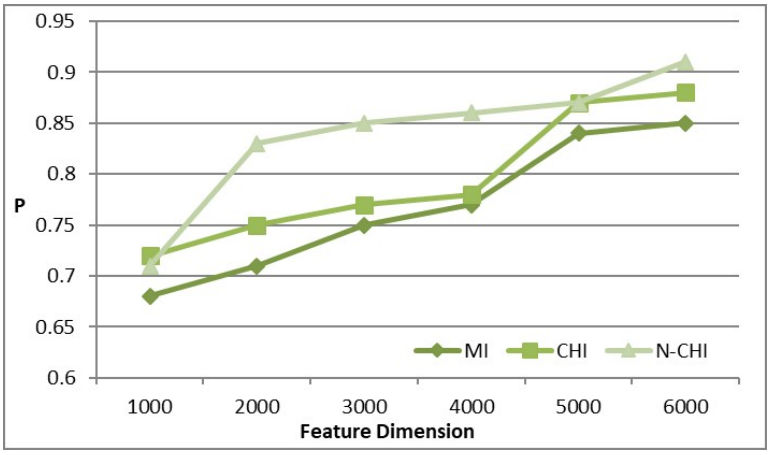

Fig. 4 *Improved verification of N-CHI algorithm -P.*

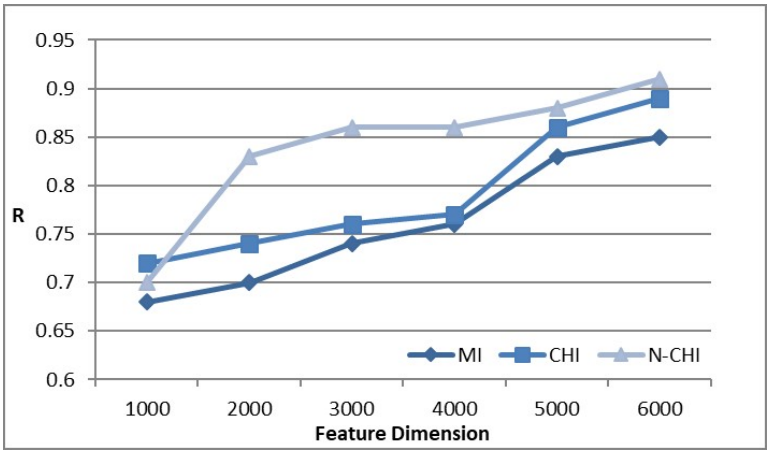

Fig. 5 *Improved verification of N-CHI algorithm -R.*

It can be seen from the above figure that this method is effective for the improvement of the method of feature extraction. Although the effect at lower latitudes is slightly lower than that of CHI, with the increase of dimension, the N-CHI algorithm is in P, R and F above point mutual information and CHI.

Experiment 3: improved verification of W-TF-IDF algorithm In order to verify the improvement of the W-TF-IDF algorithm, three sets of comparative experiments were designed in this experiment. Three sets of experiments all adopted

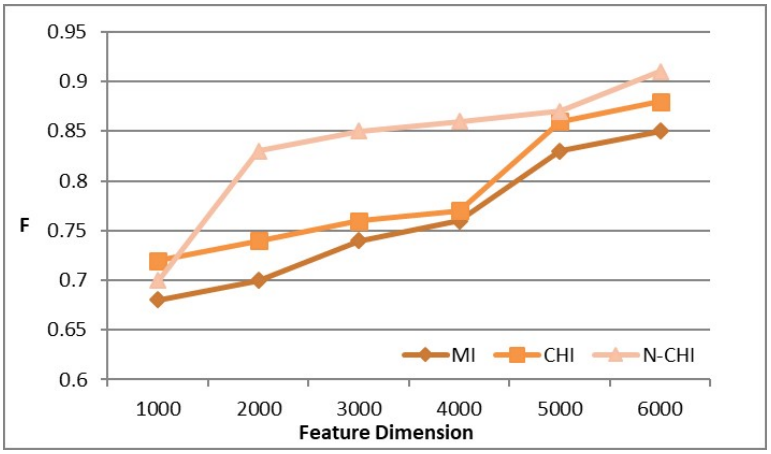

Fig. 6 *Improved verification of N-CHI algorithm -F.*

CHI as the method of feature selection: The first group uses the word frequency as the weight calculation method, named TF; The second group uses TF-IDF as the weight calculation method, named TF-IDF; The third group uses a modified W-TF-IDF as the weight calculation method, named W-TF-IDF. The experimental results are as follows in Fig. 7.

