A Data Processing Method Based on Sequence Labelling and Syntactic Analysis for Extracting New Sentiment Words from Product Reviews

Shunxiang Zhang [1,*], Hanqing Xu [1], Guangli Zhu [1], Xiang Chen, KuanChing Li[2,*]

[1] School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan 232001, China;

[2] Dept of Computer Science and Information Engr. (CSIE), Providence University, Taichung 43301, Taiwan

*Corresponding author

Shunxiang Zhang (e-mail: sxzhang@aust.edu.cn)，KuanChing Li (e-mail: kuancli@pu.edu.tw).

Abstract: New sentiment words in product reviews are valuable resources that are directly close to users. The data processing of new sentiment word extraction can provide information service better for users, and provide theoretical support for the related research of edge computing. Traditional methods for extracting new sentiment words generally ignored the context and syntactic information, which leads to the low accuracy and recall rate in the process of extracting new sentiment words. To tackle the mentioned issue, we proposed a data processing method based on sequence labeling and syntactic analysis for extracting new sentiment words from product reviews. Firstly, the probability that the new word is a sentiment word is calculated through the location rules derived from the sequence labeling result, and the candidate set of new sentiment words is obtained according to the probability. Then, the candidate set of new sentiment words is supplemented with the method of matching appositive words based on edit distance. Finally, the final set of new sentiment words is collected through fine-grained filtering, including the calculation of Point Mutual Information (PMI) and difference coefficient of positive and negative corpus (DC-PNC). The experimental results illustrate the effectiveness of new sentiment words extracted by the proposed method which can obviously improve the accuracy and recall rate of sentiment analysis.

Keywords: product reviews, new sentiment words, sequence labeling, syntactic analysis.

1 Introduction

With the application and development of e-commerce on the Internet, a critical mass of users tend to post product reviews on shopping platforms. Product reviews can provide consumers or companies with a wealth of information, including objective product descriptions, accurate data statistics and product popularity [1-4]. It has great practical value to perform sentiment analysis on product reviews accurately and effectively. As a useful prior knowledge, sentiment words can pave the way for subsequent sentiment analysis. Since the sentiment word is the basic language unit for people to express opinions or attitudes, the extraction of new sentiment words is undoubtedly a crucial field.

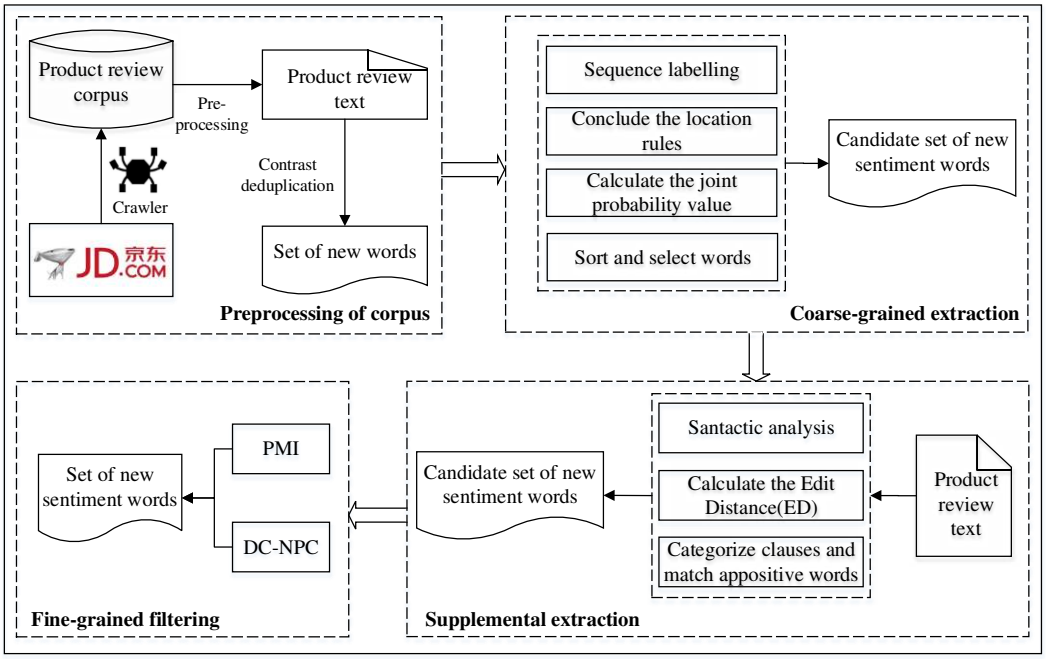
For the previous work on sentiment word extraction [5-6], in (Zhang and Wei, 2018), a method for constructing microblog sentiment dictionary is proposed, and the sentiment analysis of microblog texts is achieved. In (Zhu and Pan, 2020), two coefficients (i.e., microblog importance and time decay) are combined to extract the highlighted words, and the correlation strength between any two highlighted words is measured via the compound co-occurrence rates. Different from the previous work, the process

of extracting new sentiment words is gradually fine-grained. According to the framework of this paper, we can summarize the extraction task as follows: dig out a set of candidate new sentiment words at a coarse-grained level, and then filter out new sentiment words at a fine-grained level. Considering the problem that traditional methods generally ignore the context and syntactic information, a novel approach to extracting new sentiment words from product reviews based on sequence labelling and syntactic analysis is proposed.

At present, to evaluate whether the approach to extracting new sentiment words is effective, the following two aspects need to be considered. On the one side, more new sentiment words are retained in the process of extraction. On the other side, it is ensured that the extracted words have a clear sentiment polarity. For the method in this paper, we retain more new sentiment words in the process of coarse-grained extraction and supplemental extraction, and ensure that the extracted words have the sentiment polarity in fine-grained filtering.

To overcome the above issue, a new data processing method based on sequence labelling and syntactic analysis for extracting new sentiment words from product reviews is proposed. The process of extracting new sentiment words is divided into three main steps: coarse-grained extraction, supplemental extraction and fine-grained extraction. The system framework is shown in Figure 1.

Figure 1. System Framework



- **Coarse-grained extraction.** Two kinds of sequence labeling are performed on the pre-processing corpus to conclude the location rules of old sentiment words. The probability that the new word is a sentiment word is calculated by virtue of the location rules. Then, words that meet the set threshold are selected into the candidate set of new sentiment words.
- **Supplemental extraction.** Syntax trees are generated from product review texts by means of syntactic analysis, and the syntax tree is traversed to generate strings. The edit distance between strings is utilized to measure the similarity of syntactic structure. Furthermore, the candidate new sentiment words are extracted by the method of matching appositive words based on edit distance.
- **Fine-grained filtering.** The final set of new sentiment words is collected by fine-grained filtering, which includes the calculation of point mutual information (PMI) and difference coefficient of

positive and negative corpus (DC-PNC). Then, the sentiment polarity of words are classified into positive and negative respectively.

