Comput Syst Sci & Eng (2020) 3: 223–232
© 2020 CRL Publishing Ltd

International Journal of
Computer Systems
Science & Engineering

The Method for Extracting New Login Sentiment Words From Chinese Micro-Blog Based on Improved Mutual Information

Guangli Zhu, Wenting Liu, Shunxiang Zhang*, Xiang Chen and Chang Yin

State Key Laboratory of Mining Response and Disaster Prevention and Control in Deep Coal Mines (Anhui University of Science and Technology), Anhui 232001, China

The current method of extracting new login sentiment words not only ignores the diversity of patterns constituted by new multi-character words (the number of words is greater than two), but also disregards the influence of other new words co-occurring with a new word connoting sentiment. To solve this problem, this paper proposes a method for extracting new login sentiment words from Chinese micro-blog based on improved mutual information. First, micro-blog data are preprocessed, taking into consideration some nonsense signals such as web links and punctuation. Based on preprocessed data, the candidate strings are obtained by applying the N-gram segmentation method. Then, the extraction algorithm for new login words is proposed, which combines multi-character mutual information (MMI) and left and right adjacent entropy. In this algorithm, the MMI describes the internal cohesion of the candidate string of multiple words in a variety of constituted patterns. Then, the candidate strings are extended and filtered according to frequency, MMI, and right and left adjacency entropy, to extract new login words. Finally, the algorithm for the extraction of new login sentiment words is proposed. In this algorithm, the Sentiment Similarity between words (SW) is determined in order to measure the sentiment similarity of a new login word to other sentiment words and other new login sentiment words. Then, the sentiment tendency values of new login words are obtained by calculating the SW to extract new login sentiment words. Experimental results show that this method is very effective for the extraction of new login sentiment words.

Keywords: Chinese micro-blog; New login sentiment words; N-gram segmentation; Right and left adjacency entropy; Multi-character Mutual Information; Sentiment Similarity between the Words

INTRODUCTION

Chinese micro-blog is a service-oriented social networking site that has openness, timeliness and diversity. Users can express their thoughts with words, pictures or videos anytime and anywhere. At the same time, micro-blog has become the main place for the emergence of new login words which tend to be related to a hot event, and therefore are expected to contain certain sentiment. These new login words with sentiment are not included in the existing dictionary. This will reduce the accuracy of the Chinese word segmentation of the text, which in turn affects the analysis of the part of speech of new login sentiment words and, subsequently, the sentiment analysis of the text after the word segmentation.

At the same time, there are several problems with micro-blog data: First, the authenticity of micro-blog content is an issue worth discussing, since there can be several duplicate pages (in order to improve ranking, artificially repeated references). Moreover, the content of the micro-blog page (including text, pictures, etc.) is not standardized, which seriously affects the quality of the extracted new login words. Second, because the micro-blog text involves a wide range of domains, the extraction

**Corresponding author, sxzhang@aust.edu.cn

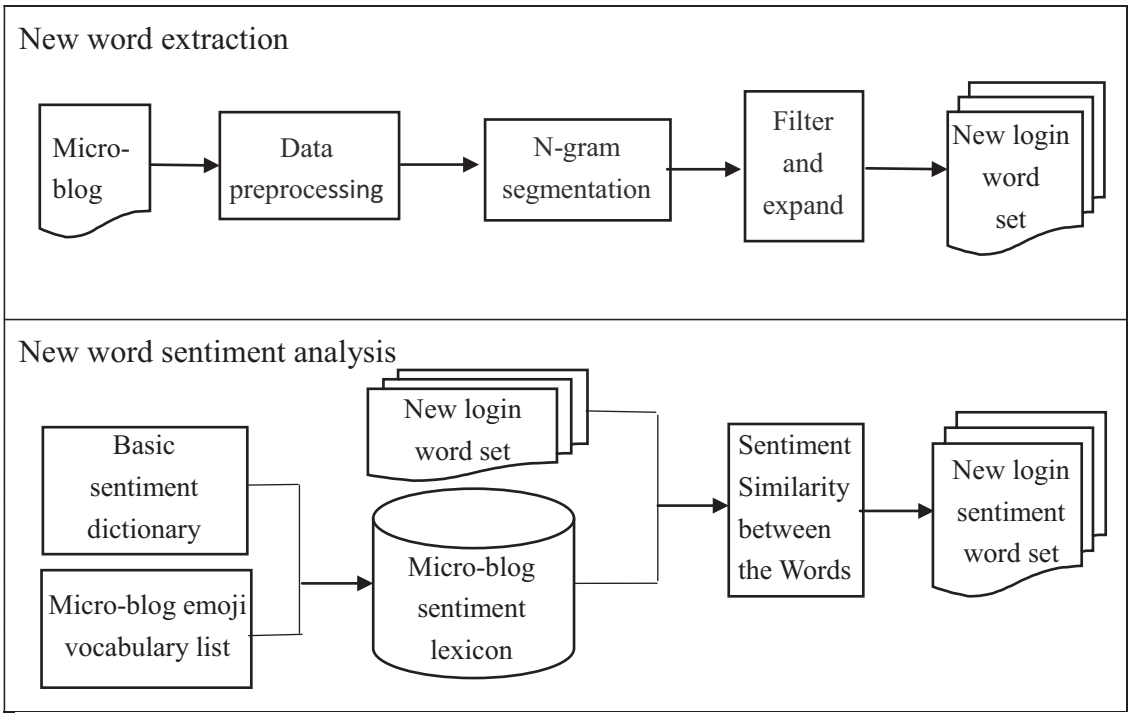

Figure 1 Extraction process of new login sentiment words in micro-blog

method for a certain domain will greatly affect the accuracy of the extraction of new words. Third, since the new micro-blog login word does not exist in the dictionary, its sentiment tendencies cannot be determined immediately.

