Comput Syst Sci & Eng (2020) 6: 399–410

International Journal of
Computer Systems
Science & Engineering

A Discovery Method for New Words From Mobile Product Comments

Hai Yang Zhu[1,*,†], Xiaobo Yin[1,*,‡], Shunxiang Zhang[1,*,§], Zhongliang Wei[1,¶], Guangli Zhu[1,‖] and Meng-Yen Hsieh[2,*,]**

[1] *School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan, China*
[2] *Department of Computer Science and Information Engineering, Providence University, Taichung 43301, Taiwan*

A large number of new words in product reviews generated by mobile terminals are valuable indicators of the privacy preferences of customers. By clustering these privacy preferences, sufficient information can be collected to characterize users and provide a data basis for the research issues of privacy protection. The widespread use of mobile clients shortens the string length of the comment corpus generated by product reviews, resulting in a high repetition rate. Therefore, the effective and accurate recognition of new words is a problem that requires an urgent solution. Hence, in this paper, we propose a method for discovering new words from product comments based on Mutual Information and improved Branch Entropy. Firstly, by calculating the Co-occurrence Frequency and Mutual Information between words and adjacent words, the character strings of words after pre-processing and word segmentation are expanded left and right respectively to discover the potential word set. The candidate set of new words is obtained by means of an improved support filtering algorithm. Finally, a new word set is built by applying an improved Branch Entropy filtering algorithm and removing old words. The experimental results show that this method can accurately and effectively identify new words in product comments.

Keywords: privacy preference; new word identification; product comments; Mutual Information; Branch Entropy

1. INTRODUCTION

People are free to post comments on online shopping platforms to express personal opinions. However, most product reviews are fragmented and written in colloquial language. Because of this informal expression, new words continue to emerge and spread rapidly throughout the network. These salient features make it challenging to recognize new words that are being used in product reviews.

Some researchers have found that many new words often refer to user's privacy preferences [2,3], and that some new words are indicative of the user's age, personality, occupation, etc. For example, young people like to use popular new words online, and people in computer-related industries like to use computer terminology, etc. Hence, the clustering of words that a user tends to use for online comments can identify and summarize the characteristics of users, which is an important aspect of processing privacy protection [4,5].

Moreover, new words often reveal a lot of critical information about the user such as emotions, attitudes, and opinions. The new words extracted from online shopping platforms [6–8] can provide the basis for market analysis and user satisfaction surveys. It is also a crucial part of determining the quality of work. Therefore, how to effectively recognize new words in the review text according to the characteristics of product reviews is an important task in natural language processing.

To solve this problem, it is necessary to conduct a comprehensive examination of the characteristic of product reviews and to identify new words according to these characteristics. When attempting to recognize new words in product reviews, traditional new-word recognition methods have the following shortcomings:

1. According to a large number of corpus statistics, the word-formation patterns of new words are divided into 11 types. These three single-word phrases, "1 + 1", "1 + 1 + 1"

*Correspondence: xbyin@aust.edu.cn; sxzhang@aust.edu.cn; mengyen@pu.edu.tw Tel.: +88-642-6328-001(M.H.).
†838412840@qq.com(H.Z.)
‡xbyin@aust.edu.cn(XY.)
§sxzhang@aust.edu.cn(S.Z.)
¶zhlwei@aust.edu.cn(Z.W.)
‖glzhu@aust.edu.cn(G.Z.)
**mengyen@pu.edu.tw(M.H.)

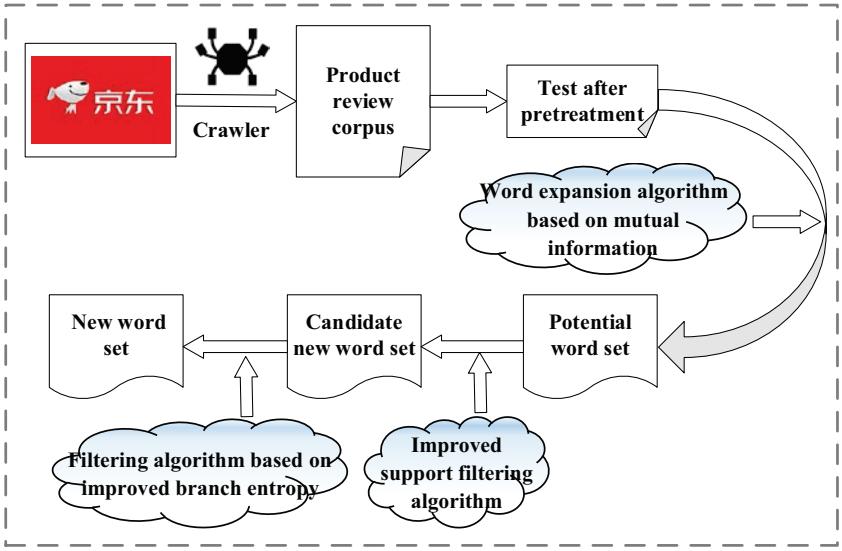

Figure 1 Procedure for new word recognition in product reviews.

and "1 + 1 + 1 + 1", are commonly used in word-formation patterns. They account for most of the new word formation, but other multi-word word formation patterns, such as "1 + 2", "2 + 1", and "3 + 1", also account for a large part of the formation patterns of new words. Consequently, effectively identifying multi-word formations of new words is a crucial issue.

2. In the process of word formation, the difficulty of word formation for different numbers of words is different. However, the standard filtering method has a good effect on high-frequency two-word new words, but not on low-frequency multi-word new words.

3. The language used in product reviews tends to be too colloquial and fragmented, which leads to many new words often being located at the sentence breaks, or they are formed separately, which makes the commonly used method of calculating Branch Entropy with external statistical information not very suitable.

Based on the above considerations, this paper proposes a method for discovering new words from product comments using Mutual Information and improved Branch Entropy, which take into account all the characteristics of product reviews. Unlike the traditional statistical neologism recognition method, this method combines three statistical methods—Mutual Information, Support, and Branch Entropy—with rule-based methods such as symbol, stop word, emoji, and dictionary filtering.