The main contributions of our work can be summarized as the following three points:

- This paper proposed a data processing method based on sequence labelling and syntactic analysis for extracting new sentiment words, which can detect new sentiment words from product reviews effectively.
- This paper proposed a method of judging the sentiment polarity of words based on PMI and DC-NPC, which can determine the sentiment polarity of candidate words accurately.
- The new sentiment words extracted in this paper are applied to multiple datasets, and good experimental results are obtained, thus verifying the effectiveness of the proposed method.

The rest of this paper is organized as follows. The related works are introduced in Section 2. The specific process of coarse-grained extraction and supplemental extraction of candidate new sentiment words are described in Section 3. The fine-grained filtering of new sentiment words is discussed in Section 4. The experimental design and analysis are explained in Section 5. The conclusion of full text and the outlook for future work are summarized in Section 6.

2 Related works

The goal of extracting new sentiment words is to identify new sentiment units in the process of data processing, so that the subsequent sentiment analysis can be performed more accurately and effectively. In this section, we briefly review the related work from two perspectives, the recognition of new words and the judgment of sentiment polarity.

2.1 New word recognition

Regarding the method of new word recognition, Li et al. proposed a DWWP system and used the combined mutual information technology to solve the user's invention of new words and conversion of sentimental words [7]. Sarna et al. applied probabilistic methods to identify new keywords and assign groups correspondingly, and make decisions based on existing keywords and new keywords extracted [8]. He et al. associated the word co - occurrence probability with the words similarity, and assumed that the most semantically different words are potential candidates for the anchor words [9]. Yan et al. proposed an iterative method to extract new words, through which it was possible to extract distinguishable seed context patterns [10]. Li et al. took new word recognition as a binary classification task and proposed a new effective classification feature including word embedding, activation distance, and statistical conversion probability [11]. Lee et al. regarded mutual information and entropy as a basis for an algorithm and identified unknown words from multilingual code-switching sentences [12]. Yan et al. proposed an iterative scheme to extract new words and introduced dynamic features that characterize the similarity of context patterns [13]. Shan et al. proposed a new word discovery algorithm based on the principle of similarity judgment, combined the similarity and mutual information as an indicator to measure internal integration [14].

2.2 Sentiment polarity judgment

Regarding the method of judging the polarity of sentiment words, Darwich et al. overcomed the inherent problems of dictionary-based generation models, and derived the sentiment polarity of term senses by the context- dual-step aware in-gloss matching [15]. Li et al. performed word embedding based

on a set of seed words and inferred multi-dimensional affective representation of words by a regression-based method automatically [16]. Masiri et al. considered the part-of-speech tags, specified potential terms and employed a comprehensive sentiment lexicon to compute the polarity of the sentences [17]. Wu et al. proposed a new method of merging specific sentiment classifiers in the field of multi-source emotional knowledge training, extracting emotional information from four information sources and fusing them [18]. Deng et al. proposed a novel hierarchical supervision topic model which can capture the sentiment polarity of each word in different topics under the hierarchical supervision [19]. Wu et al. proposed a sentiment classification task of words, and classified the sentiment of words according to the hidden representation of words in sentences [20]. Zhao et al. applied sentiment-oriented point mutual information (SO-PMI) to judge the sentiment polarity of sentiment words and calculated the emotional intensity of sentiment words [21]. Lee et al. utilized association rule mining technology to extract words that have the sentiment polarity [22]. Beigi et al. proposed a noval approach to constructed domain-specific sentiment lexicon, in which the combination of neural network and a sentiment lexicon can adapt word polarities to the target domain without supervision [23]. Deng et al. trained a classifier to predict the sentiment polarity of words, which choosed sentiment-aware word embedding as features [24].

Based on the existing research, in this paper, the process of extracting new sentiment words is regarded as a gradually refined process. Firstly extract candidate new sentiment words at a coarse-grained level, and then filter candidate new sentiment words with fine-grained. It is found that the product review corpus has the following characteristics: (1) the syntactic structure of product review texts are highly similar; (2) new sentiment words often appear around product names, product attributes or four parts-of-speech words (adjectives, adverbs, nouns and verbs). In this paper, we proposed a data processing method based on sequence labeling and syntactic analysis for extracting new sentiment words from product reviews, and also newly defined the concept of DC-PNC to judge the sentiment polarity. The method improved the extraction effect of new sentiment words. To some extent, it solved the problems of unobvious polarity and low accuracy of the extracted sentiment words.

3 Coarse-grained and supplemental extraction of candidate new sentiment words

A product review corpus is constructed by crawling four kinds of product reviews from the JD Mall platform, including "computer reviews", "Laundry detergent reviews", "drawing board reviews" and "tracksuit reviews". After the pre-processing of product review texts, two steps of extracting candidate new sentiment words are conducted, including coarse-grained extraction and supplemental extraction.

3.1 The pre-processing of product review texts

The pre-processing of the raw product review corpus is shown as Algorithm 1.

Algorithm 1: Pre-processing of the raw product review corpus

Input: raw product review corpus T
Output: product review corpus after pre-processing N
1: T1=Remove "garbage" comments;
2: T1=Remove useless characters and special symbols (e.g., '!"#$%&()*+);
3: T1=Remove stop words;
4: N=ICTCLSA(T1);
5: Return N;

(1) Normalized processing:

(i) Removing some "garbage" comments, including texts that are not related to the product and texts containing slogans or improper intent.

(ii) Removing useless characters and special symbols (e.g., '!"#$%&()*+).

(iii) Filtering stop words in combination with the stop word list.

(iv) Correction of typos, conversion of simplified and traditional characters, etc.

(2) Chinese word segmentation: using the word segmentation tool ICTCLSA to segment the product review corpus.

(3) Contrast deduplication: the results of word segmentation are combined with old sentiment words for comparison and deduplication in order to obtain a set of new words.

3.2 Coarse-grained extraction of candidate new sentiment words

Based on a large amount of corpus statistics, this rule can be obtained. If the context of a word is similar to the context of another old sentiment word, the possibility that the word is a sentiment word will increase. So the main idea of this step is to obtain the location rules of old sentiment words firstly, and then take advantage of the location rules to extract new sentiment words. Specifically, that is to count the frequency of old sentiment words appeared around two kinds of labels, and calculate the probability that the new word appeared around two kinds of labels is a sentiment word, then extract words that meet the set threshold as candidate new sentiment words.

There are four main steps in this part of work, which includes sequence labeling, concluding the location rules, calculating the joint probability value, and selecting candidate new sentiment words according to the probability value.

(1) Sequence labeling

Sequence labeling problems [25-27] in natural language processing include word segmentation, part-of-speech (POS) tagging, named entity recognition [28-30], keyword extraction, etc.. As long as a specific label set is given, sequence labeling can be performed. Sequence labeling means that for an input sequence: $X = x_1, x_2, x_3, \mathrm{L}\ , x_i, \mathrm{L}\ , x_n$, x_i in the input sequence X is labeled with a certain label, then the sequence is output: $Y = y_1, y_2, y_3, \mathrm{L}\ , y_i, \mathrm{L}\ , y_n$.