In order to solve the above problems and meet the requirements of quick and accurate extraction of new login sentiment words, this paper proposes a new method for extracting new login sentiment words based on Multi-character Mutual Information and Sentiment Similarity between Words. The first problem can be effectively resolved in the data preprocessing step. Thus, the method is divided into two stages to solve the other two problems: new login word extraction and new login word sentiment analysis, as shown in Figure 1.

(1) **The new login word extraction**. First, the preprocessed micro-blog data is split into N-gram strings to obtain the candidate strings. Then, the internal and external statistics of the candidate string are calculated according to the MMI and the left and right adjacent entropy to obtain a candidate new login word set. Finally, the obtained candidate new login word set is compared with the dictionary, and the existing words in the dictionary are deleted to obtain a new login word set.

(2) **The new login word sentiment analysis**. According to the general sentiment words and other new words, Semantic Orientation Pointwise Mutual Information (SO-PMI) is improved to obtain SW. Then the sentiment tendency values of the new login words are calculated according to SW to judge their sentiment tendency. This judgement yields neutral new login words that are subsequently deleted, leaving the new login sentiment words.

The rest of the paper is organized as follows: Section 1 introduces related work. Section 2 introduces the preprocessing of micro-blog data and the extraction method of new login words in micro-blog. Section 3 introduces the sentiment tendency analysis method of new login words and the extraction algorithm of new login sentiment word. The experimental and some analysis results are given in section 4. Finally, conclusions are presented in section 5.

1. RELATED WORK

1.1 Methods for Extraction of New Login Words

Currently, there are three methods for recognising new login words: rule-based, statistics-based methods, and methods based on statistics and rules.

Rule-based methods generally construct templates based on the principles of word formation, with semantic information or lexical information. Then the new login word is matched. Yanzhao Han et al. [1] used a noise reduction algorithm and a conditional random field model to mark the parts of speech in micro-blog data, and used the Bayesian method to correct the parts of speech of a large proportion of homophonic words. Tang et al. [2] extracted new words using the rules formed by the conditional random field model. He et al. [3] systematically studied seven sub-model modeling methods based on the morphology of a low-resource language.

Statistics-based methods often use statistical linguistic features or combine machine learning methods to extract new words. Usually, both are involved when using statistical methods to extract new words. In regard to feature selection, in general, statistical features can be measured by calculating the phoneme posterior probabilities [4], string frequency, string cohesion, string liberalization [5,6], and Assembled Mutual Information

[7]. In terms of machine learning strategies, a variety of them have been applied to new word extraction tasks. These include LSTM [8], recurrent neural networks [9], Semantic and Topic Context Models [10], and CRF [11].

One method based on statistics and rules combines the two methods to complement each other. Sheikh et al. [12] combined CBOW and Skipgram neural networks with contextual semantic information. Kim et al. [13] used n-gram and fingerprint-based filtering chain rules and robust regression to improve efficiency. Zhang et al. [14] proposed a novel unsupervised new word recognition method based on Weak word string function and traditional statistics. Yan et al. [15] proposed a new dynamic feature, SD-SPP, based on the existing static features and SVM classifiers. Liang et al. [16] designed a new method for Chinese word boundary detection based on edge likelihood (EL) and proposed a domain-independent Chinese New Word Detector (DICND). Using the knowledge in BabelNet, Du et al. [17] proposed three different methods - direct training data, domain adaptive technology and BabelNet API - to obtain OOV translations. Wang [18] used suffix tree method to identify new login sentiment words.

1.2 Micro-Blog Sentiment Analysis Method

Micro-blog sentiment analysis involves the judgement of vocabulary and sentences in micro-blog, and then determining the user's emotional tendency. The current methods used for microblog sentiment analysis are mainly divided into dictionary-based methods, machine learning-based methods and methods based on dictionary and machine learning.

The dictionary-based method uses a series of sentiment lexicons and rules to deconstruct the sentence, analyze and match words with the dictionary (generally with part of speech analysis, syntactic dependency analysis), calculate the emotional value, and finally use the emotional value to indicate the emotional tendency of the text. Dey et al. [19] analysed the sentiment in user reviews through rules and built a new dictionary (called Senti-N-Gram dictionary). Wu et al. [20] compared words in multiple source domains with words in specific domains using a sentiment graph. Zhao et al. [21] constructed a micro-blog sentiment dictionary through rules involving parts of speech, position of words and context information. Lyu et al. [22] used BAWL as a seed dictionary and used SO-NPMI to analyze the polarity of words in the data set; Milagros et al. [23] added semantic dependence to the sentiment classification provided by the emoji creators in the encyclopedia.

A machine learning-based approach treats sentiment analysis as a classification problem. Shuang et al. [24] proposed a neural network based on the sentiment information collector-extractor architecture (SICENN) for sentiment analysis. Kong et al. [25] constructed a neural architecture to train emotionally-conscious word embedding by integrating three kinds of knowledge: context words and their constituent characters, the polarity of sentences, and labeled words. Khan et al. [26] proposed a new method for calculating the similarity of English word pairs, which can be adapted to a single embedding of word pairs to determine similarity. Liao et al. [27] proposed a multi-level semantic fusion method based on representation learning to learn recognition features, focusing on the emotions implicit in facts presented in sentences. Yoo et al. [28] designed the Polaris system and improved the accuracy of its sentiment analysis and prediction by using the latest deep learning techniques. Zhou et al. [29] introduced deep learning theory and used the LSTM model based on an attention mechanism to conduct sentiment analysis in order to better grasp the sentiment information in the text.

