**Academic Journal of Science and Technology**
ISSN: 2771-3032 | Vol. 7, No. 3, 2023

# Weibo New Word Recognition Based on the Fusion of Word Frequency Features and Adjacent Changes

**Yuanfang Xu**

Inner Mongolia Normal University, Hohhot, China

**Abstract:** Weibo as an informal communication and information dissemination tool, has constantly emerged with a large number of new words. Identifying new words in Weibo is of great significance for in-depth understanding and research on social hotspots and language changes. This article proposes a method for integrating word frequency features and adjacency changes to recognize new words in Weibo text. The experimental results show that this method has achieved good results in Weibo new word recognition tasks.

**Keywords:** Weibo new word recognition; Word frequency characteristics; Number of adjacency changes; language change.

## 1. Introduction

Weibo, as an informal communication and information dissemination tool, has a large number of short texts emerging on its platform [1]. Weibo text not only contains a large amount of news, entertainment gossip, and personal opinions, but also reflects the latest trends in social hotspots and language changes. Therefore, identifying new words in Weibo text is of great significance for in-depth understanding of social hotspots and language changes [2].

Traditional word recognition methods mainly use statistical features such as word frequency, word length, information entropy, etc. to identify new words with lower frequency of occurrence. However, relying solely on statistical features makes it difficult to accurately identify new words in Weibo texts, as the content and expression of Weibo texts are relatively informal and often contain a large number of unfinished words, pinyin words, and internet buzzwords. Therefore, it is necessary to consider other features to help identify new words.

This article proposes a method for integrating word frequency features and adjacency changes to recognize new words in Weibo text [3]. This method first identifies candidate new words with lower frequency by counting word frequency features, and then uses adjacency changes to further filter out the true new words. Adjacency changes refer to the changes in the adjacency of a word over different time periods [4]. By observing the number of adjacency changes, one can determine whether the word has novelty and uniqueness [5].

## 2. Related Work

### 2.1. New words identification

New word recognition refers to identifying new words with lower frequency in each text set [6]. Traditional new word recognition methods mainly rely on word frequency features, word length features, and information entropy features to identify new words. Word frequency feature refers to the frequency at which a word appears in a text, word length feature refers to the length of a word, and information entropy feature refers to the amount of information contained in a word. However, these methods cannot accurately identify new words in Weibo text, as the content and expression of Weibo text are relatively informal and often contain a large number of unfinished words, pinyin words, and internet buzzwords.

### 2.2. Number of adjacency changes

Adjacent changes refer to the changes in adjacency of a word over different time periods. By counting the number of adjacency changes, one can discover whether a word has novelty and uniqueness. In text, the adjacency changes of a word can be a change in part of speech, a change in word meaning, or a change in word combination. The number of adjacency changes can be obtained by calculating the co-occurrence of words within two adjacent time periods, and the specific method can be based on the co-occurrence matrix.

## 3. Method

The method proposed in this article for integrating word frequency features and adjacency changes includes the following steps: firstly, using statistical features to identify candidate new words with lower frequency in the text; Then, based on word frequency characteristics and adjacency changes, candidate new words are further filtered to obtain the final set of new words.

### 3.1. Statistic

Word frequency feature refers to the frequency at which a word appears in a text. When calculating word frequency features, it is necessary to preprocess the text, including word segmentation, removing stop words, and removing punctuation marks. Then, the number of occurrences of each word in the text can be counted and sorted according to the number of occurrences, in order to identify candidate new words with lower frequency of occurrence.

### 3.2. Adjacency variation

Adjacent changes refer to the changes in adjacency of a word over different time periods [7]. When calculating the number of adjacency changes, it is necessary to first divide the text into time periods, and then use the word frequency matrix to calculate the co-occurrence matrix [8]. The co-occurrence matrix can represent the degree of correlation between words, and observing the co-occurrence matrix can determine whether a word has novelty and uniqueness. The

specific method can use the word bag model and cosine similarity to calculate the co-occurrence matrix.

### 3.3. Weibo New Word Recognition Method Process

The method in this article combines information from two aspects: word frequency feature and adjacency change number. By using word frequency feature, candidate new words with lower occurrence frequency can be selected. By using adjacency change number, the novelty and uniqueness of words can be further judged, thereby improving the accuracy of new word recognition. The Weibo new word recognition method that integrates word frequency features and adjacency changes includes the following steps:

(1) Data preprocessing: Preprocess Weibo text, including removing special characters, punctuation marks, emoticons, etc., to ensure text purification and consistency.

(2)Word segmentation: Use word segmentation tools to segment the pre processed Weibo text into word sequences.

(3)Statistical Word Frequency Features: Based on the word sequence obtained from word segmentation, count the number of occurrences of each word in Weibo text to obtain word frequency features. This feature can reflect the importance and universality of a word in Weibo text.