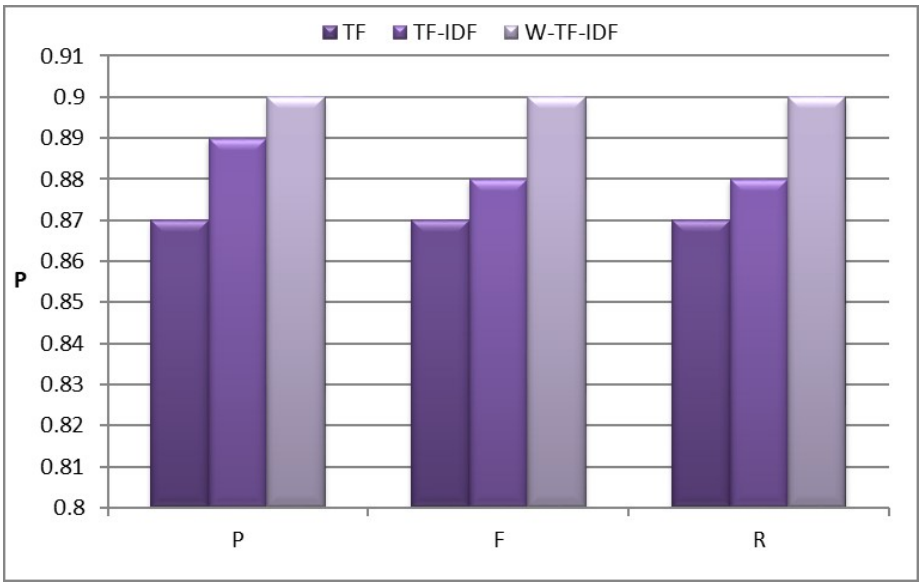

Fig. 7 *Improved verification of W-TF-IDF algorithm.*

It can be seen from the figure that the text is effective for the method of weight calculation. Although the effect of promotion is not obvious, the values of P, R and F are all improved.

Experiment 4: improved verification of comprehensive the algorithm In this paper, the method of feature selection and weight calculation is improved. In order to verify its improvement effect, a total of four sets of experiments were designed: The first group without any improvement, named CHI-IDF; The second group is a method that only improves the feature extraction; The third group is a method that only improves the weight calculation; The fourth group made two improvements to the text, named C-CHI-IDF, The experimental results are shown in the following Tab. VI.

	POS-P	NEG-P	F	R	F
CHI-IDF	0.86	0.92	0.89	0.88	0.88
N-CHI	0.89	0.94	0.91	0.90	0.90
W-TF-IDF	0.86	0.94	0.90	0.90	0.90
C-CHI-IDF	0.91	0.95	0.93	0.93	0.93

Tab. VI *Comparison of experimental results.*

It can be seen from the above table that the text achieves good results for the improvement of feature extraction and weight calculation, both in terms of positive, negative and overall precision.

The above experimental results are the average of multiple experiments. By comparison, it can be found that the improvement of feature extraction and weighting calculation is effective, the classification accuracy rate is significantly higher, and the weighting improvement is better than that of feature selection The effect is more obvious, and the improved algorithm has significantly improved the accuracy of the positive class.

Experiment 5: Selection of classification algorithms Using different classification algorithms to construct the classifier will result in different classification results. According to the classification task of text sentiment analysis, this paper uses the improved C-CHI-IDF method to select naive Bayes (NB) which is more suitable for text classification and KNN, and artificial neural network (ANN) to construct the classifier, and compare the classification results, the classification results are shown in Tab. VII.

	POS-P	NEG-P	F	R	F
NB	0.92	0.94	0.93	0.93	0.93
KNN	0.90	0.94	0.92	0.92	0.92
ANN	0.91	0.95	0.93	0.93	0.93

Tab. VII *Selection of classification methods.*

According to the above table, there is no significant difference in the classification effect between NB and KNN when the sample size is small. However, the amount of data continues to increase. The classification effect of NB is significantly higher than that of KNN, and it is better for classification effect of NB in positive.