Given the characteristics of product reviews, the two methods based on statistical filtering support and Branch Entropy are improved. Firstly, the word expansion algorithm based on Mutual Information is performed on the preprocessed and segmented strings in order to calculate the co-occurrence word frequency with the left and right adjacent words. When the co-occurrence word frequency reaches a threshold, the Mutual Information is calculated. If both the Co-occurrence Frequency and Mutual Information reach the threshold, they will be merged together, and continue to calculate outwards until the judgment conditions are no longer being met. The word expansion procedure is completed, and inputs the expanded words into the potential word set. This method of potential word expansion has improved the extraction efficiency of new words formed by multi-word word formation patterns such as "1 + 2", "2 + 1", "3 + 1". Secondly, an improved support filtering algorithm is proposed to define various filtering thresholds for words of different lengths. The method filters potential words to obtain a candidate new words set To some extent, this solves the problem where the filtering accuracy of the existing new word recognition methods is not high for low-frequency and multi-word new words. Thirdly, an improved method for calculating Branch Entropy is proposed for optimizing the calculations on the left and right branch entropies of the candidate new words set, taking into account that the candidate new words are located at the broken sentence. The method also solves the problem of inaccurate Branch Entropy calculation caused by a part of new words often being located in broken sentences of product comments. Finally, the method uses a dictionary to filter old words and generate a new word set. The proposed method mainly includes two aspects, shown as Figure 1.

- **Obtain a potential word set.** The pre-processing and word segmentation of the corpus are required. The commodity comment corpus is crawled for the pre-process of de-noising and sentence break labelling, and the NLPIR Chinese word segmentation system is used to segment the corpus. A word extension algorithm based on Mutual Information is used to obtain a potential word set. After word segmentation, all word strings are regarded as possible to form a part of the new word, and the left and right word expansion based on the Co-occurrence Frequency and Mutual Information are cycled to obtain the potential word set. This process fully considers the various combinations of new words that may appear, thereby improving the efficiency of new word extraction.

- **Obtaining a new word set.** Candidate new word sets are obtained by means of the improved support filtering algorithm. This algorithm handles the various frequencies of new words with different word lengths in the corpus

and uses different standard filters to obtain candidate new word sets for potential words with different word lengths. Based on the algorithm of improved Branch Entropy filtering for further filtering product reviews tend to be short and fragmented, which inevitably leads to a special condition Some new words are often located at the sentence breaks. We propose an improved method of Branch Entropy to address this particular situation. This filtering process given the characteristics of commodity reviews improves the overall performance of the method for newword recognition.

This paper is organized as follows: In section 2, a brief review of related works is given. The mechanism for acquiring potential word sets is described in Section 3, new word sets acquisition is described in Section 4, the experiments and analysis are given in Section 5, and conclusions are presented in Section 6

2. RELATED WORKS

This section reviews the current works on new word recognition, roughly divided into three methods: statistics-based, rule-based, and a combination of the two.

2.1 Rule-Based New Word Recognition Method

Rule-based new word recognition methods generally statistically summarize word formation rules through word features and use linguistics to analyze new words' possible word formation patterns to build models. Mei L.L. *et al*. [1] extracts the top new words by combining multiple statistical language knowledge. Qin Y. *et al*. [2] uses feature templates to extract local context features. Yan L. *et al.* [3] proposed dynamic features that characterize the similarity of context patterns. Cabrera O. *et al*. [4] analyzed the current state of the art of context models and articulated different phases of a context life cycle to change such services' behavior. Huang M. *et al.* [5] designed statistical measures to quantify the utility of a lexical pattern and measure the possibility of a word being a new word. Zhang M. *et al.* [6] presents an informal word for detecting joint segmentation models' integration. Finnimore P. *et al.* [7] completes single-language and multi-language complex word identification (CWI) through selected features and simple learning models to provide a strong baseline for this field's future development. This kind of method generally has a strong domain and may have higher accuracy in specific fields, but less portable and manual labeling of the corpus is time-consuming and laborious. If low-frequency new words are filtered to ensure accuracy, it will be challenging to identify low-frequency new words.

2.2 Statistics-Based New Word Recognition Method

The new word recognition method reveals new words through various statistical strategies with statistical information. It is generally used in recognition of new words under large-scale corpus. Tang Z. *et al.* [8] proposed SCRFs, a parallel optimization of CRFs based on the Resilient Distributed Datasets (RDD) in the Spark computing framework. He K. *et al.* [9] proposed an anchor word selection method, which linked the co-occurrence probability of words with similar words. Wang Y. W. *et al.* [10] proposed a new simple feature selection method that can effectively filter redundant features. Sun X. *et al.* [11] proposed a new training method, adaptive online gradient descent based on feature frequency information. Using the time distribution information of words and users' attention, Zhu Z. *et al*. [12] proposed an improved TF-IDF algorithm to identify new words in network news. Wei W. *et al*. [13] extracts strongly semantically related compound words according to the conditional co-occurrence in the document, effectively avoiding polysemy. In addition to social media [14] such as Weibo and Post Bar, new word recognition is also widely used in other fields. Li X. *et al*. [15] proposed an improved support vector machine algorithm to detect unknown words in Song poetry. Wang W. *et al*. [16] proposed a framework of recurrent neural networks with morpheme representation for Mongolian named entity recognition. Li W. *et al*. [17] proposed an improved statistical method based on EMI to detect new words in the tourism field [18]. This kind of method has strong adaptability and portability, but it must have a large-scale corpus to ensure training accuracy.

2.3 Statistics-Based and Rule-Based New Word Recognition Method

The method based on the combination of statistics and rules is currently the most common new word statistical method. Zhao L. *et al*. [19] combined the bidirectional LSTM segmentation model with two character-level language models that use a gate mechanism. Chen X. *et al*. [20] proposed a knowledge sharing trans-multi-criteria learning algorithm based on multiple heterogeneous segmentation. Zhang M.Z. *et al*. [21] transformed the problem of new word discovery into calculating the probability of word formation and the annotation of word position. Tian J. *et al.* [22], a convolutional neural network model with part-of-speech tagging and word double embedding is proposed to deal with text multi-classification problem. Chen Z. *et al.* [23] proposed a fault knowledge extraction method based on deep learning. Wang L. *et al.* [24] proposed a document-level optimization method named entity recognition (NER) and used LSTM to classify words. Zhang J. *et al.* [25] introduced the parameter λ during an active learning iteration process to control the number of selected repeating samples to ensure the diversity of the selected samples, and at the same time according to the uncertainty of the word annotation results and the diversity of the context in the samples Sex. This kind of method can complement the previous two methods and give play to their respective advantages.