Based on the combination of dictionary and machine learning, the dictionary is integrated into the machine learning model. Kamper et al. [30] trained a convoluted neural network (CNN) using auxiliary information in the form of known word pairs. Jeremy et al. [31] proposed a bilingual dictionary-based bilingual sentiment embedding model (BLSE). Fu et al. [32] combined/utilzed an emotional dictionary to calculate sentiment decision scores, and proposed an improved supervised learning improvement model based on language rules and sentiment scores. Zeng et al. [33] combined traditional shallow text part-of-speech features with sentiment features, and applied this to XGBoost to generate a microblog sentiment analysis integration model. Hung [34] proposed three WSD technologies based on WOM document context to build a WSD-based SentiWordNet dictionary. Abdi et al. [35] proposed a method based on deep learning to classify user opinions expressed in comments (called RNSA).

At present, most scholars pay attention only to new login word extraction and micro-blog sentiment analysis, and lack of relevant research on Micro-blog new emotional words. However, micro-blog new login sentiment word recognition is based on the extraction of new words, and it is also a vital part of the micro-blog sentiment analysis process. Therefore, this paper proposes a method for improving mutual information to extract new login sentiment words from the micro-blog.

2. MICRO-BLOG NEW LOGIN WORD EXTRACTION

2.1 Micro-Blog Data Preprocessing

Due to the large number of micro-blog users, the wide range of fields, the irregular writing of texts, and the limited number of words, etc., there is a lot of noise in the crawled micro-blog data. To solve this problem, the micro-blog data needs to be preprocessed, which can effectively reduce the text processing time and the storage space required by the data. Data preprocessing consists of deletion and replacement.

1) **Deletion.** Links, repeat punctuation in the micro-blog data are deleted. The fixed string that comes with the micro-blog program is deleted, such as "#***#" and "@+username", etc.

2) **Replacement.** Punctuation marks and stop words obtained by using the ICTCLAS word segmentation in the micro-blog text are replaced with spaces. The traditional characters are replaced with simplified Chinese.

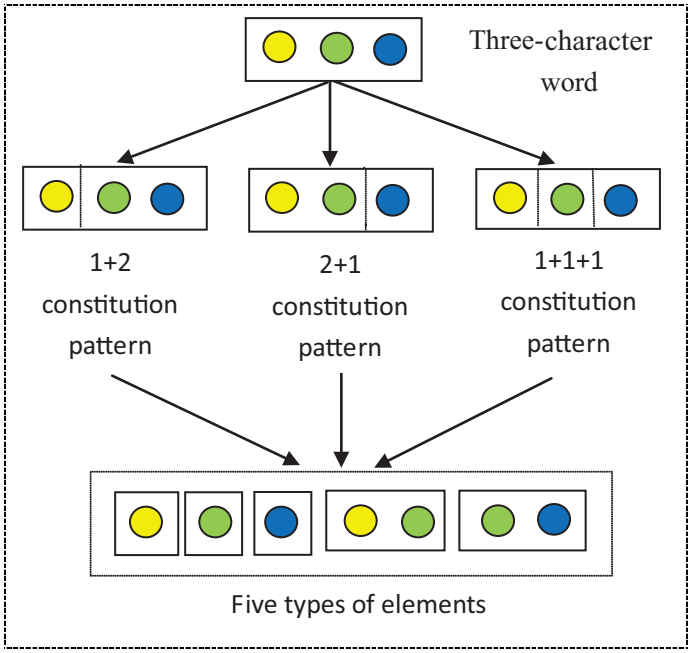

Figure 2 The pattern constitution of three-character words

2.2 Micro-Blog New Login Words Extraction

2.2.1 N-Gram Segmentation Method

The general word segmentation system divides a sentence according to existing dictionaries. However, this method may cause incorrect segmentation and failure to recognise new words. For example, "王经/理/了/理/袖口", the word segmentation system may be divided this sentence into "王/经理/了/理/袖口". The new login word "王经" will be missed. Thus, the segmentation method used in this paper is the N-gram segmentation method.

The basic idea of N-gram segmentation is to scan the text word-by-word and divide the N words; each N word after the segmentation constitutes a string. At present, the bigram and trigram segmentation techniques are relatively mature. Also, the research finds that the new login words generally consist of 2 to 4 words, so N in this paper is 2 and 3.

2.2.2 Multi-Character Mutual Information

Mutual information is a representation of the degree of dependency between two words. The traditional formula for mutual information is defined as:

$$MI(x, y) = \log \frac{P(xy)}{P(x)P(y)} \tag{1}$$

Where $P(xy)$ represents the ratio of co-occurrence frequency of x and y in the corpus to the total number of words in the corpus. $P(x)$ and $P(y)$ represent the ratio of frequency of x and y appearing in the corpus in proportion to the total number of words, respectively.

According to the formula, it can be intuitively understood that the traditional mutual information has the following two problems when new login word extraction is applied: 1) candidate words are considered to be divided only into two parts of the constituted pattern; and 2) only one type of constituted pattern is considered. In order to solve these two problems, it is necessary to consider different constituted patterns of multi-character candidate words (the number of candidate words are more than two words). Figure 2 shows the patterns constituted of three-character words. Similarly, there are seven types of patterns constituted of four-character words, and nine types of all the elements contained in all the constitution patterns. Based on this, all the elements included in the constitution patterns can be considered in the denominator of the mutual information, so that the calculation of the internal cohesion of the new login words is more accurate. The relevant definition and formulas are given below.