Table 1. A brief description of sequence labeling

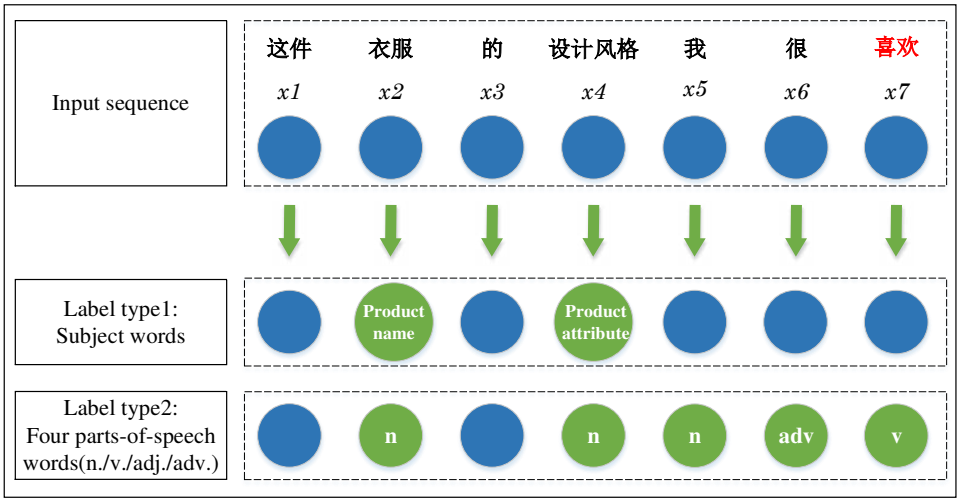
Input sequence	product review text after preprocessing
Label type	①subject words (including product names and product attributes); ②four parts-of-speech words (adjectives, adverbs, nouns and verbs);
Output sequence	text after labeling

According to the experience of language expression and the rules of part of speech collocation, we find that four parts-of-speech words, product names and product attributes often modify sentiment words. Consequently, we choose them as labels of sequence labeling. Here, both of product names and product attributes are collectively called “subject words”. Table 1 explained a brief description of sequence labeling and Figure 2 shows a brief schematic diagram of sequence labeling in this paper.

Based on two types of labels, two kinds of sequence labeling tasks need to be performed, including subject words tagging and part-of-speech tagging. Subject words tagging is to mark specific categories of entities, including product names and product attributes. Here, we manually constructed a collection of subject words for labeling. Part-of-speech tagging [31-32] is to mark the part of speech of each word,

which can also be used to analyze the syntactic structure of a sentence. Harbin Institute of Technology's NLP toolkit is utilized to part-of-speech tagging.

Figure 2. A brief schematic diagram of sequence labeling



(2) Concluding the location rules

According to the output sequence in the previous step, the statistics-based method is utilized to calculate the frequency of the old sentiment words appeared around two kinds of labels respectively (the following "around" means "within 4 characters"). Here, the distribution of the old sentiment words around two kinds of labels is called "location rules". To a certain extent, the location rules of old sentiment words reflect the context in which sentiment words often appear. Therefore, the purpose of this step is to utilize the location rules of old sentiment words to pave the way for mining new sentiment words.

The ratio of the frequency of the old sentiment words appeared around the two kinds of labels and the total number of times they appeared in the corpus is respectively $P(a)$ and $P(b_i)$ .The formula is shown in (1), (2).

$$P(a) = t_a / T \tag{1}$$

$$P(\mathrm{b}_i) = t_{bi} / T \tag{2}$$

Where, t_a and t_{bi} represent the frequency of old sentiment words appeared around subject words and four parts-of-speech words (i=1, 2, 3, 4 represent adjectives, nouns, adverbs and verbs respectively). T is the total number of times that old sentiment words appeared in the raw product review corpus.

(3) Calculating the joint probability of new words

The probability of old sentiment words appeared around subject words and four parts-of-speech words is $P(a)$ and $P(b_i)$ .Therefore, the probability of new words appeared around subject words and four parts-of-speech words being sentiment words is also set as $P(a)$ and $P(b_i)$. Since each new word may appear either around subject words or four parts-of-speech words. So the "weighted summation" strategy is adopted to set the following formula. The formula aimed to calculate the joint probability of new words being sentiment words. The calculation formula of the joint probability is shown in (3).

$$P(word) = w_a \times P(a) + w_b \times \sum_{i=1}^{4} P(b_i) \tag{3}$$

Where, $P(a)$ and $P(b_i)$ represent the probability that the new word is a sentiment word when it appears around the subject words and four parts-of-speech words respectively. w_a and w_b represent the weights of $P(a)$ and $P(b_i)$ respectively, and the formulas are shown in (4) and (5).

$$w_a = \frac{t_a}{t_a + \sum_{i=1}^{4} t_{bi}} \tag{4}$$

$$w_b = \frac{\sum_{i=1}^{4} t_{bi}}{t_a + \sum_{i=1}^{4} t_{bi}} \tag{5}$$

(4) Selecting candidate new sentiment words

The greater the joint probability of a word, the more likely it is to be a new sentiment word. Therefore, the goal of this step is to extract words with higher joint probability. Then, the joint probability calculated in the previous step is compared with the set threshold. If the joint probability value exceeds the set threshold, it is added to the candidate set of new sentiment words. Otherwise, the word is removed.

Algorithm 2 shows the procedure for extracting new sentiment words with coarse-grained according to the results of sequence labeling.

Algorithm 2: A procedure for extracting new sentiment words with coarse-grained

Input: Product corpus after pre-processing N, the set of new words $\{n_i\}$
Output: candidate set of new sentiment words $\{w_i\}$

```
1.  N1= Mark the product name(N);
2.  N1= Mark the product attribute(N);
3.  N2= Mark four parts-of-speech words(N);
4.  N3= Mark old sentiment word(N);
5.  Return N1, N2, N3;
6.  For (each word n∈N) {
7.      If (N3∈ N1+- four-words) {
8.          t_a+=1;}
9.      Else if (N3∈ N2+- four-words) {
10.         t_bi+=1;}
11.     Return t_a, t_bi;
12. }
13. calculate P(a), P(b_i);
14. calculate the weights w_a, w_b;
15. For each word n_1∈{n_i} {
16.     calculate P(n_1);
17.     If P(n_1)> threshold{
18.         add n_1 to {w_i};
19. }
20. Return {w_i};
```

Algorithm 2 is comprised of four parts. The first part (Step 1-5) respectively annotates four parts-of-speech words, product names, product attributes and old sentiment words in the product review corpus after preprocessing. The second part (Step 6-13) is to conclude the location rules of old sentiment words. If an old sentiment word appears within 4 characters of four part-of-speech words, product names or product attributes, increase the corresponding frequency by one. The probability of old sentiment words appeared around two kinds of labels are calculated by formulas. The third part (Step14-16) calculates the

joint probability that the new word appeared around two kinds of labels is a sentiment word, according to the location rules of old sentiment words obtained in the previous step. In the fourth part (Step17-20), add the words that meet the set threshold to the candidate set of new sentiment words.