Based on the existing theoretical basis, we consider the short text, fragmented content, high repetition rate, and other characteristics of the product review corpus. Additionally, the difficulty of word formation is different for the different number of words and new words are often located at sentence breaks.

Table 1 Examples of pre-processing and word segmentation.

Example of pre-processing and word segmentation process	
New words: 奈斯(Nice)	
(1) Sourcetext	节假日快递也非常快，奈斯！ (Holiday delivery is also very fast, Nice!)
(2) Pre-processed text	卍节假日快递也非常快卍奈斯卍
(3) The text after participle	卍/节假日/快递/也/非常/快/卍/奈/斯/卍/

This article proposed a mechanism of improved support and improved Branch Entropy. It advances the recognition effect of multi-word neologisms and neologisms, which are often located at sentence breaks, and dramatically improves the problem of low recognition rate and recall rate of standard new word recognition methods in specific areas product reviews.

3. ACQUISITION OF POTENTIAL WORD SETS

First, we need to preprocess 100,000 product reviews crawling from JD Mall, including the five categories: “food reviews”, “shoe reviews”, “clothes reviews”, “cosmetics reviews”, and “game software reviews”. The processed corpus is adopted in the NLPIR Chinese word segmentation system for word segmentation. A word extension algorithm based on Co-occurrence Frequency and Mutual Information is run with the word strings, while expanded word strings are the input into the potential word set after word segmentation.

3.1 Product Comments Pre-Processing and Segmentation

(1) **Product comments pre-processing:**

The product review text is different from the ordinary text and contains a lot of noise data. In order to improve the accuracy of new word recognition, the text must be preprocessed in advance. The pretreatment work is mainly divided into the two steps.

(i) Delete redundant data. This step deletes stop words, spaces, punctuation marks, special symbols, emoticons, links and other useless symbols

(ii) Mark the sentence breaks. This article treats English letters, spaces, punctuation marks, special symbols, and emoticons as broken sentences, and replaces them with the uncommon word "卍" as a label.

(2) **Chinese word segmentation:**

This paper uses the NLPIR Chinese word segmentation system to segment the corpus of pre-processed product comments. The NLPIR Chinese word segmentation system is a popular word segmentation tool for word segmentation with superior performance in terms of both accuracy and speed.

Algorithm 1 The procedure for product comment corpus pre-processing and word segmentation

Input: Product comment corpus M

Output: A string list, WordList, after word segmentation

1. M1=Split&MarkSymbol(M)
2. M1=Split&MarkAlpha(M)
3. M1=Split&MarkSpace(M)
4. M1=Split&MarkEmoticons(M)
5. Return M1
6. WordList=NLPIR(M1)
7. Return WordList

The specific pre-processing and word segmentation process are shown in Table 1.

The procedures for specific corpus pre-processing and word segmentation are described in Algorithm 1:

Algorithm 1 is comprised of two parts The first part (Steps 1–5) filters out the punctuation marks, spaces, emoticons in the corpus, and marks the deleted position. The second part (Steps 6 and 7) runs the NLPIR Chinese word segmentation system and obtains the word string after word segmentation.

The time complexity of Algorithm 1 can be calculated. Only the crawled original product review text is traversed once, and the time complexity is O(n). The preprocessed text deletes and marks the English letters, spaces, punctuation marks, special symbols, emoticons, and other data not related to the recognized new words. After pre-processing, the time required for the subsequent scanning of the corpus is significantly reduced.

3.2 Obtain Potential Word Sets Through Co-Occurrence Frequency and Mutual Information

In order to ensure the efficiency of extracting new words formed by multi-word word formation patterns such as “1 + 2”, “2 + 1”, “3 + 1”, etc., this paper the extraction of new words in an expanded manner. Firstly, we should determine whether the Mutual Information (between the word string and its adjacent words) reaches a given threshold. If the Mutual Information reaches the threshold, the adjacent words are merged into the word string as a new word string. Then, the merging process is repeated for this new word string with its adjacent words until the Mutual Information less than the threshold. At last, we get the final expanded word string, and export this final word string into the potential word set. Since Mutual Information calculation methods are sensitive to data sparseness, this paper focuses on filtering out low-frequency combinations through

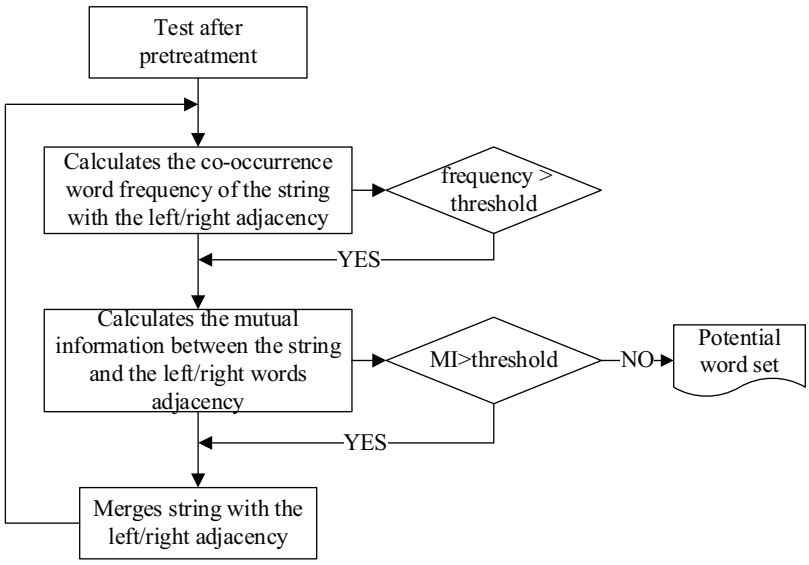

Figure 2 A basic flow to acquire potential word sets.

Co-occurrence Frequency before calculating Mutual Information. The Mutual Information (MI) and Co-occurrence Frequency (CF) are calculated as follows.

(1) Mutual Information (MI):

Mutual Information is used as an internal statistic to calculate the degree of internal combination of new words in the task of new-word recognition. The calculation formula of Mutual Information is:

$$MI(X, Y) = \log \frac{p(x, y)}{p(x)p(y)}, \tag{1}$$

where $p(x, y)$ represents the probability, while x and y appear together in the corpus; meanwhile, $p(x)$ and $p(y)$ represent the probability of x and y appear alone in the corpus, respectively.