Definition 1 For multi-character mutual information (MMI) to measure the internal cohesion of the string in a variety of word-constituted patterns, the formula is as follows:

$$MMI\ (a_1a_2\cdots a_n) = \log \frac{P(a_1a_2\cdots a_n)}{\sqrt[T]{\prod_{j=1}^{T} P_j(a_1a_2\cdots a_n)}} \tag{2}$$

$$T = \sum_{i=1}^{n-1} (n-i+1) \tag{3}$$

$$P(a_1a_2\cdots a_n) = \frac{F(a_1a_2\cdots a_n)}{N} \tag{4}$$

$$P_j(a_1a_2\cdots a_n) = \frac{F_j(a_1a_2\cdots a_n)}{N} \tag{5}$$

Where N is the total number of micro-blog text in the corpus. $F(a_1a_2\cdots a_n)$ is the frequency of string $a_1a_2\cdots a_n$ that appears in the corpus. $F_j(a_1a_2\cdots a_n)$ represents the frequency of occurrence of the $j-$th element in all constituted patterns in the corpus. T is the number of all element types in all constituted patterns of the string $a_1a_2\cdots a_n$.

Algorithm 1 Micro-blog new login word extraction algorithm

Input: micro-blog text set F, candidate string set FC, micro-blog candidate new login word set WC candidate string frequency threshold θ_1, multi-character mutual information threshold θ_2, left and right adjacent entropy threshold θ_3, θ_4

Output: micro-blog new login words W

1: Pre-processed micro-blog text set F
2: Use n-gram segmentation method to segment F, and count the frequency f_i of each string c_i
3: if $f_i \geqslant \theta_1$, put the string c_i into FC
4: else delete the string
5: end if
6: Calculate $MMI(a_1a_2\cdots a_n)$ of each string c_i in FC
7: if $MMI(a_1a_2\cdots a_n) < \theta_2$, remove string c_i from FC
8: end if
9: Perform steps 10 to 12 for bigram strings in FC
10: Calculate left and right adjacency entropy $\{E_l, E_r\}$ of bigram string in FC
11: if $E_l \geqslant \theta_3$ & & $E_r \geqslant \theta_4$, the left and right boundaries are determined, and the string enters WC
12: else delete the bigram string
13: end if
14: Perform steps 14 to 19 on the trigram string in FC
15: if $E_l \geqslant \theta_3$, the left boundary is determined, calculate E_r of the candidate string and perform step 17.
16: else the string is extended one word to the left, calculate E_r of the expanded candidate string
17: if $E_r \geqslant \theta_4$, the right boundary is determined, the string enters WC
18: else the string is extended one word to the right, the expanded candidate string enters WC
19: end if
20: end if
21: if compare WC with the dictionary, and filter out new login words that are not in the dictionary to construct W
22: end if

2.2.3 Left and Right Adjacency Entropy

In the context adjacency analysis, different adjacent feature quantities may be adopted, such as adjacency type, adjacency entropy, adjacency pair type, and adjacency pair entropy. The comparison of P@N indicators shows that the effect of adjacency entropy is generally better than that of adjacency type. The effect of adjacency entropy is better than that of adjacency pairs. Therefore, entropy is more effective than types for measuring the diversity of a pragmatic environment. At the same time, the adjacency entropy is better than the adjacency pair entropy. Thus, this paper chooses the adjacent entropy as the external statistic to extract the new word.

The left adjacency entropy formula is defined as:

$$E_l = -\sum_{i=1}^{|V_l|} \frac{n_i}{n} \log\left(\frac{n_i}{n}\right) \tag{6}$$

The right adjacency entropy formula is defined as:

$$E_r = -\sum_{j=1}^{|V_r|} \frac{m_j}{\mathrm{m}} \log\left(\frac{m_j}{m}\right) \tag{7}$$

Where, $|V_l|$ and $|V_r|$ are the number of types of left adjacent words and right adjacent words, respectively. n and m are the total number of left adjacent words and right adjacent words, respectively. n_i and m_j are the number of a certain left adjacent words and right adjacent words, respectively.

2.3 New Word Extraction Algorithm

According to the discussion above, in order to extract the new words of micro-blog, the preprocessed text is first divided into candidate strings by N-gram (N = 2, 3) segmentation. And then the candidate strings are multi-filtered, including multi-character mutual information threshold filtering, left and right adjacency entropy threshold filtering, and dictionary contrast filtering. The dictionary here uses the collection of words that expanded the HowNet text lexicon through the synonym word forest. This step is used to continuously expand and filter the candidate strings. The specific algorithm is as follows.

1) Step 3–5 are used to determine whether the string is a candidate string; that is, whether the word frequency of the string is greater than a preset threshold. If so, the string enters the candidate string set. Otherwise, it is deleted.

2) Steps 6–8 are used to determine whether the multi-character mutual information of each string is less than the threshold. If so, it is deleted; if not, it is saved.

3) Steps 10–13 are used to determine whether the left and right adjacent entropy of the bigram string are greater than preset thresholds. If so, the bigram string is entered in the new login word set.

4) Steps 15–20 are used to calculate the left and right adjacent entropy of the trigram string. If both are greater than the threshold, they are placed in the new login word set. Otherwise, the string is expanded to the left or right, and then is entered in the new login word set.

5) Steps 21–22, are used to determine whether the candidate new login words are in the dictionary. If so, they are deleted. Otherwise, they are saved.

The time complexity of the algorithm is as follows. To begin, the first scan of the micro-blog text is performed to determine and

Table 1 18 groups of micro-blog positive and negative emoticons.