The classification accuracy rate of ANN is not significantly different from the other two, but it has better performance for the sample imbalance of ANN.

4.3 Experimental analysis

Through the above experimental results and brief analysis, we can see that the fine-grained text sentiment classification method based on machine learning proposed by the text is effective. In order to reflect the experimental results more intuitively, we convert some data into the form of graphs. As shown in Fig. 8.

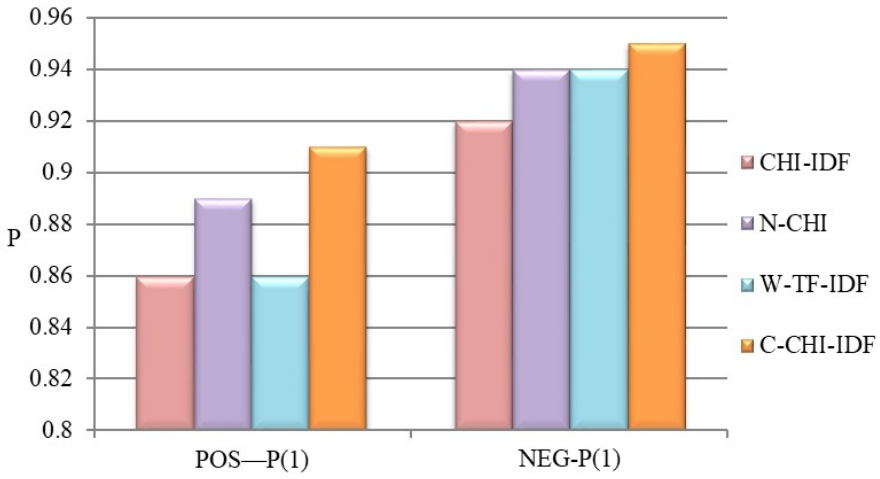

Fig. 8 *Comparison of experimental results.*

As can be seen from Fig. 8, we also found that the correct rate of the method for the negative class is significantly higher than the correct rate of the active class, although the number of negative class sentences is much lower than the active class. This shows that words in negative sentences are more likely to be selected as features, and in Chinese natural language, people are more likely to carry negative words when expressing negative emotions.

In this paper, we used a general evaluation data set. Other sentiment analysis methods of the data set and some evaluation results are compared with the results of this paper. The comparison results are shown in Tab. VIII.

5. Conclusion

According to existing method of the feature extraction and weight calculation without considering the influence for the sentiment words, this paper proposed a method of fine-grained Chinese sentiment analysis for short text in Micro-blog which proposed the N-CHI and W-TF-IDF two kinds of new algorithm based on improved the algorithm of CHI and TF-IDF. This method effectively improves the accuracy of text sentiment information classification and has certain universality.

	POS-P	NEG-P	P
Literature [42]	–	–	0.900
Literature [6] Method 1	0.895	0.837	–
Literature [6] Method 2	0.919	0.886	–
COAE2014-sjtu	0.954	0.885	0.919
Bjut-coae2014	0.914	0.915	0.914
scool	0.769	0.800	0.785
WB-SA	0.892	0.914	0.903
hut	0.901	0.848	0.875
LEO-WH_Run1	0.964	0.791	0.877
Medians	0.891	0.850	0.877
C-CHI-IDF	0.910	0.950	0.930

Tab. VIII *Comparison of normalization methods.*

However, this method has some limitations that does not consider the influence of the relationship between part-of-speech and words on affective information, which will be the focus of the fourth part of this article.

Acknowledgement

This research is supported by National Natural Science Foundation of China under the Grant 61672210 and supported by the Henan Research Program of Foundation and Advanced Technology under the Grant 162300410183.

References

[1] ZKIK K., ORHANOU G., HAJJI S.E. Secure Mobile Multi Cloud Architecture for Authentication and Data Storage. *IGI Global*. 2017, 7(2), pp. 62–76, doi: 10.4018/IJCAC.2017040105.

[2] JAIN A.K., GUPTA B.B. PHISH-SAFE: URL Features based Phishing Detection System using Machine Learning. In: *Cjc, Csi*, 2015.

[3] VElURU S., GUPTA B.B., RAHULAMATHAVAN Y., et al.Privacy Preserving Text Analytics: Research Challenges and Strategies in Name Analysis. *Handbook of Research on Securing Cloud-Based Databases with Biometric Applications*. 2015, pp. 67–92, doi: 10.13140/2.1.3017.3760.