(2) Co-occurrence Frequency (CF):

Co-occurrence Frequency refers to the frequency of the word w adjacent to the left (right) adjacent word in the text field.

The flow chart for the extraction of latent words is shown in Figure 2.

The procedure for potential word acquisition is shown as Algorithm 2:

Algorithm 2 A procedure of acquiring potential word sets based on Co-occurrence Frequency and Mutual Information.

Input: A word list, denoted as WordList
Output: A potential word list denoted as CanList1

```
1. for (int i=1; i<=WordList.length; i++) {
2.     CanWord=WordList[i]
3.     for (int j=i-1;j>=i-5;j–){
4.         for (char c : j){
5. if(Freq(CanWord,c)>FTH)&&MI(CanWord, c)>MTH){
6.             CanWord=c+CanWord
7.             }else{
8.             goto step12;
9.           }
10.        }
11.    }
12. for (int j=i+1; j <=i+6; j++) {
13.    for (char c : j){
14.        if(Freq(CanWord,c)>FTH)&&M1(CanWord,
           c)>MTH){
15.          CanWord=CanWord+c
16.        }else{
17.        goto step21;
18.      }
19.    }
20.  }
21. CanList1.add(CanWord)}
22. Return CanList1
```

Algorithm 2 for extracting potential words uses the Co-occurrence Frequency (CF) and Mutual Information (BE) to expand the WordList after word segmentation. If both CF and BE reach the set threshold, the extension condition is satisfied. During the combination of a string and its adjacent words, the judgment continues with the next adjacent word until the string no longer meets the condition. The result is added to the potential word string CanList1. In Algorithm 2, the steps from Step 1 to Step 11 are the expansion of the left adjacent word. First, a loop is set to iteratively obtain the word string "CanWord" from the word list after word segmentation The characters in the *jth* string on the left side of "CanWord" are repeatedly obtained from right to left (at most 5 characters are obtained). And afterwards, we calculate the Co-occurrence Frequency and Mutual Information of the "CanWord" and the adjacent word c on the left, combine *c* into "CanWord" if a threshold is reached, and continue to judge the selection condition on Step 5. If the threshold is not met, the program will jump out of the loop and expand the right adjacent word from Step 12 to Step 20. The method is the same as the extension of the left adjacency. When the right adjacent word's expansion is completed, an add function is applied to output the potential word string CanList1.

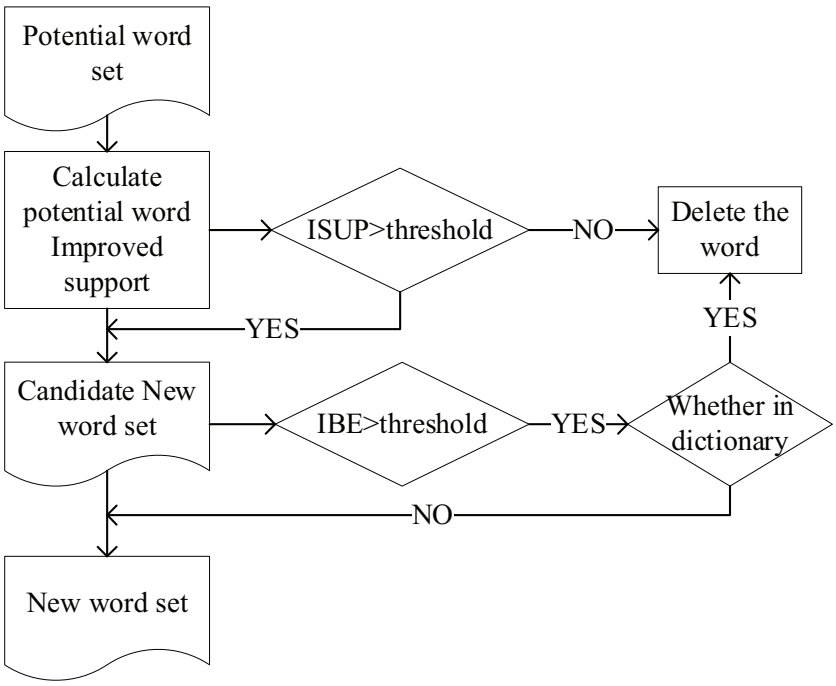

Figure 3 The basic procedure for acquiring a new word set.

We analyze the time complexity of Algorithm 2 as follows:

As Step 1 and 2 traverse the word segmentation WordList to obtain "CanWord", the time complexity is O(n). In Step 3 for getting the adjacent words from the right to left side, the algorithm with the limitation of word length is set to get at most 5 characters for each side.

According to the limitation, the maximum time complexity is O(5*n). Step 4 and 5 run a loop to get *Freq* and *MI*. The loop is only performed up to 5 times, so the complexity is O(25*n). The steps from Step 12 to Step 20 are the same as above. The steps take 5 adjacent characters on the right side, and obtain Freq and MI. This part of algorithm complexity is O(50*n) Consequently, the overall algorithm exhibits a linear-order time complexity.

4. ACQUISITION OF NEW WORD SETS

Since there are many potential words with errors as well as words in the dictionary of the potential word set obtained through Mutual Information expansion, further filtering is required. This paper proposes the filtering method of improved support and improved Branch Entropy to filter potential word sets. The method filters out the old words through the existing dictionary to complete the newword recognition of product reviews. The specific process is shown in Figure 3.

4.1 Potential Word Filtering Based on Improved Support

According to statistics, many new words in commodity reviews, especially in the same kind of commodity reviews, are often used repeatedly. Because of the high repetition rate of new words in the same kind of commodity reviews corpus, this paper proposes a filtering method based on improved support for the coarse-grained filtering of potential words.

According to the rules of word formation in Chinese, we find that the difficulty of word formation is different for a different number of words, so the number of new words with a different number of words in the corpus also varies greatly. Hence, this paper proposed an improved support calculation method which compares the frequency of candidate words in product reviews with the frequency of words of the same length as the candidate words in the dictionary. The method of random selection according to the relevance of the product is adopted to reduce the impact of data sparseness. The annotation set under similar commodities is extracted as a supplement for the calculation of improved support for candidate words. The specific definitions are as follows.