Positive	Negative
[可爱][威武][鼓掌][嘻嘻]	[哼][怒][酸] [泪] [挖鼻]
[哈哈][爱你][亲亲][抱抱]	[晕][衰][吐][白眼][鄙视] [悲伤]
[酷][心][耶][good][加油]	[伤心][弱] [吃惊] [可怜] [怒骂]
[赞啊][给力][笑哈哈]	[生病][抓狂]
[太开心][好爱哦]	

set up the candidate string set. At this time, the time complexity is $O(n)$, and n is the number of candidate strings. Candidate strings are filtered by the calculation of MMI, and the time complexity is $O(n)$. The total number of bigram strings is k, and the time complexity of the extension is $O(k)$. The number of trigram strings is $n-k$, and the time complexity of the extension is $O(n)$. Therefore, the maximum time complexity of the algorithm is $O(n)$.

3. THE ANALYSIS OF THE SENTIMENT TENDENCY OF NEW LOGIN WORDS

3.1 Constructing Micro-Blog Basic Sentiment Thesaurus

The quality of the selected micro-blog sentiment words has a great influence on the discrimination of new login words' sentiment tendency. This paper uses the dictionary-based method. The sentiment word set of the de-recombined of Hownet sentiment dictionary and NTSUSD (simplified Chinese sentiment polarity dictionary of Taiwan university) are selected as the basic sentiment dictionary. Furthermore, the emojis with obvious sentiment tendencies and the frequency of occurrence of the first 18 groups of positive and negative emojis on the micro-blog are selected as the word list of micro-blog emojis, as shown in Table 1. Then, the basic sentiment dictionary and the micro-blog emoji vocabulary list are de-recombined to obtain the micro-blog emoji vocabulary.

3.2 The Analysis of the Sentiment Tendency of New Login Words

In general, new login words in the micro-blog are a catharsis or expression of users' sentiments, and sentiment words are a direct reflection of users' sentiments. Both express the users' views of an event, and the sentiment tendencies of new login words and sentiment words appearing in the micro-blog text posted by a certain user are similar. Hence, the sentiment tendencies of new login words can be analyzed from the sentiment words that the users use.

Considering the limited length of micro-blog text and the sparse data and so on, the traditional pointwise mutual information of sentiment tendency is improved to obtain sentiment similarity between the words. Sentiment similarity between the words not only considers the existing sentiment words in the micro-blog sentiment lexicon, but also considers the influence of the sentiment of a new login word on the sentiments expressed through other new login words. The relevant definitions and formulas are given below.

Definition 2 SW (Sentiment Similarity between the Words) measures the sentiment similarity of a word to other sentiment words and other new login words in the same text. The formula is:

$$SW_j = \alpha(PA\_PMI_j - NA\_PMI_j) \tag{8}$$

$$PA\_PMI_j = \frac{1}{|C^+|}\sum_{i=1}^{|C^+|}\log\frac{P\left(X, C_i^+\right)}{P(X)P\left(C_i^+\right)} \tag{9}$$

$$NA\_PMI_j = \frac{1}{|C^-|}\sum_{i=1}^{|C^-|}\log\frac{P\left(X, C_i^-\right)}{P(X)P\left(C_i^-\right)} \tag{10}$$

$$\alpha = \frac{\sum_j PA\_PMI_j}{\sum_j NA\_PMI_j} \tag{11}$$

Where $P(X, C_i^+)$ represents the co-occurrence probability of the new word X and *i-th* positive sentiment word C_i^+. $P(X)$ represents the probability that the new word X appears alone. $P(C_i^+)$ represents the probability that the *i-th* positive sentiment word C_i^+ appear alone. $|C^+|$ indicates the total number of positive sentiment words. $\sum_j PA\_PMI_j$ represents the sum of the average point mutual information of new login words and positive sentiment words. $\sum_j NA\_PMI_j$ represents the sum of the average point mutual information of a new login words and negative sentiment words. α represents the intensity ratio of positive and negative sentiment tendency in the corpus.

It can be seen from the above that in the context of the constructed micro-blog basic sentiment dictionary, the sentiments expressed by the obtained micro-blog new login words can be determined through the calculation of SW, and then the new login sentiment words that meet the conditions are obtained. The algorithm for extracting new login sentiment words is as follows.

Steps 3–7 are used to determine the sentiment tendency of new login words. If the value of SW is greater than a certain threshold, it is positive. If the value of SW is less than a certain threshold, it is negative. Others are neutral and, therefore, are deleted. In step 8, the new login sentiment words and the negative new login sentiment words are combined to form a new login

Algorithm 2 Micro-blog new login sentiment word extraction algorithm

Input: micro-blog text set (F), micro-blog sentiment word set (E), threshold of SW θ_5, θ_6, positive new login sentiment words set PE, negative new login sentiment words set NE

Output: micro-blog new login sentiment word (EW)

The new word extraction algorithm gets W of micro-blog
Calculate SW for each word in W
If $SW_j > \theta_5$, the new login word enters PE
else if $SW_j < \theta_6$, the new login word enters NE
else delete the word
end if
end if
Combine PE and NE to get EW

sentiment word for the micro-blog. It can also be seen from the above algorithm that only one scan is needed for the micro-blog's new login word set. Hence, the time complexity is $O(n)$, where n is the number of new login words.

4. EXPERIMENTS

4.1 Experimental Data

1) The crawled data/Raw data: all micro-blog text on three different hot topics from November 2018 to March 2019 (they are "军训式应援 (military training-style support)", "杨超越登上人物杂志 (Yang appears in the magazine "People")" and "翟天临学术造假 (Zhai academic fraud)"), about 1.3 G, is used for the extraction of micro-blog new login sentiment words.