3.3 Supplemental extraction of candidate new sentiment words

If the syntactic structure of a word is similar to the syntactic structure of another old sentiment word, the possibility of the word being a sentiment word will increase. The syntax tree is a graphical representation of sentence structure, which is helpful to understand the syntactic structure of words. So the main idea of this step is to find candidate new sentiment words with the help of syntax tree and syntactic structure similarity. In this step, we introduced the concept of "appositive words" and proposed the method of matching appositive words based on edit distance to extract candidate new sentiment words.

Definition 1: Appositive words

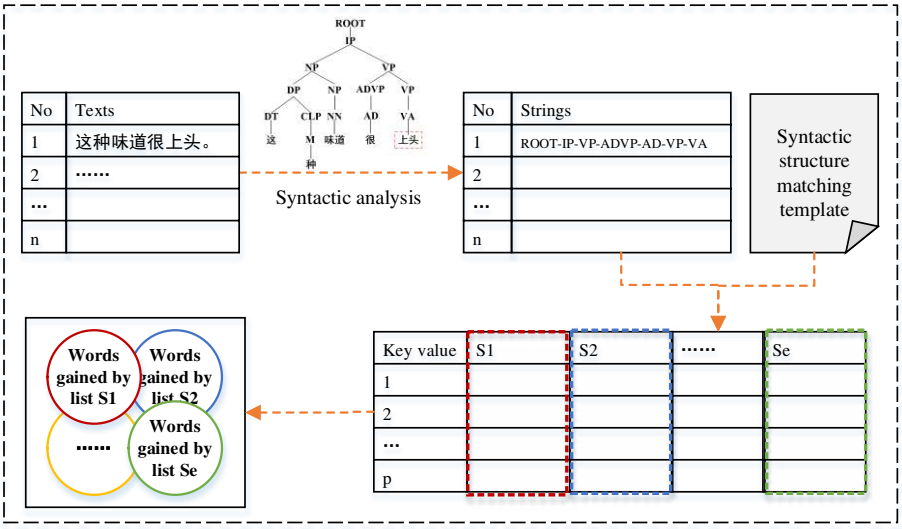
Appositive words are words that occupy the same position component in sentences with similar syntactic structures. They belong to the same category concept, and their meanings are equal and irreplaceable. The appositive word is defined by the formula (6):

$$appositive\_word = \{word \mid word \in N, ED < k\} \tag{6}$$

In formula (6), N is the product review corpus after preprocessing, ED is edit distance between strings, which is utilized to measure the structure similarity, k is the set threshold.

When the ED is smaller than the threshold value k, it is deemed that these sentences have similar syntactic structure, and the word occupied the same position with old sentiment word is considered to be a candidate new sentiment word. For example, "精美(exquisite)" in "这份礼物很精美(this gift is exquisite)" and "上头(a momentary impulse)" in "这种味道很上头(this taste is very high)" are appositive words with each other.

Figure 3. The process of supplemental extraction of new sentiment words



The process of supplemental extraction of new sentiment words is shown in Figure 3. The syntax tree of product reviews is generated based on the technology of syntactic analysis. Syntactic analysis can reflect the semantic modification relationship between sentence components, which can obtain long distance collocation information. The traversal path from the root node to the node of old sentiment word can generate a string, which reflects the syntactic structure component in which it is located. The edit distance (ED) refers to the times of editing operations required to convert two strings from one to the other, which can be utilized to measure the similarity of two strings. Similarly, the edit distance of

traversal path strings can be used to measure the similarity of syntactic structure. The smaller the edit distance, the more similar the syntactic structure. Hence, a new method of matching appositive words based on edit distance is proposed, which can be applied to extract candidate new sentiment words. The specific steps of the method of matching appositive words are as follows.

Step1: Generate a syntax tree. All of the product reviews are split into sentences firstly, which is the basis of constructing a syntax tree. Stanford University's natural processing toolkit Stanza [33] is utilized to perform syntactic analysis on sentences. Then, the structured information of the syntax tree of each sentence is obtained by the software package.

Step2: Establish a syntactic structure matching template. A syntax tree structure table is created as a matching template. The table stores multiple common string representations (e.g. ROOT-IP-VP-ADVP-AD-VP-VA), which is the traversal path of old sentiment words in the syntax tree corresponding to the comment text. The string reflects the syntactic structure component in which the word is located.

Step3: Calculate the edit distance (ED). Traverse the syntax tree of each sentence and generate the subtree. For each subtree generated, the string S is also generated by traversal. Calculate the edit distance (ED) between the string S of the subtree and the existing string in the matching template. When the edit distance is less than the threshold value, it is deemed that the syntactic structure of two strings is similar. Otherwise, the traversal string S of the subtree is added to the matching template.

Step4: Categorize clauses with similar syntactic structure. Build a result table and create multiple new keys in the result table. Below the column corresponding to each key value, a list of clauses with similar syntactic structure is stored.

Step5: Extract candidate new sentiment words. The sentences in the same list of clauses are aligned according to the syntactic structure. The words occupying the same position as the old sentiment words are regarded as candidate new sentiment words.

Step6: Remove old sentiment words. After the above steps are completed, combining with the existing sentiment dictionary, remove the repeated words from the candidate set of new sentiment words.

4 Fine-grained filtering of candidate new sentiment words

The words extracted by the above steps may have no sentiment polarity, so the judgment of sentiment polarity is still required. Hence, we proposed a new method of judging the sentiment polarity of words based on PMI and DC-NPC.

4.1 Point mutual information (PMI)

Point mutual information (PMI) is utilized to calculate the semantic similarity of two words. The larger the value of PMI, the higher the relevance of two words. The calculation formula is shown in (7).

$$PMI(word_1, word_2) = log_2 \frac{P(word_1 \& word_2)}{P(word_1)P(word_2)} \tag{7}$$

Where, $P(word_1 \& word_2)$ represents the probability of two words appeared in a review at the same time, $P(word_1)$ and $P(word_2)$ represent the probability that $word_1$ and $word_2$ appear in reviews separately.

The semantic similarity of candidate new sentiment words and commendatory benchmark words and derogatory benchmark words are calculated respectively, and the sentiment polarity of words can be determined by the difference. The calculation formula is shown in (8).

$$SO\_PMI(word) = \sum_{i=1}^{n} PMI(word, P_{wi}) - \sum_{i=1}^{n} PMI(word, N_{wi}) \tag{8}$$

Where, P_{wi} is a set of commendatory benchmark words, N_{wi} is a set of derogatory benchmark words. $PMI(word, P_{wi})$ represents the semantic similarity of the candidate new sentiment words and commendatory benchmark words, $PMI(word, N_{wi})$ represents the semantic similarity of the candidate new sentiment words and derogatory benchmark words. *SO_PMI(word)* represents the sentiment polarity of the word.