Definition 1 Improved Support (ISUP)

$$ISUP = \frac{P(c1)}{P_n(c)} + \frac{P(c2)}{P_n(c)}, \tag{2}$$

where $P(c1)$ represents the frequency of candidate word c appearing in the comment of the commodity, $P(c2)$ represents the frequency of candidate word c appearing in the comment of 300 items of 10 similar commodities, and $P_n(c)$ represents the frequency of a candidate word with the same length as c appearing in the dictionary.

The support filtering algorithm for potential words is:

Step 1 establishes the support thresholds for the improved support for the length of from 2 to 5 word strings in this article. After many experiments, it was judged that when the support thresholds for strings with the assigned lengths were set as STH(2)=0.45, STH(3)=0.25, STH(4)=0.4, STH(5)=0.15, the F value of the new word recognition algorithm in this paper reaches 70.80 as the highest value. Step 2–5 are focused on filtering the improved support for the candidate word "canWord" in the

Algorithm 3 A new word candidate acquisition algorithm based on improved support filtering.

Input: Potential word string CanList1
Output: Candidate new words string CanList2

```
1. STH(length):STH(2)=0.45,STH(3)=0.25,STH(4)=0.4,
   STH(5)=0.15
2. for (canWord : CanList1) {
3.         if(ISUP(canWord)>STH(canWord.length)){
4. CanList2.add(canWord) }
5. Return CanList2
```

potential word string CanList1, and finally, the candidate new word string CanList2, is returned.

We analyze the time complexity of Algorithm 3 as follows:

By traversing the potential word string CanList1 once, and calculating the support of the potential word "canWord", the algorithm's time complexity is O(n) after counting the word frequency.

4.2 Candidate New Word Filtering Based on Improved Branch Entropy

Two of the most commonly-used methods for calculating the boundaries of new words in existing new word recognition methods are: Branch Entropy (BE) and Adjacency Variety (AV). Branch Entropy, as a kind of external statistic, can be used to calculate the uncertainty of new words and their adjacent words. If the entropy of adjacency is higher, the uncertainty is more significant, the degree of free use is greater, and it can be used with more strings, and the probability of word formation is also greater.

In this paper, the advanced calculation method of Branch Entropy is adapted to the characteristics of product reviews. Because the text of product reviews is colloquial and non-standard, a great deal of fragmented review information appears as disorderly. Many new words are located at the sentence breaks or are used alone. When calculating the Branch Entropy of candidate new words, there is a lack of adjacent words or even no adjacent words, and the calculation method of ordinary Branch Entropy does not specify how the candidate words are calculated at the sentence boundary. Hence, this paper proposes an improved method for calculating Branch Entropy.

In this study, the stop words, spaces, punctuation marks, special symbols, and emojis are regarded as sentence breaks. When one of candidate new words is at the sentence break, the method of using the adjacent word cannot calculate the true degree of freedom of the candidate new word, such as these sentences, "red is very nice", "It is very nice!", and "This shoe is really nice." The candidate's new word "nice" is mostly located at the sentence break, and it will be minimal when calculating its adjacent entropy, which is a severe error. According to statistics, most new words can be formed when candidate new words are located at sentence breaks, and when a new candidate word is often located at a sentence break, the situation of adjacency contains information that can independently form new candidate words. Therefore, we make the following improvements to the Branch Entropy.

Algorithm 4 A new words string acquisition algorithm based on improved support filtering.

Input: Candidate new words string CanList2
Output: new words string CanList4

```
1. for(X:CanList2){
2.     Calculate the Branch Entropy H_L(X) and H_R(X)
3.       if(H_L(X) > HL_TH&&H_R(X) > HR_TH{
4.       CanList3.add(X) } }
5. for(X:CanList3){
6.       if(!OldWordList.contain(X))
7.       CanList4.add(X) } }
8. Return CanList4
```

In the Branch Entropy calculation, the frequency of the candidate new words located at the broken sentence is counted, and the Branch Entropy located at the broken sentence is increased. The specific definitions are:

Definition 2 Improved Branch Entropy (IBE)

$$H_L(X) = \sum_{Xl \in Dl} P(Xl|X) \log P(Xl|X) + w1 * H_Z(X), \quad (3)$$

$$H_R(X) = \sum_{Xr \in Dr} P(Xr|X) \log P(Xr|X) + w2 * H_Z(X), \quad (4)$$

$$H_Z(X) = \frac{N(Z, X)}{N(X)}, \quad (5)$$

where $H_Z(X)$ is the frequency of the candidate new word, X, at the sentence break, $w1$ and $w2$ are the enhancement coefficients of the left and right sentence information entropy respectively, X represents the candidate new word, Dl represents the left adjacent word set of X except X at the sentence break, Dr represents the right adjacent word set of X except X at the sentence break, $P(Xl|X)$ represents the conditional probability of Xl is left adjacent word of X, and $P(Xr|X)$ represents the conditional probability of Xr is right adjacent word of X.

The candidate's external statistics were obtained by calculating the improved Branch Entropy of candidate new words. When the left and right Branch Entropy of candidate new words reached the set threshold, the filter conditions were satisfied. Then the old words were filtered out through existing dictionaries to obtain the new word set. The specific filtering algorithm based on improved Branch Entropy is:

Algorithm 4 is comprised of two parts. The first part (Steps 1–4) calculates the left and right improved Branch Entropy (IBE) of the candidate new word string. When both are greater than the set threshold, pass the filter. The second part (Steps 5–7) removes old words, applies the existing dictionary to filter out old words in the dictionary, and finally obtains the new word set.

We analyze the time complexity of Algorithm 4 as follows:

The steps (Steps 1–4) traverse the candidate's new word string CanList2 and calculate the improved Branch Entropy to obtain CanList3. Because the word frequency information is already available, the time complexity is O(n). In Steps 5–7 we compare and filter CanList3 with the existing dictionary, import the dictionary into the hashset, and use the hashset method to check the existing dictionary with the time complexity of O(1), so the CanList3 is traversed only once. The required time complexity

Table 2 Examples of new words for product reviews.

Some examples of neologisms
坑爹(cheating), 奈斯(nice), 欧凯(ok), 炒鸡(super), 柠檬精(green with envy), 上头(anxious), 男票(boyfriend), 雨女无瓜(none of your business), 杠精(argumentative person), 渣男(playboy), 神马(anything)

Table 3 Newword recognition steps and quantity changes.