2) Stop-words thesaurus: a stop-words thesaurus (excluding a large number of English words and Chinese punctuation) is obtained by recombining the dictionaries, including "Harbin Institute of Technology stop-words vocabulary", "Sichuan University machine learning smart laboratory stop-words vocabulary", "Baidu stop-words vocabulary", etc. A total of 1598 stop-words were obtained and then used for the preprocessing of the micro-blog data.

3) Dictionary: this comprised a collection of words that expanded the HowNet text lexicon through the synonym word forest for the screening of new candidate words.

4) Micro-blog basic sentiment vocabulary: the sentiment word set obtained by combining the sentiment dictionaries is used as the basic sentiment dictionary of this paper. The sentiment dictionaries include HowNet Sentiment Dictionary, the Taiwan University Simplified Chinese Sentiment Polarity Dictionary (NTUSD), and the list of micro-blog esoteric emojis in this article. It is used to determine the sentiment tendency of new login words.

4.2 Experimental Method

To extract the new login sentiment words, firstly, all the crawled microblog data are preprocessed, and the candidate string is obtained by means of N-gram segmentation. All candidate strings whose frequencies satisfy the threshold are sequentially filtered out. Then, three statistical features are calculated for each candidate string, including MMI, the left adjacent entropy E_l and the right adjacent entropy E_r. A new login word set is obtained by combining the dictionaries. Finally, the sentiment tendency of new login words are determined by the SW results, and the superiority and inferiority of the algorithm are evaluated by the evaluation index.

The method commonly-used to evaluate new login sentiment word extraction generally draws on the evaluation index in the information retrieval model, and P (precision rate), R (recall rate), and the comprehensive index F_1 (F-Measure) are used to evaluate the accuracy of the algorithm. The following formula is used for the calculation:

$$P = TP/(TP + FP) \tag{12}$$

$$R = TP/(TP + FN) \tag{13}$$

$$F = 2PR/(P + R) \tag{14}$$

where TP indicates that the obtained new login (sentiment) word is actually a new login (sentiment) word, predicted as a new login (sentiment) word. FP indicates that the obtained new login (sentiment) word is actually a non-new login (sentiment) word, and predicted as a new login (sentiment) word. FN indicates that the obtained new login (sentiment) word is actually a new login (sentiment) word, and predicted as a non-new login (sentiment) word. The values of P and R are between 0 and 1. The closer the value is to 1, the higher is the precision or recall. The value of F-measure is the harmonic mean of the precision rate and the recall rate.

4.3 Experimental Analysis

In order to verify the validity of the algorithm, two experiments have been carried out in this paper. One is the validity verification of the new login word extraction algorithm of micro-blog. And the other is the verification of the validity of the new login sentiment word extraction of micro-blog.

Because the content of each topic is different, each is discussed separately. The first topic is "军训式应援 (military training-style support)". This topic is proposed for a certain star in China. Because of the good image of the star, all the words appearing in micro-blog have positive connotations. The second

Table 2 Samples of high-frequency new login word recognition results.

	bigram	trigram	4-gram
Number	1632	753	232
Topic one	开挂、打 call、牛逼	吊炸天、嗨上天	C 位出道、前方高能
Topic two	冲鸭、杠精、控评	键盘侠、走花路	炒鸡厉害、喜大普奔
Topic three	学霸、脱粉、实锤	零容忍、演技派	顶级流量、压力山大

Table 3 Comparison of extraction results of new words.

	P (%)	R (%)	F (%)
Traditional N-gram method	24.65	12.03	16.17
Literature [7]	30.35	12.08	17.28
Algorithm in paper	35.51	20.55	26.03

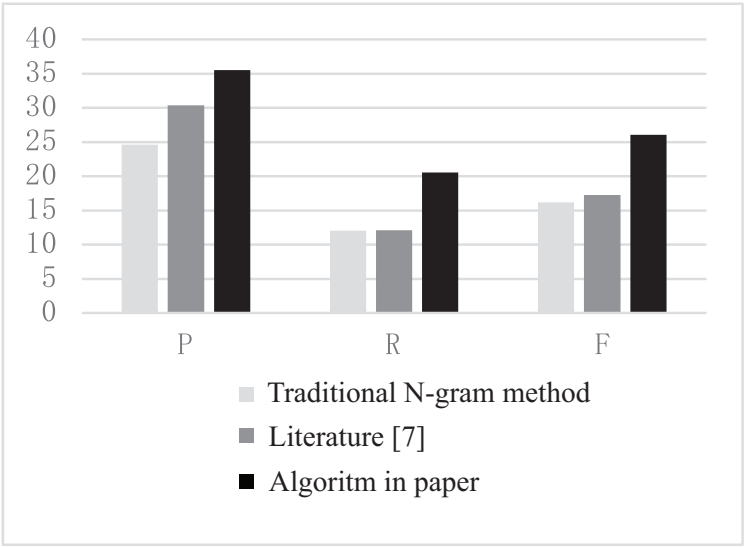

Figure 3 Comparison of new word extraction results

topic is "杨超越登上人物杂志 (Yang appears in the magazine "People")". Since "People" magazine usually publishes articles about highly influential people, members of the public will, naturally, have different opinions about the appearance of a new star in the public domain. The third topic is "翟天临学术造假 (Zhai academic fraud)". Academic fraud is an intolerable practice. Therefore, most of the words appearing in the micro-blog under this topic are words with negative connotations.

Table 2 shows some examples of high-frequency new login words obtained for the three topics by algorithm 1.