4.2 The difference coefficient of positive and negative corpus (DC-PNC)

In this part, we proposed a new method for judging the sentiment polarity of words. The sentiment polarity of candidate words is determined by the ratio of frequency difference and frequency sum in the positive and negative corpus, which is defined here as the difference coefficient of positive and negative corpus (DC-PNC). If the frequency of a word appeared in the positive corpus is high and this word rarely appears in the negative corpus, then we believe that the sentiment polarity of the word is positive. Otherwise, the opposite is true. The value of DC-NPC ranges from -1 to 1. The closer its absolute value is to 1, the more likely it is to have sentiment polarity. The specific definition is:

Definition 2: the difference coefficient of positive and negative corpus (DC-PNC)

$$DC(word) = \frac{F_{pos}(word) - F_{neg}(word)}{F_{pos}(word) + F_{neg}(word)} \tag{9}$$

Where, $F_{pos}(word)$ and $F_{neg}(word)$ represent the number of times that the word appeared in the positive corpus and the negative corpus respectively.

$$\delta(word) = \begin{cases} 1, & \alpha \le DC(word) < 1 \\ -1, & -1 < DC(word) \le \beta \end{cases} \tag{10}$$

Where, $\delta(word)$ represents the sentiment polarity of candidate new sentiment words. If $\delta(word) = 1$, the candidate new sentiment word is added into the positive set of new sentiment words. If $\delta(word) = -1$, the candidate new sentiment word is added into the negative set of new sentiment words. Otherwise, we believe that the candidate new sentiment word cannot be collected into the final set of new sentiment words.

In the formula (10), there are certain underlying parameters that need to be tuned. In Section 5, the experiment of tuning parameters of α and β was performed with ten groups of parameter values. Since the best performance of 75.6% (F-measure) was obtained when $\alpha = 0.8$, $\beta = -0.8$, this will be our choice for α and β.

Algorithm 3 shows the procedure for filtering new sentiment words based on PMI and DC-NPC.

Algorithm 3: A procedure for filtering new sentiment words based PMI and DC-NPC

Input: candidate set of new sentiment words $\{w_i\}$,
commendatory benchmark vocabulary $\{P_{wi}\}$,
derogatory benchmark vocabulary $\{N_{wi}\}$
Output: set of new sentiment words $\{W_i\}$,
positive set of new sentiment words $\{W_{+i}\}$,
negative set of new sentiment words $\{W_{-i}\}$

```
For each word w1 ∈ {wi}{
    For each word p1 ∈ {Pwi}, n1 ∈ {Nwi}{
        calculate SO_PMI(w1);
        If SO_PMI(w1) >0{
            add w1 to {W+i};}
        Else if SO_PMI(w1) < 0{
            add w1 to {W−i};}
    }
```

9: }
10: For each word $W_{+1} \in \{W_{+i}\}$ {
11: calculate $DC(W_{+1})$;
12: If $DC(W_{+1}) < \alpha$ {
13: remove W_{+1};}
14: }
15: Return $\{W_{+i}\}$;
16: For each word $W_{-1} \in \{W_{-i}\}$ {
17: calculate $DC(W_{-1})$;
18: If $DC(W_{-1}) > \beta$ {
19: remove W_{-1};}
20: }
21: Return $\{W_{-i}\}$;
22: $\{W_i\} = \{W_{+i}\} + \{W_{-i}\}$;
23: Return $\{W_i\}$;

Algorithm 3 is comprised of two parts. In the first part (Step 1-9), the $SO\_PMI(word)$ of the candidate set of new sentiment words were calculated separately. If the $SO\_PMI(word) > 0$, add it into the positive set of new sentiment words $\{W_{+i}\}$. Otherwise, add it into the negative set of new sentiment words $\{W_{-i}\}$. The second part (Step 10-23) is to filter the $\{W_{+i}\}$ and $\{W_{-i}\}$ respectively. Firstly, remove words that $DC(word)$ does not meet the set thresholds. Finally, add the filtered $\{W_{+i}\}$ and $\{W_{-i}\}$ to the set of new sentiment words $\{W_i\}$.

5 Experiment and analysis of results

5.1 Experimental data

There are various e-commerce platforms where users can express comments on a product or service. The JD Mall website is used as the source for crawling product review data. Considering the diversity of products and the coverage of customers, four kinds of product review, including "computer review", "laundry detergent review", "drawing board review", and "tracksuit review", are collected to construct a product review corpus.

In this paper, the product review corpus of 20000 reviews is devided into two disjoint sets: Training Set and Test Set. The Training Set of 4000 reviews is used to extract new sentiment words, and the Test Set of 16000 reviews is used for sentiment classification to verify the effectiveness of new sentiment words extracted by the method in this paper.

In the meantime, product reviews of Training Set and Test Set need to be marked in advance. Generally, sentiment classification consists of three sentiment polarities: positive, negative and neutral. Since comments only contain sentiment words with positive and negative polarity, in this paper, all comment texts are classified into two sentiment polarities: positive and negative. Comments under the "praise review" label are divided into the positive review data set, and comments under the "bad review" label are divided into the negative review data set. In addition, as for comments under the "middle review" label and other reviews, the labeling of sentiment polarity is completed manually. Table 2 and Table 3 show the distribution of Training Set and Test Set respectively.

Table 2. The distribution of Training Set

Name	Category	Positive	Negative	Total

Training Set	Computer review	632	368	1000
	Laundry detergent review	703	297	1000
	Drawing board review	585	415	1000
	Tracksuit review	611	389	1000

Table 3. The distribution of Test Set

Name	Category	Positive	Negative	Total
Test Set	DataSet1: Computer review	2538	1462	4000
	DataSet2: Laundry detergent review	2022	1978	4000
	DataSet3: Drawing board review	2693	1307	4000
	DataSet4: Tracksuit review	2435	1565	4000
	DataSet5: DataSet1+DataSet2+ DataSet3+ DataSet4	9688	6312	16000

5.2 Experimental performance evaluation index

Precision (P), recall (R) and F-measure (F) are utilized as experimental performance evaluation indexes. The formulas are (11), (12) and (13).

$$P = j_t / j_f \tag{11}$$

$$R = j_t / j_s \tag{12}$$

$$F = \frac{2 * P * R}{P + R} \tag{13}$$

Where, j_t represents the number of product reviews of the category judged correctly, j_f represents the number of product reviews judged as the category, and j_s represents the number of product reviews that should be judged as the category. The category includes positive and negative.

5.3 Experimental methods

Product review data provided by JD Mall shopping platform are applied to experiments. The specific steps of the experimental design are as follows:

Step1: Constructing the product review dataset. Product reviews were crawled from JD Mall by the crawler as experimental dataset, including "computer review", "laundry detergent review", "drawing board review", and "tracksuit review".

Step2: Pre-processing the product review corpus. Corresponding to Algorithm 1, the product review corpus is normalized by removing some “garbage” comments, special symbols and stop words; the ICTCLSA Chinese word segmentation tool is utilized for word segmentation.

Step3: Extracting new sentiment words from the Training Set. This step applied the new method proposed in this paper to extract new sentiment words from the Training Set, including coarse-grained extraction, supplemental extraction and fine-grained filtering.