Name	Amount
Canlist1	1673
Canlist2	1069
Canlist3	768
Canlist4	432

of the steps is O(n). The overall time complexity of the algorithm is O(2n).

By means of the proposed four algorithms, new-word recognition was carried out with 100,000 reviews in the product review corpus. A total of 432 new words were recognized. Some examples are shown in Table 2.

5. EXPERIMENTS AND RESULTS ANALYSIS

5.1 Experimental Data

Considering the diversity of reviews and users' coverage, we chose JD Mall as the experimental platform. This experiment used the crawler tool to crawl the comment information under JD Mall.

5.2 Evaluation Criteria

In this experiment, the accuracy rate (P), recall rate (R), and F value are adopted as the performance evaluation indicators for the task of discovering new words. The specific calculation formula is:

$$P = \frac{TN}{N} * 100\%, \tag{6}$$

$$R = \frac{TN}{M} * 100\%, \tag{7}$$

$$F = \frac{2 * P * R}{P + R}, \tag{8}$$

where TN represents the number of new words correctly recognized by the algorithm, N represents the total number of new words acquired by the algorithm, and M represents the number of new words in the experimental corpus.

5.3 Experimental Method

To verify the effect of product review new word recognition method based on Mutual Information and improved Branch Entropy on product review newword recognition, this study conducted the experiments using product reviews obtained from JD Mall.

Step 1: Obtain product review corpus. Web crawlers are used to crawl about 100,000 product reviews as experimental data from JD Mall, and to divide them into five categories: "food review", "shoe review", "clothing review", "cosmetics review", and "game software review".

Step 2: Pre-process the experimental corpus. The experimental corpus was used to filter out stop words, English letters, and other noise reduction operations; the NLPIR Chinese word segmentation system was used for word segmentation corresponding to Algorithm 1.

Step 3: Obtain potential word sets. The word expansion algorithm is adopted based on word frequency and Mutual Information to filter and expand the word strings obtained after word segmentation to obtain potential word sets corresponding to Algorithm 2.

Step 4: Obtain new word sets. Filtering the improved latent words with improved support filtering algorithm and filtering algorithm with improved Branch Entropy are proposed to avoid duplication with the existing dictionary to obtain new word sets corresponding to Algorithm 3 and Algorithm 4.

To verify the effectiveness of the new-word recognition algorithm, two additional experiments were conducted for the purpose of comparison.

- **Experiment 2** adopts a newword recognition method without a word segmentation tool. Firstly, the corpus is pre-processed, and then the word expansion is performed on the uncut word corpus from left to right word frequency and Mutual Information calculation to obtain candidate new words. The adjacent entropy and dictionary are filtered to obtain new words.
- **Experiment 3** uses traditional N-Gram-based segmentation of experimental corpus and then recognizes new words by calculating Mutual Information and Branch Entropy.

5.4 Experimental Results and Analysis

The new word recognition results obtained by the above experiments are shown in Table 3 and Table 4. Examples of

Table 4 Examples of extracting new words.

Reviews	New words
Food review	美滋滋，杠杠的，跳楼价，士力架，下饭，爱了，一脸懵逼，给力，炒鸡，抢购，价格，美味…
Shoe review	阿迪王，坑货，男票，马丁靴，炒鸡，踩屎感，上脚，倒闭款，舒服，欧巴，童鞋，颜值，超值…
Clothing review	靓仔，国潮，山寨版，落伍，地摊货，精神小伙，奥力给，硬核，巨划算，质感，面料 …
Cosmetics review	康康，美美哒，亚子，棒棒哒，我枯了，好闻，囤货，抢购，纯白，系列，喜欢，掉色…
Game software review	碉堡了,白给，弟中弟，菜鸟，捡漏，洪荒之力，奶思，卧槽，无情，网游，流弊，画质，枯燥，脱坑,满意，种草，经典，卡带，攻略，游戏粉…

Table 5 Influence table of support threshold parameters.

Topic	Support threshold (STH)	Precision	Recall	F
First	STH(2)=0.4 STH(3)=0.2 STH(4)=0.35 STH(5)=0.1	66.47	60.34	63.26
Second	STH(2)=0.45 STH(3)=0.25 STH(4)=0.4 STH(5)=0.15	72.33	69.34	70.80
Third	STH(2)=0.5 STH(3)=0.3 STH(4)=0.45 STH(5)=0.2	68.20	63.91	65.98

new words obtained from experiments are shown in Table 5, and the comparative experimental results are shown in Table 6

In this experiment, the number is counted during the entire process of newword recognition. First, expand the word by Algorithm 2 based on Co-occurrence Frequency and Mutual Information. The number of possible word strings in the word set after expansion is 1673. Then 1069 candidate new word strings are obtained by Algorithm 3 that improves support filtering. Finally the number of new words is 432 obtained by Algorithm 4 after improved Branch Entropy filtering and dictionary filtering. The results are shown in Table 3.

From the examples in Table 4 showing the extraction of new words, it can be seen that most of the new words are correctly recognized, and some three-character new words and four-character new words that are difficult to recognize using traditional methods are also accurately recognized.

In Step 4, the improved support method proposed in this paper is used to filter the potential word set. This method filters words of different lengths with different thresholds. Table 5 and Figure 4 clearly show the influence of the improved support parameter on the final experiment. The final experiment shows that when STH(2)=0.45, STH(3)=0.25, STH(4)=0.4, STH(5)=0.15 the proposed method works best.

It can be seen from the experimental results in Table 4 that the new word recognition method based on Mutual Information and improved Branch Entropy proposed in this paper has better performance in recognizing new words in product reviews. Compared with results from Experiment 2, the experimental precision, recall rate, and F value have been improved with the proposed method. Compared with results from Experiment 3, although the recall rate is slightly reduced, the precision and F value is significantly improved.

In the experiment the new word recognition method does not use the word segmentation tool, but directly expands the corpus of uncut words from left to right based on MI to obtain new words. Although this method avoids the garbage word string generated by word cutting, there is poor recognition of low-frequency words Although there is a reasonable accuracy rate, the recall rate is low. Experiment 2 tests the traditional new word recognition method (i.e., N-Gram + MI + BE) that uses the N-Gram to perform word segmentation and then makes judgments by the internal formation probability of a word and the external formation probability of a word. However, due to the N-Gram method's low recognition accuracy for a new word made up of multiple words, the overall accuracy is relatively low.