The micro-blog data crawled by the web crawler is processed by means of algorithm 1 to obtain a new login word set, and the evaluation index of the new login word set is calculated. Algorithm 1 in this paper is compared with the traditional N-gram method and the literature [7]. The results are shown in Table 3.

It can be seen from Table 3 that the algorithm proposed in this paper has a better precision rate, recall rate and F-measure compared with the traditional N-gram method and the method reported in the literature [7], achieving good results. The traditional N-gram method has a low recognition rate for multi-character new login words, and it does not take into account the importance of new login words' internal and external statistics in regard to new login word recognition, so its precision is relatively low. The method proposed in the literature [7] relies on word segmentation, which can cause some words to be segmented wrongly.

Table 5 shows the extraction results of the new login sentiment word compared with the method proposed in the literature [18].

As can be seen from Figure 4, the method proposed in the literature [18] uses the mutual information and adjacency information entropy to filter the candidate strings. However, it does not take into account the semantic information of the new login word, which leads to many of the new login words not being sentiment words. The method in this paper adds the semantic information of the sentiment dictionary of words on the basis of statistics, thus improving Precision rata, recall rate and F-measure compared with the literature [18].

Table 4 Sample of new login sentiment words extraction.

Positive	Negative
给力、牛逼、开挂、比心、冲鸭	杠精、尼玛、学渣、辣鸡、凉凉
吊炸天、演技派、走花路	伤不起、键盘侠、辣眼睛
C位出道、家里有矿、宝藏少女	天凉王破、一脸懵逼

Table 5 Comparison of new login sentiment words extraction results.

	P (%)	R (%)	F (%)
Literature [18]	47.10	36.54	41.15
Text method	60.24	42.35	49.74

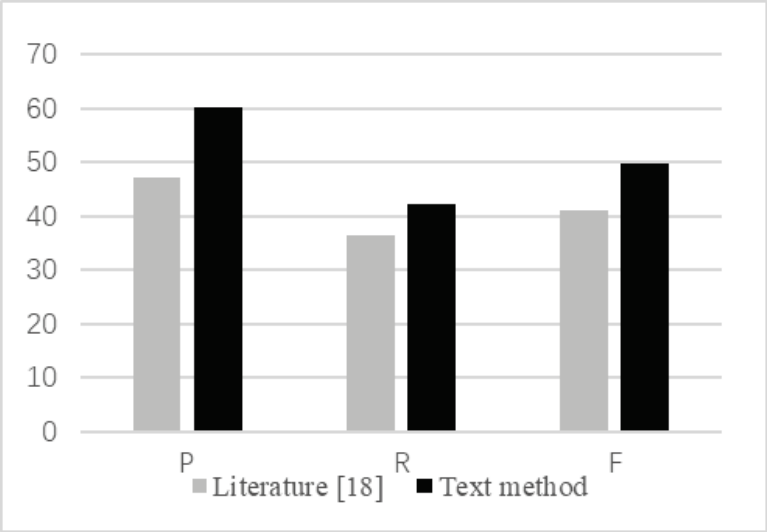

Figure 4 Comparison of new login sentiment word extraction results

5. CONCLUSIONS

The combination of rules and statistics proposed in this paper can improve the extraction of new login sentiment words in a microblog. The algorithms proposed in this paper can more accurately solve the problem of new login word extraction and the sentiment judgment of new login word. The main contributions of this paper are:

1) Multi-character mutual information and left and right adjacency entropy extend and filter the candidate string. The multi-character mutual information addresses the drawback of the traditional mutual information that considers only one type of pattern constituting multi-character strings. The left and right adjacency entropy further extends and filters the candidate word set obtained in the previous step, which effectively prevents the omission of some new login words.

2) The general sentiment words and other new login sentiment words in the context can influence the determination of the sentiment implied by a new login word. So the improved/modified traditional SO-PMI has been presented, which is combined the obtained intensity ratio of positive and negative sentiment tendency. It can effectively improve the accuracy of determining the sentiment in a new login word.

The feasibility of this algorithm also shows that it has made some improvement to the method used for extracting new login sentiment words. At the same time, the accurate extraction of new login sentiment words is of great help to Chinese word segmentation, the sentiment analysis of the text and the recognition of the part of speech of a new login sentiment word after word segmentation.

6. ACKNOWLEDGMENT

This research work was supported in part by the Anhui Provincial Natural Science Foundation Project (No.: 19808085 MF189) and the Anhui University Top Talent Cultivation Project (No. gxbjZD15). Natural Science Research Project of Anhui College (No. KJ2018A0285).

REFERENCES

1. Han, Y.Z., Qiao, Y.N., Fan, Y.P., et al.: Research on new word-of-speech recognition based on conditional random field model and text error correction[J]. Journal of Nanjing University (Natural Science), 52(02): 353–360(2016).