Step4: Performing sentiment classification on the Test Set. This step adopts the sentiment lexicon-based method to perform sentiment classification [34-35], which is to accumulate the weights of sentiment words appeared in the sentence, and determine the inclination of opinions to be positive or negative according to the accumulated value.

Step5: Comparing the experimental results. Based on the results of sentiment classification on the Test Set, a comparison was made with the previous labeled results on the Test Set to measure the accuracy.

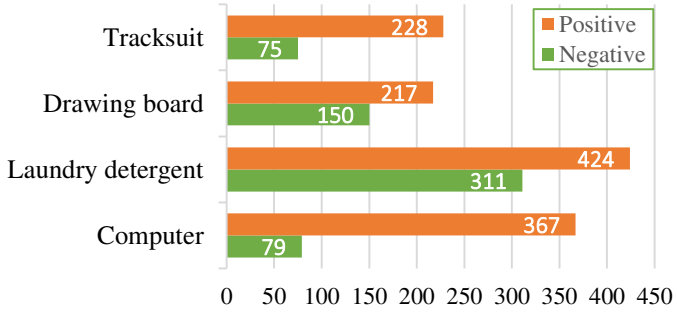
To verify the validity of the method of extracting new sentiment words, we designed two kinds of experiments for comparison. Here, the sentiment dictionary of Dalian University of Technology is abbreviated as DUTIR, and the set of new sentiment words extracted from Step3 of the experiment in this paper is called NSW.

- **Experiment Ⅰ (DUTIR)** makes use of the sentiment dictionary of Dalian University of Technology (DUTIR) to perform sentiment classification on product reviews in the Test Set. The precision, recall and F-measure are calculated respectively.
- **Experiment Ⅱ (DUTIR+NSW)** makes use of the new sentiment dictionary to perform sentiment classification on product reviews in the Test Set, which combines the sentiment dictionary of Dalian University of Technology (DUTIR) with the set of new sentiment words (NSW) extracted by the new method. The precision, recall and F-measure are calculated respectively again.

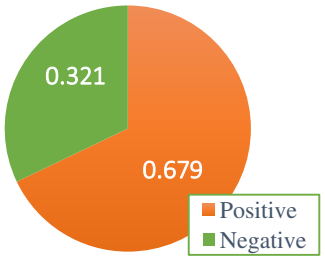
5.4 Experimental results and analysis

The distribution of positive and negative new sentiment words extracted from each product reviews of Training Set is shown in Figure 4(a). A total of 1,851 words were extracted from the four kinds of product reviews, and 1,311 new sentiment words were left after removing the repeated words. The overall proportion of positive and negative words in the set of new sentiment words is shown in Figure 4(b). It can be seen that the number of positive new sentiment words accounted for 0.679 and the number of negative new sentiment words accounted for 0.321.

Figure 4. Polarity distribution of new sentiment words extracted from Training Set



(a)Distribution of positive and negative new sentiment words extracted from each product reviews of Training Set



(b)Overall proportion of positive and negative words in the set of new sentiment words

Table 4. Examples of new sentiment words extracted from Training Set

Reviews	New sentiment words
Computer review	牛批，尚可，颜值，体验感，惊艳，超薄，淘汰，捡漏，下单，爱京东，开机，简约，一流，护眼，野兽，不重，手感好，真香，抢购，流畅，保护，办公，影响，评价，异响，网游，种草，轻薄，保真…
Laundry detergent review	再囤，候用，出彩，中买，出众，倍儿香，洁净，易漂，机洗，好好闻，到位，蓬松，手洗，损坏，褪色，亲民，洗涤，聚划算，上门，焕然一新，回购，拔草，粗糙，温和，去污，除螨，护色…
Drawing board review	没坏，反复，耐造，看不厌，品质，过关，标签，破损，发货，童趣，值得入手，态度好，解答，验证，外观，能用，严实，大牌子，丝滑，关键，撕贴，建议，不粘，美腻，吐槽，点赞，正品…

Tracksuit review	透气，触感，料子，收藏，质感，粘毛，次品，脱线，亲肤，面子，给力，倍儿棒，上身，客服，严实，新潮，怀疑人生，报废，一言难尽，缩水，潮流，耐看，贴身，档次，很有，舒适，关键…

Table 4 lists some examples of new sentiment words extracted from Training Set. As can be seen from Table 4, most of new sentiment words are extracted correctly and most of words express obvious sentiment polarity, including some novel Internet buzzwords. The experiment result illustrates that the method in this paper can effectively extract new sentiment words.

Influence of α and β: In Section 4.2, a new method of judging the sentiment polarity of words is proposed. In formula (10), the judgement of DC-NPC involves the setting of parameters, which refers to α and β. Hence, combining with ten groups of parameters setting, the influence of parameter α and β on experiment can be revealed clearly in Table 5 and Figure 5.

Figure 5. The influence graph of threshold parameters

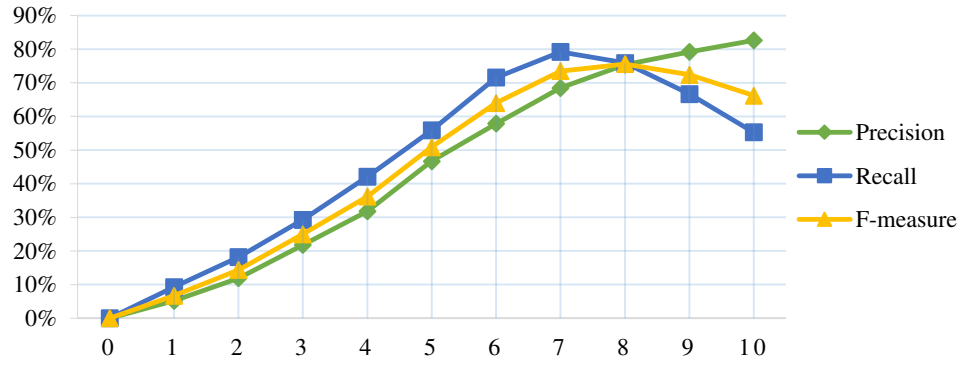


Table 5. The influence table of threshold parameters

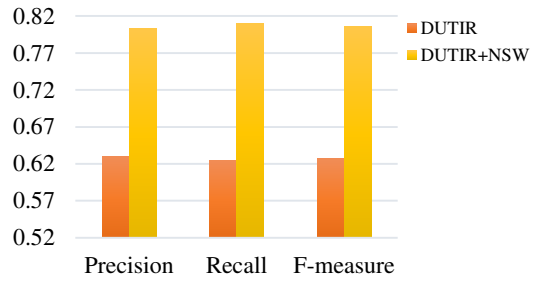
Group number	Parameter value	Precision	Recall	F-measure
First	$\alpha=0.1, \beta=-0.1$	0.052	0.093	0.067
Second	$\alpha=0.2, \beta=-0.2$	0.119	0.182	0.144
Third	$\alpha=0.3, \beta=-0.3$	0.218	0.293	0.250
Fourth	$\alpha=0.4, \beta=-0.4$	0.318	0.421	0.362
Fifth	$\alpha=0.5, \beta=-0.5$	0.467	0.559	0.509
Sixth	$\alpha=0.6, \beta=-0.6$	0.579	0.716	0.640
Seventh	$\alpha=0.7, \beta=-0.7$	0.685	**0.792**	0.735
Eighth	$\alpha=0.8, \beta=-0.8$	0.754	0.759	**0.756**
Ninth	$\alpha=0.9, \beta=-0.9$	0.792	0.667	0.724
Tenth	$\alpha=1.0, \beta=-1.0$	**0.826**	0.553	0.662