In terms of method, word expansion is conducted using Algorithm 2 to obtain a word set. Then, Algorithm 3 is applied to filter the words based on the improved support method to obtain the candidate new word set, by adjusting the filtering requirements for the potential words of a different number of

Table 6 Experiment results for new word discovery.

Method	Precision	Recall	F-Measure
Proposed method	72.33	69.34	70.80
Experiment 2	70.45	63.79	66.95
Experiment 3	53.22	71.45	61.00

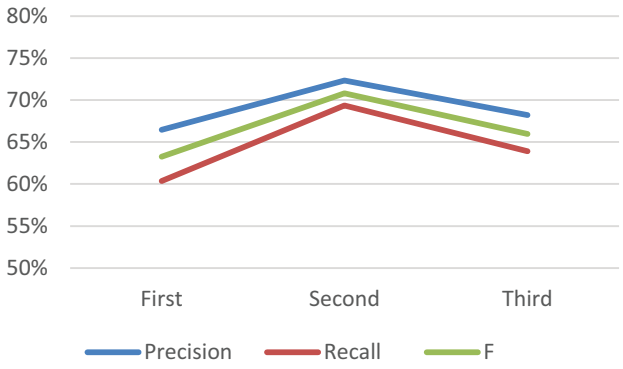

Figure 4 The influence graph of improved support threshold parameter.

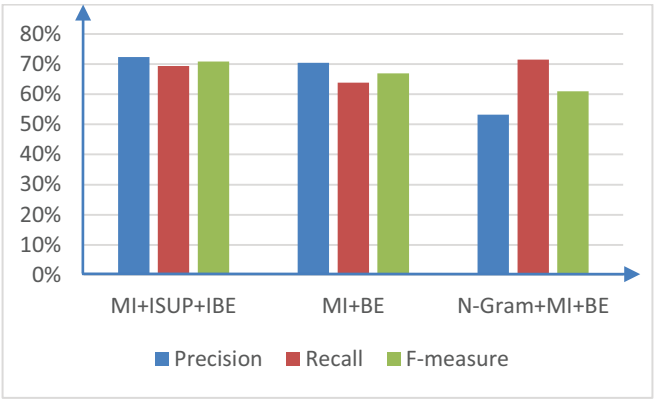

Figure 5 Comparison of experiment results.

words and to reduce the filtering requirements of three-character and five-character words. As can be seen from Figure 4, the recall rate of multi-word new words has been dramatically improved. Finally, the improved support filtering method of Algorithm 4 is used to refine the candidate new word set This method changes the procedure for calculating the Branch Entropy, and increases the candidate new word Branch Entropy according to the frequency of the candidate new word at the sentence breaks. To some extent, it solves the problem of new words often being located in broken sentences, which leads to inaccurate filtering based on adjacent entropy, and further improves the accuracy of new words recognition in this paper. As a whole, the experimental results show that, in terms of product reviews, the new word recognition method proposed in this paper is superior to the other two methods tested for comparison.

6. CONCLUSIONS

Based on the characteristics of product reviews, this paper proposes a method of discovering new words from product comments based on Mutual Information and improved Branch Entropy. By means of the proposed method, new words in product reviews can be extracted quickly and accurately. The acquired new words can be used for research on users' privacy preferences, privacy boundaries, etc. The words can also provide important data for research work on privacy as a basic element of users' privacy protection The main contributions of this paper are as follows:

(1) Combining the frequency and Mutual Information judgment method, the word string after word segmentation is expanded wordbyword to extract possible new words. This method of word expansion has improved the extraction efficiency of new words formed by multi-word word formation patterns such as "1 + 2", "2 + 1", and "3 + 1".

(2) This paper proposes a filtering method based on improved support, which sets different filtering thresholds for words with different word lengths. This method is used to filter words to obtain the candidate set of new words and, to some extent, it solves the problem that the filtering accuracy of the existing new word recognition methods is not high for low-frequency multi-word new words.

(3) This paper proposes a filtering method based on improved Branch Entropy that calculates the left and right Branch

Entropy of the candidate set of new words and takes into account cases where the candidate new words are located in broken sentences. Also, it solves the problem of inaccurate Branch Entropy calculation caused by parts of new words often being located in broken sentences in product comments.

The experimental results show that the proposed mechanisms significantly improve the precision, recall rate, and F value of new word discovery. In the future, the new word recognition of product reviews will provide a valuable basis for researches on user privacy issues such as privacy preference, privacy protection, and information privacy boundaries, to name a few.

Author Contributions: Conceptualization, H.Z.; methodology, H.Z., X.Y., and M.H.; validation, Z.W., G.Z. and M.H.; formal analysis, X.Y. and S.Z.; writing—original draft preparation, H.Z., X.Y. and S.Z.; writing—review and editing, H.Z., Z.W. and M.H.; visualization, H.Z. and G.Z.; supervision, S.Z. and Z.W.; funding acquisition S.Z. and Z.W. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

This research work was supported in part by the National Natural Science Foundation of China (Grant No. 62076006), in part by the 2019 Anhui Provincial Natural Science Foundation Project (Grant No. 1908085MF189), and in part by the 2018 Cultivation Project of Top Talent in Anhui Colleges and Universities (Grant No. gxbjZD15).

Conflict of Interests: The authors declare no conflict of interests.