2. Tang, Z., Fu, Z.M., Gong, Z.R., et al.: A parallel conditional random fields model based on spark computing environment. Journal of Grid Computing, 15(3): 323–342(2017).
3. He Y.Z., Baumann P, Fang H, et al.: Using Pronunciation-Based Morphological Subword Units to Improve OOV Handling in Keyword Search. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(1): 79–92(2016)
4. Khokhlov Y, Tomashenko N, Medennikov I, et al.: Fast and Accurate OOV Decoder on High-Level Features. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp2884–2888, 2017.
5. Mei L, Huang H, Wei, X., et al.: A Novel Unsupervised Method for New Word Extraction. Science China Information Sciences, 59(9): 92102(2016)
6. Zhang S, Zhu H, Xu Z. The extraction method of new logining word/term for social media based on statistics and N-increment. Journal of Ambient Intelligence and Humanized Computing, 30: 1–11(2017)
7. Li W, Guo K, Shi Y, et al.: DWWP: Domain-specific New Words Detection and Word Propagation System for Sentiment analysis in the Tourism Domain. Knowledge-Based Systems, 146, 203–214(2018).
8. Lv, Z., Kang, J., Zhang, W.Q., et al.: An LSTM-CTC based verification system for proxy-word based OOV keyword search. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5655–5659, 2017.
9. Gundogdu, B., Yusuf, B., Murat, S.: Generative RNNs for OOV keyword search. IEEE Signal Processing Letters, 26(1): 124–128(2019).
10. Sheikh, I., Fohr, D., Illina, I., et al.: Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition. IEEE/ACM Transactions on Audio Speech and Language Processing, 25(3): 598–610(2017).
11. Liu, W., Wu, B., Xie, W., et al.: A New Word Discovery Method Based on Ancient Chinese Corpus[J]. Journal of Chinese Information Processing, 33(01): 46–55(2019).
12. Sheikh, I., Illina, I., Fohr, D., et al.: Document Level Semantic Context for Retrieving OOV Proper Names. IEEE International Conference on Acoustics. IEEE, pp. 6050–6054 ,2016.
13. Kim, S., Shin, H., Baek, C.H., et al.: Learning New Words from Keystroke Data with Local Differential Privacy. IEEE Transactions on Knowledge and Data Engineering, 1–1(2018).
14. Zhang, J., Huang, K.Y., Liang, C., et al.: Unsupervised neologism recognition research for Chinese social media corpus [J]. Journal of Chinese Information Processing, 32(03): 17–25+33(2018).
15. Yan L, Bai, B, Chen W, et al.: New Word Extraction from Chinese Financial Documents. Signal Processing Letters, 24(6): 770–773 (2017)
16. Liang, Y.Z., Yang, M., Zhu, J., et al.: Out-domain Chinese new word detection with statistics-based character embedding Natural Language Engineering, 25(2): 239–255(2019).
17. Du, J.H., Way, A., Zydron, A.: Using BabelNet to improve OOV coverage in SMT 10th International Conference on Language Resources and Evaluation, LREC 2016, pp. 9–15, 2016.
18. Wang F.: Research on Affective New Word Discovery Based on Weibo [J]. Journal of Software, 36(11): 6–8, 2015
19. Dey, A., Jenamani, M., Thakkar, J.J.: Senti-N-Gram: An n-gram lexicon for sentiment analysis. Expert Systems with Applications, 103: 92–105(2018).
20. Wu, F.Z., Huang, Y.F.: Sentiment domain adaptation with multiple sources. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers. v 1, pp. 301–310, 2016.
21. Zhao, C.J., Wang, S.G., Li, D.Y.: Exploiting social and local contexts propagation for inducing Chinese microblog-specific sentiment lexicons. Computer Speech and Language, 55: 57–81(2019).
22. Lyu, K., Kim, H.: Sentiment Analysis Using Word Polarity of Social Media. Wireless Personal Communications, v 89, n 3, pp. 941–958, 2016.
23. Milagros, F.G., Jonathan, J.M., Silvia, G.M., et al.: Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Systems with Applications, v 103, pp. 74–91, 2018.
24. Shuang, K., Zhang, Z.X., Guo, H., et al.: A sentiment information Collector–Extractor architecture based neural network for sentiment analysis. Information Sciences, 467, 549–558 (2018).
25. Kong, L., Li, C.Y., Ge, J.D., et al. Construction of Microblog-Specific Chinese Sentiment Lexicon Based on Representation Learning. 15th Pacific Rim International Conference on Artificial Intelligence, PRICAI, v 11012, pp. 204–216, 2018.
26. Khan, N., Shaukat, A.: New Word Pair Level Embeddings to Improve Word Pair Similarity. 2017 14th IAPR International Conference on Document Analysis & Recognition, ICDAR, V 05, pp. 57–62, 2018.
27. Liao, J., Wang, S.G., Li, D.Y.: Identification of fact-implied implicit sentiment based on multi-level semantic fused representation. Knowledge-Based Systems, 165: 197–207(2019).
28. Yoo, S.Y., Song, J.I., Jeong, O.R.: Social Media Contents based Sentiment Analysis and Prediction System. Expert Systems with Applications, 105: 102–111(2018).
29. Zhou, W., Liu, Y., Cai, J.: Analysis of Weibo Emotion Based on Attention Mechanism. Information Studies: Theory & Application, 41(03): 89–94(2018).
30. Kamper, H., Wang, W.R., Livescu, K.: Deep convolutional acoustic word embeddings using word-pair side information. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4950–4954, 2016.
31. Jeremy, B., Roman, K., Sabine, S.I.W.: Bilingual sentiment embeddings: Joint projection of sentiment across languages. ACL 2018–56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), v 1, pp. 2483–2493, 2018.
32. Fu, Y., Shi, W.: Study on the Sentiment Analysis of Weibo Based on Enhanced Supervised Learning. Journal of Intelligence, 37(12): 130–134+167(2018).
33. Zeng, Z.M., Wan, P.Y.: Analysis of Public Security Events Based on Fusion Evolution. Information Science, 36(12): 3–8+51(2018).
34. Hung, C., Chen, S.J.: Word sense disambiguation-based sentiment lexicons for sentiment Classification. Knowledge-Based Systems, 110: 224–232 (2016).
35. Abdia, A., Shamsuddina, S.M., Hasana, S., et al.: Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Information Processing and Management. 56, 1245–1259(2019).