Figure 5 shows the precision, recall and F-measure of the experiment with different α and β. Considering the three evaluation indicators, we can see that the best performance of 75.6% (F-measure) was obtained when $\alpha=0.8$, $\beta=-0.8$. When $\alpha=0.7$, $\beta=-0.7$, the recall rate is the best and more new sentiment words can be extracted, but the accuracy is second best. When $\alpha=1.0$, $\beta=-1.0$, the reason of the decline may be that there are too few sentiment words that meet the threshold condition and some words that are originally sentiment words are filtered out, which leads to the poor effect in the application of sentiment classification.

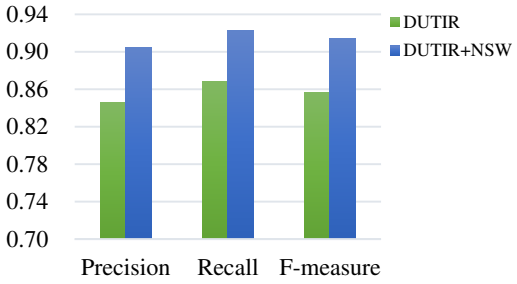
The experimental comparison results on positive comments of DataSet5 are shown in Figure 6(a), and the experimental comparison results on negative comments of DataSet5 are shown in Figure 6(b),

including three experimental performance evaluation indexes: precision, recall and F-measure. Table 6 shows the experimental comparison results obtained by using two methods (DUTIR and DUTIR+NSW) respectively on five datasets.

Figure 6. The experimental comparison results on DataSet5



(a) The experimental comparison results on positive comments of DataSet5



(b) The experimental comparison results on negative comments of DataSet5

Table 6. The experimental comparison results on five datasets

Test Set	Experiment	Positive comments			Negative comments		
		P	R	F	P	R	F
Dataset 1	DUTIR	0.545	0.536	0.540	0.846	0.865	0.855
	DUTIR+NSW	**0.754**	**0.759**	**0.756**	**0.886**	**0.903**	**0.894**
Dataset 2	DUTIR	0.647	0.643	0.645	0.819	0.847	0.833
	DUTIR+NSW	**0.832**	**0.841**	**0.836**	**0.916**	**0.935**	**0.925**
Dataset 3	DUTIR	0.625	0.619	0.622	0.844	0.871	0.857
	DUTIR+NSW	**0.765**	**0.773**	**0.769**	**0.883**	**0.904**	**0.893**
Dataset 4	DUTIR	0.764	0.763	0.763	0.875	0.888	0.881
	DUTIR+NSW	**0.810**	**0.816**	**0.813**	**0.908**	**0.919**	**0.913**
Dataset 5	DUTIR	0.630	0.625	0.627	0.846	0.868	0.857
	DUTIR+NSW	**0.803**	**0.810**	**0.806**	**0.905**	**0.923**	**0.914**

According to the comparison results of two methods in Figure 6 and Table 6, it can be concluded that the method of extracting new sentiment words based on sequence labeling and syntactic analysis has a great performance on the extraction of new sentiment words from product reviews. Specific analysis can be made as follows.

(1) As can be seen from Figure 6(a) and (b), compared with the results of Experiment Ⅰ (DUTIR), the overall effect of experiment Ⅱ (DUTIR+NSW) were improved after adding new sentiment words extracted by the proposed method. For positive comments in dataset5, the precision, recall rate and F-measure were significantly improved. Specifically, precision increased by 17.3%, recall rate increased by 19.5%, and F-measure increased by 17.9%. In view of the negative comments in dataset5, the precision, recall rate and F–measure were slightly increased. Specifically, precision increased by 5.9%, recall rate increased by 5.5%, and F-measure increased by 5.7%. Thus, the validity of new sentiment words extracted by the method is proved.

(2) In terms of recall rate, the candidate word set was obtained with a hybrid of Algorithm 2 and the newly proposed method of matching appositive words based on edit distance. Multiple feature information was taken into account, such as word position, context and syntactic structure, which

is helpful to extract potential new sentiment words more comprehensively. As can be seen from Table 6, the recall rate of the experiment has been improved to some extent.

(3) In terms of accuracy, PMI and DC-NPC in Algorithm 3 are combined to filter the candidate set of words, which aims to get the final set of new sentiment words. The combination solved the issue that the screening precision of existing methods was not high to some extent. As can be seen from Table 6, the accuracy of applying new sentiment words extracted to sentiment classification has been improved.

In general, the experimental results indicate that in the field of product reviews, the data processing method based on sequence labeling and syntactic analysis can extract new sentiment words effectively, and provide effective help for sentiment analysis of product reviews.

6 Conclusions

In regard to the issue that traditional methods for extracting new sentiment words generally ignore the context and syntactic information, a data processing method based on sequence labelling and syntactic analysis for extracting new sentiment words from product reviews is proposed. The method is mainly divided into three stages: coarse-grained extraction, supplemental extraction and fine-grained filtering. The main contributions of this paper can be summarized as three aspects.

(1) Subject words tagging and part-of-speech tagging are combined to label old sentiment words, which can capture the location rules of old sentiment words, and the location rules can be used to extract new sentiment words. This step considered the context of sentiment words, which can extract candidate new sentiment words at a coarse-grained level more accurately.

(2) This paper proposed a method of matching appositive words based on edit distance, which mainly uses the similarity of syntactic structure to extract more new sentiment words. This method solved the problem of ignoring structured syntactic information in traditional methods. Meanwhile, the scale of the set of new sentiment words has also been expanded.

(3) This paper introduced the new concept of the difference coefficient of positive and negative corpus (DC-PNC) to judge the sentiment polarity of words. To a certain extent, the combination of PMI and DC-PNC improved the screening precision of new sentiment words.

In the later work, an unsupervised approach will be attempted to realize automatic recognition of new sentiment words, which aims to improve the effect of recognizing new sentiment words. As a related research work of edge computing, the extraction of new sentiment words based on product reviews will effectively utilize data resources to provide information service for users in the future, so as to maximize the value of data.

Conflict of Interest:

Author Shunxiang Zhang declares that he has no conflict of interest.
Author Hanqing Xu declares that she has no conflict of interest.
Author Guangli Zhu declares that she has no conflict of interest.
Author Xiang Chen declares that he has no conflict of interest.
Author Kuan-Ching Li declares that he has no conflict of interest.