REFERENCES

1. Barmpatsalou K., Cruz T. J., Simoes P., &Monteiro, E. (2018). Current and future trends in mobile device forensics: a survey. *ACM Computing Surveys*. Vol. 51, No. 3, pp. 1–31.
2. Kuzmanovic M., & Savic G. (2020). Avoiding the Privacy Paradox Using Preference-Based Segmentation: A Conjoint Analysis Approach. *Electronics*. Vol. 9, No. 9, pp. 1382.
3. Lai J., Mu Y., Guo F., Jiang P., & Susilo W. (2018). Privacy-enhanced attribute-based private information retrieval. *Information Sciences*. Vol. 454, pp. 275?291.
4. Gao C.Z., Cheng Q., He P., Susilo W., & Li J. (2018). Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. *Information Sciences*, Vol. 444, pp. 72–88.
5. Zhang S., Yao T., Liang W., Sandor V.K.A., & Li K.C. (2020). An Efficient Privacy-Preserving Multi-keyword Query Scheme in Location Based Services. *IEEE Access*.
6. Zhang K., Zhu Y., Maharjan S., & Zhang Y. (2019). Edge intelligence and blockchain empowered 5G beyond for the industrial Internet of Things. *IEEE Network*. Vol. 33, No. 5, pp. 12–19.
7. Zhang K., Zhu Y., Leng S., He Y., Maharjan S., & Zhang Y. (2019). Deep learning empowered task offloading for mobile edge computing in urban informatics. *IEEE Internet of Things Journal*. Vol. 6, No. 5, pp. 7635–7647.
8. Zhang K., Leng S., Peng X., Pan L., Maharjan S., & Zhang Y. (2018). Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks. *IEEE Internet of Things Journal*. Vol. 6, No. 2, pp. 1987–1997.
9. Mei L.L., Huang H.Y., Wei X.C., & Mao X.L. (2016). A novel unsupervised method for new word extraction. *Science China Information Sciences*. Vol. 59, No. 9, pp. 92–102.
10. Qin Y., Shen G. W., Zhao W. B., Chen Y. P., Yu M., & Jin X. (2019). A network security entity recognition method based on feature template and CNN-BiLSTM-CRF. *Frontiers of Information Technology & Electronic Engineering*. Vol. 20, No. 6, pp. 872–884.
11. Yan L., Bai B., Chen W., & Wu D. O. (2017). New word extraction from Chinese financial documents. *IEEE Signal Processing Letters*. Vol. 24, No. 6, pp. 770–773.
12. Cabrera O., Franch X., & Marco J. (2017). Ontology-based context modeling in service-oriented computing: a systematic mapping. *Data & Knowledge Engineering*, Vol. 110, pp. 24–53.
13. Huang M., Ye B., Wang Y., Chen H., Cheng J., & Zhu X. (2014). New word detection for sentiment analysis. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, MD, USA, June 22–27, 2014, pp. 531–541.
14. Zhang M., Fu G., & Yu N. (2017). Segmenting Chinese Microtext: Joint Informal-Word Detection and Segmentation with Neural Networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI).* Melbourne, Australia, August 19–25, 2017 pp. 4228–4234.
15. Finnimore P., Fritzsch E., King D., Sneyd A., Rehman A. U., Alva-Manchego F., & Vlachos A. (2019). Strong Baselines for Complex Word Identification across Multiple Languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA, June 2–7, 2019, Vol. 1, pp. 970–977.
16. Tang Z., Fu Z.M., Gong Z.R., Li K.L. (2017) A Parallel Conditional Random Fields Model Based on Spark Computing Environment. *Journal of Grid Computing*. Vol. 15, No. 3, pp. 1–20.
17. He K., Wang W., Wang X., & Hopcroft J. E. (2019). A new anchor word selection method for the separable topic discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. Vol. 9, No. 5, pp. e1313.
18. Wang Y. W., & Feng L. Z. (2018). A new feature selection method for handling redundant information in text classification. *Frontiers of Information Technology & Electronic Engineering*. Vol. 19, No. 2, pp. 221–234.
19. Sun, X., Wang, H., & Li, W. (2012) Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* Jeju Island, Korea, July 8–14, 2012, Vol. 1 pp. 253–262.
20. Zhu Z., Liang J., Li D., Yu H., & Liu G. (2019). Hot topic detection based on a refined TF-IDF algorithm. *IEEE Access*, Vol. 7, pp. 26996–27007.
21. Wei W., Guo C.H. (2019). A text semantic topic discovery method based on the conditional co-occurrence degree. *Neurocomputing*. Vol. 368, pp. 11–24.
22. Sarna G, Bhatia M P S. A (2016). A probabilistic approach to automatically extract new words from social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining*. San Francisco, CA, USA, August 18–21, 2016, pp. 719–725.
23. Li X., Wu B., & Zhang B. (2016). Unknown Word Detection in Song Poetry. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, Changsha, China, June 13–16, 2016, pp. 544–549.

24. Wang W, Bao F, & Gao G. (2019) Learning Morpheme Representation for Mongolian Named Entity Recognition. *Neural Processing Letters*, Vol. 50, No. 3, pp. 2647–2664.
25. Li W., Guo K., Shi Y., Zhu L., & Zheng Y. (2017). Improved New Word Detection Method Used in Tourism Field. *Procedia Computer Science*. Vol. 108, No. 8, pp. 1251–1260.
26. Li W., Guo K., Shi Y., Zhu L.Y., Zheng Y.C. (2018) DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. *Knowledge-Based Systems*. Vol. 146, No. 15, pp. 203–214.
27. Zhao L., Zhang Q., Wang P., & Liu X. (2018). Neural Networks Incorporating Unlabeled and Partially-labeled Data for Cross-domain Chinese Word Segmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018),* Stockholm, Sweden, July 13–19, 2018, pp. 4602–4608.
28. Chen X., Shi Z., Qiu X., & Huang X. (2017). Adversarial multi-criteria learning for Chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)* Vancouver, Canada, July 30 August 4, 2017, pp. 1193–1203.
29. Zhang M.Z., Cui X.Q., Chen Y., Liu Y.X., Zhao J.K., Hong O.Y., & Yuan B. (2019). Research on Key Technologies of Customer Consultation Hotspots Mining. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*. March 15–18, 2019, Suzhou, China, pp. 162–166.
30. Tian J., Zhu D., & Long H. (2018). Chinese Short Text Multi-Classification Based on Word and Part-of-Speech Tagging Embedding. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*. December, 2018, Sanya China. No. 62 pp. 1–6.
31. Chen Z., Liu X., Yin Y., & Lu H. (2020). Named Entity Recognition Method for Fault Knowledge based on Deep Learning. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing (ICMLSC 2020)*, Haiphong City, Viet Nam, January 17–19, 2020, pp. 1–4.
32. Wang L., Li S., Yan Q., & Zhou G. (2018). Domain-specific named entity recognition with document-level optimization. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 17, No. 4, pp. 1–15.
33. Zhang J., Huang D.G., Huang K.Y., Liu Z., & Meng X.Z. (2018). λ-active learningbased microblog-oriented Chinese word segmentation. *Journal of Tsinghua University (Science and Technology)*. Vol. 58, No. 3, pp. 260–265.