[4] PIRYANI R., MADHAVI D., SINGH V.K. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing and Management*. 2016, 53(1), doi: 10.1016/j.ipm.2016.07.001.

[5] ZHAO Y.Y., QIN B., LIU T. Sentiment analysis. *Journal of Software*, 2010, 21(8), pp. 1834–1848, doi: 10.3724/SP.J.1001.2010.03832.

[6] YUAN D., ZHOU Y., LI R., et al. Sentiment analysis of microblog combining dictionary and rules. In: *Ieee/acm International Conference on Advances in Social Networks Analysis and Mining*, IEEE, 2014, pp. 785–789, doi: 10.1109/ASONAM.2014.6921675.

[7] HARRAG F., HAMDI-CHERIF A., El-QAWASMEH E. Performance of MLP and RBF neural networks on Arabic text categorization using SVD *Neural Network World*. 2010, 20(4), pp. 441–459.

[8] BARIGOU F. Improving K-nearest neighbor efficiency for text categorization *Neural Network World*. 2016, 26(1), pp. 45–66, doi: 10.14311/NNW.2016.26.003.

[9] MAUTNER P., MOUCEK R. Processing and categorization of Czech written documents using neural networks. *Neural Network World*. 2012, 22(1), pp. 53–66, doi: 10.14311/NNW.2012.22.004.

[10] KUPKA J., TOMANOVA I. Some extensions of mining of linguistic associations. *Neural Network World*. 2010, 20(1), pp. 27–44.

[11] PETER TURNEY M.L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 2003, pp. 315–346, doi: 10.1145/944012.944013.

[12] KU L.W., LIANG Y.T., ChEN H.H. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *In AAAI-CAAW,*, 2006.

[13] BACCIANELLA S., ESULI A., SEBASTIANI F. Multi-facet Rating of Product Reviews. *European Conference on Ir Research on Advances in Information Retrieval*. Springer-Verlag, 2009, pp. 461–472, doi: 10.1007/978-3-642-00958-7_41.

[14] GYAMFI Y., WIEBE J., MIHALCEA R., et al. Integrating knowledge for subjectivity sense labeling. *Human Language Technologies: the 2009 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 10–18, doi: 10.3115/1620754.1620757.

[15] WIEBE J., BRECK E., BUCKLEY C., et al. *NRRC summer workshop on multi-perspective question answering*, 2002, doi: 10.1017/CBO9781107415324.004.

[16] WANG K.E., XIA RUI. An approach to Chinese sentiment lexicon construction based on conjunction relation. *Proceedings of the 14th China National Conference on Computational Linguistics*. Guangzhou, China: CCL, 2015.

[17] KRESTEL R., SIRESDORFER S. Generating contextualized sentiment lexica based on latent topics and user ratings. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. New York, NY: ACM. 2013, pp. 129–138, doi: 10.1145/2481492.2481506.

[18] LIANG JUN, CHAI YU-MEI, YUAN HUi-BIN, ZAN HONG-YING, LIU MIN. Deep learning for Chinese micro-blog sentiment analysis. *Journal of Chinese Information Processing*. 2014, 28(5), pp. 155-61.

[19] HUANG M.L., YE B.R., WANG Y.C., CHEN H.Q., CHENG J.J., ZHU X.Y. New word detection for sentiment analysis. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, pp. 531–541, doi: 10.3115/v1/P14-1050.

[20] ZHAO YANYAN, QIN BING, SHI QIUHUI, LIU TING. Large-scale Sentiment Lexicon Collection and Its Application in Sentiment Classification. *Journal of Chinese Information Processing*, 2017, 31(2), pp. 187–193.

[21] YANG Y., XU C., RRN G. Sentiment analysis of text using SVM. *Lecture Notes in Electrical Engineering* (2012)138 LNEE, pp. 1133–1139, doi: 10.1007/978-1-4471-2467-2_134.

[22] SILVA N.F.F.D., HRUSCHKA E.R., JR E.R.H. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*.. 2014, 66, pp. 170–179, doi: 10.1016/j.dss.2014.07.003.

[23] AGARWAL A., XIE B., VOVSHA I., et al. Sentiment analysis of Twitter data. *The Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011, pp. 30–38.