Also this manuscript is approved by all authors for publication. I (Shunxiang Zhang) would like to declare on behalf of all co-authors that the work described was original research that has not been published previously. All the authors listed have approved the manuscript that is enclosed.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent: No humans or any individual participants are involved in this study.

References

[1] Zhao W, Guan Z, Chen L, He X, Cai D, Wang B, Wang Q. (2018) Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*. Vol.30, No.1, pp.185-197.

[2] S. N. Singh, T. Sarraf. (2020) Sentiment Analysis of a Product based on User Reviews using Random Forests Algorithm. *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, April 9, 2020, pp.112-116.

[3] Bi J, Liu Y, Fan Z. (2019) Representing sentiment analysis results of online reviews using interval type-2 fuzzy numbers and its application to product ranking. *Information Sciences*. Vol.504, pp.293-307.

[4] Pankaj, P. Pandey, Muskan, N. Soni. (2019) Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, February, 2019, pp.320-322.

[5] Zhang S, Wei Z, Wang Y. (2018) Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future generation computer systems.* Vol.81, pp.395-403.

[6] Zhu G, Pan Z, Wang Q, Zhang S, Li K. (2020) Building multi-subtopic Bi-level network for micro-blog hot topic based on feature Co-Occurrence and semantic community division. *Journal of Network and Computer Applications*. Vol.170, pp.102815.

[7] Li W, Guo K, Shi Y, Zhu L, Zheng Y. (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.

[8] Sarna G, Bhatia M P S. (2016) A probalistic approach to automatically extract new words from social media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, San Francisco, CA, USA, August 18-21, 2016, pp.719-725.

[9] He K, Wang W, Wang X, Hopcroft JE. (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.1313-1318.

[10] Yan L, Bai B, Chen W, Wu D. (2017) New Word Extraction from Chinese Financial Documents. *IEEE Signal Processing Letters*. Vol.24, No.6, pp.770-773.

[11] Li X, Wu B, Zhang B. (2016) Unknown Word Detection in Song Poetry. *IEEE International Conference on Data Science in Cyberspace(DSC)*, Changsha, China, June 13-16, 2016, pp.544-549.

[12] Lee CW, Wu YL, Yu LC. (2019) Combining Mutual Information and Entropy for Unknown Word Extraction from Multilingual Code-Switching Sentences. *Journal of Information Science and Engineering*. Vol.35, No.3, pp.597-610.

[13] Yan LW, Bai B, Chen W, Wu DO. (2017) New Word Extraction from Chinese Financial Documents. *IEEE Signal Processing Letters*. Vol.24, No.6, pp.770-773.

[14] Shang G. (2019) Research on Chinese New Word Discovery Algorithm Based on Mutual Information. *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2019),* New York, USA, December 2019, pp.580-584.

[15] Darwich M, Noah S.A.M, Omar N. (2020) Deriving the sentiment polarity of term senses using dual-step context-aware in-gloss matching. *Information Processing & Management*. Vol.57, No.6, pp.102273.

[16] Li M, Lu Q, Long Y, Gui L. (2017) Inferring Affective Meanings of Words from Word Embedding. *IEEE Transactions on Affective Computing*. Vol.8, No.4, pp.443-456.

[17] Basiri M.E., Abdar M, Kabiri A, Nemati S, Zhou X, Allahbakhshi F. (2020) Improving Sentiment Polarity Detection Through Target Identification. *IEEE Transactions on Computational Social Systems*. Vol. 7, No.1, pp.113-128.

[18] Wu F, Huang Y, Yuan Z. (2017) Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources. *Information Fusion*. Vol.35, pp.26-37.

[19] Deng D, Jing L, Yu J, Sun S, Michael K. Ng. (2019) Sentiment Lexicon Construction with Hierarchical Supervision Topic Model. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP).* Vol.27, No.4, pp.704-718.

[20] Wu C, Wu F, Liu J, Huang Y, Xie X. (2019) Sentiment Lexicon Enhanced Neural Sentiment Classification. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)*, Beijing, China, ACM, November 3, 2019, pp.1091-1100.

[21] Zhao M, Zhang T, Chai J. (2016) Based on SO-PMI Algorithm to Discriminate Sentimental Words' Polarity in TV Programs' Subjective Evaluation. *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, May 12, 2016, pp.38-40.

[22] Lee Y, Park S, Yu K, Kim J. (2018) Building Place-Specific Sentiment Lexicon. *Proceedings of the 2nd International Conference on Digital Signal Processing (ICDSP 2018). Association for Computing Machinery,* Tokyo, Japan, ACM, February 25-27, 2018, pp.147–150.

[23] Beigi O M, Moattar M H. (2020) Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. *Knowledge-Based Systems*, Vol.213, pp.106423.

[24] Deng D, Jing L, Yu J, Sun S. (2019) Sparse Self-Attention LSTM for Sentiment Lexicon Construction. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP).* Vol.27, No.11, pp.1777-1790.

[25] Sun X, Sun S, Yin M, Yang H. (2020) Hybrid neural conditional random fields for multi-view sequence labeling. *Knowledge-Based Systems*. Vol.189, pp.105151.

[26] Sun X, Ma S, Zhang Y, Ren X. (2019) Towards easier and faster sequence labeling for natural language processing: A search-based probabilistic online learning framework (SAPO). *Information Sciences*. Vol.478, pp.303-317.

[27] Lin C W, Shao Y, Zhang J, Yun U. (2020) Enhanced Sequence Labeling Based on Latent Variable Conditional Random Fields. *Neurocomputing*. Vol.403, pp.431-440.

[28] Chen Z, Liu X, Yin Y, Lu H. (2020) Named Entity Recognition Method for Fault Knowledge based on Deep Learning. *Proceedings of the 4th International Conference on Machine Learning and Soft Computing (ICMLSC 2020)*, Haiphong City, Viet Nam, ACM, January 17-19, 2020, pp.1-4.

[29] Wang L, Li S, Yan Q, Zhou G. (2018) Domain-specific Named Entity Recognition with Document-Level Optimization. *ACM Transactions on Asian & Low Resource Language Information Processing*. Vol.17, No.4, pp.1-15.

[30] 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.

[31] Zhou D, Zhang Z, Zhang M, He Y. (2018) Weakly Supervised POS Tagging without Disambiguation. *ACM Transactions on Asian & Low Resource Language Information Processing*. Vol.17, No.4, pp.1-19.

[32] Pota M, Marulli F, Esposito M, Pietro G D, Fujita H. (2019) Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched Word Embeddings. *Knowledge-Based Systems*. Vol.164, pp.309-323.

[33] Peng Q, Zhang Y, Zhang Y, Jason B, Christopher D. M. (2020) Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020)*, Online, July 5-10, 2020, pp.101-108.

[34] A. S. Manek, P. D. Shenoy, M. C. Mohan. (2017) Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *Word Wide Web*. Vol.20, No.2, pp.135-154.

[35] Lu K, Wu J. (2019) Sentiment Analysis of Film Review Texts Based on Sentiment Dictionary and SVM. *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence (ICIAI 2019)*, Suzhou, China, ACM, March 15, 2019, pp.73-77.