[24] HE FEIVAN, HE YANXIANG, LIU NAN, et al. A Micro-blogging Short Text Oriented Multi-class Feature Extraction Method of Fine-grained Sentiment Analysis. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 50(1), (Jan. 2014), doi: 10.13209/j.0479-8023.2014.013.

[25] ZHOU X, WANG S, XU W, et al. Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier. *International Conference on Bioinformatics and Biomedical Engineering*. Springer, Cham, 2015, pp. 201–209, doi: 10.1007/978-3-319-16483-0_20.

[26] PANG, BO, LEE, et al. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of Acl*, 2004, pp. 271–278, doi: 10.3115/1218955.1218990.

[27] TOPRAK C., GUREVYCH I. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Deft'09 Text Mining Challenge*, 2009.

[28] MORAES R., VALIATI J.F, NETO W.P.G. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*. 2013, 40(2), pp. 621–633, doi: 10.1016/j.eswa.2012.07.059.

[29] ZIRN C., NIEPERT M., STUCKENSCHMIDT H., et al. *Fine-Grained Sentiment Analysis with Structural Features*, 2011.

[30] GUXMAN E., MAALEJ W. How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. *Requirements Engineering Conference*.IEEE, 2014, pp. 153–162, doi: 10.1109/RE.2014.6912257.

[31] TEH P.L., PAK I., RAYSON P., et al. Exploring fine-grained sentiment values in online product reviews. *Open Systems*. IEEE, 2016, pp. 114–118, doi: 10.1109/ICOS.2015.7377288.

[32] FINK C.R., CHOU D.S., KOPECKY J.J., et al. Coarse-and Fine-Grained Sentiment Analysis of Social Media Text. *Johns Hopkins Apl Technical Digest*. 2011, 30(1), pp. 22–30.

[33] SHI H., ZHOU G., QIAN P., et al. An unsupervised fine-grained sentiment analysis model for chinese online reviews. *International Journal on Information*. 2012, 15(10), pp. 4277–4294.

[34] WANG S.H, LV Y.D., SUI Y., et al. Alcoholism Detection by Data Augmentation and Convolutional Neural Network with Stochastic Pooling. *Journal of Medical Systems*. 2018, 42(1), pp. 2, doi: 10.1007/s10916-017-0845-x.

[35] HUANG F., ZHANG S., ZHANG J., et al. Multimodal Learning for Topic Sentiment Analysis in Microblogging. *Neurocomputing*2017, 253(C), pp. 144–153, doi: 10.1016/j.neucom.2016.10.086.

[36] BAECCHI C., URICCHIO T., BERTINI M., et al. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications*2016, 75(5), pp. 2507–2525, doi: 10.1007/s11042-015-2646-x.

[37] CHEN F., GAO Y., CAO D., et al. Multimodal hypergraph learning for microblog sentiment prediction. *IEEE International Conference on Multimedia and Expo*. IEEE, 2015, 1-6, doi: 10.1109/ICME.2015.7177477.

[38] PORIA S., PENG H., HUSSAIN A., et al. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, 2017, doi: 10.1016/j.neucom.2016.09.117.

[39] KAPUKARANOV B., NAKOV P. Fine-grained sentiment analysis for movie reviews in Bulgarian. *Proceedings of Recent Advances in Natural Language Processing*. 2015, 9(7), pp. 266–274, doi: 10.1126/science.aaa8685.

[40] LUO YAWEI, TIAN SHENGWEI, YU LONG, TURGUN·IBRAHIM, ASKAR·HAMDULLA. *Sentiment Analysis of Uyghur Text for Fine-grained Opinion Mining* 2016, 30(1), pp. 140–148.

[41] AL-SMADI M., QAWASMEH O., AL-AYYOUB M., et al. Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. *Journal of Computational Science*, 2017.

[42] ZHANG XIAOMEI, LI RU, WANG BIN, WU DI, GAO JUNJIE. Subjective and Objective Classification of Micro-blog Based on Feature Fusion. *Journal of Chinese Information Processing*. 2014, 28(4), pp. 50–57.

[43] TANG D., WEI F., QIN B., et al. Sentiment Embeddings with Applications to Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*. 2016, 28(2), pp. 496–509, doi: 10.1109/TKDE.2015.2489653.