(4)Screening candidate Weibo new words: Based on the set threshold or rules, select candidate new words with lower word frequency from the word frequency features. Common screening methods include setting a threshold and retaining only words with word frequencies below this threshold as candidate Weibo new words.

(5)Calculate adjacency changes: Divide Weibo text into adjacent time periods using a sliding window. Then, calculate the number of adjacent changes for each word within adjacent time periods. The number of adjacent changes can be measured by comparing the frequency differences of words in adjacent time periods, and words with significant frequency differences may have novelty.

(6)Filter to obtain Weibo new words: Taking into account word frequency characteristics and adjacency changes, filter candidate new words to obtain the final Weibo new word recognition result. A common filtering method is to set word frequency threshold and adjacency change threshold, and only retain words with lower word frequency and greater adjacency change as Weibo new words.

## 4. Experimental Results and Analysis

### 4.1. Experimentation

To verify the effectiveness of the method of integrating word frequency features and adjacency changes in Weibo new word recognition tasks, we used a dataset containing a large amount of Weibo text. Firstly, preprocess the dataset, including word segmentation, removing stop words, and removing punctuation marks. Then, use statistical features to identify candidate new words with lower frequency of occurrence. Next, calculate word frequency features and adjacency changes, and filter to obtain the final set of new words. Finally, the accuracy of the recognition results is evaluated through manual annotation.

### 4.2. Experimental result

As shown in Table 1, some Weibo new words identified using this method are listed below. The word frequency feature represents the number of occurrences of each new word in Weibo text, while the number of adjacency changes represents the degree of adjacency changes of new words in different time periods. It can be observed that the frequency characteristics of these new words are relatively low, indicating that they appear relatively infrequently on Weibo. The number of adjacency changes can be used to determine the uniqueness and novelty of a new word, and a higher value indicates that the new word is more prominent in terms of adjacency changes.

**Table 1.** Weibo New Word Recognition Results Based on the Fusion of Word Frequency Features and Adjacent Changes

| Numble | Recognized Weibo New Words | Word frequency characteristics (number of occurrences) | Adjacency variation | Manual review results | Judgment results |
|---|---|---|---|---|---|
| 1 | Super Like | 314 | 0.68 | Yes | correct |
| 2 | Laugh and Cry | 189 | 0.56 | No | error |
| 3 | Strike Call | 122 | 0.42 | No | error |
| 4 | Moving bricks | 112 | 0.42 | Yes | correct |
| 5 | Boomerang Kids | 87 | 0.31 | Yes | correct |

For example, in the experimental results, "Super Like" is a new word that frequently appears on Weibo, appearing 314 times. At the same time, its adjacency changes are relatively high, at 0.68, indicating that it has significant adjacency changes in different time periods. Another example is "Boomerang Kids", which although it appears 87 times with a low word frequency, has a relatively high number of adjacency changes of 0.31. This may mean that this term is a recent emerging social phenomenon or concept.

Through calculation, the accuracy of the recognition results is 0.6, the recall rate is 1.0, and the F-value is 0.75. This indicates that when recognizing new words, the algorithm can accurately identify a portion of them, but there are also some cases of misidentification.

The experimental results show that the method of integrating word frequency features and adjacent changes has achieved good results in Weibo new word recognition tasks. Compared to traditional word frequency feature methods, this method can more accurately identify new words in Weibo text and has a higher accuracy and recall rate. Therefore, this method can serve as an effective tool for in-depth understanding and research of social hotspots and language changes.

# 5. Summary

This article proposes a method for integrating word frequency features and adjacency changes to recognize new words in Weibo text. The experimental results show that this method has achieved good results in Weibo new word recognition tasks. Integrating word frequency features and adjacency changes can more accurately identify new words in Weibo text, and has a high accuracy and recall rate. Therefore, this method can serve as an effective tool for in-depth understanding and research of social hotspots and language changes.

Although this method has achieved good results in Weibo new word recognition tasks, there are still some limitations. For example, this method relies on statistical features and text preprocessing, and may not accurately recognize new words in Weibo texts that contain a large number of unfinished words, pinyin words, and internet buzzwords. Therefore, future research can consider other features, such as contextual features and social network features, to improve the accuracy and robustness of new word recognition.

# Acknowledgment

Fund projects: Research Project of Inner Mongolia Higher Education Institutions (NJZY21549)

# References

[1] Su Ning. Based on word features and search engine for Chinese new word identification [J], Journal of Wuhan University, 2010, volume fifty-sixth, issue sixth, 704-710.

[2] Li Chengcheng,Xu Yuanfang, Based on support vector and word features new word discovery research, proceedings of 2012 IEEE International Conference on Computer Science and Automation Engineering ,2012,166-168.

[3] Feng Yong, Li Hua. Based on Adaptive Chinese word segmentation and approximation of SVM text classification algorithm [J], computer science, volume thirty-seventh, 2010, first, 251-254, 293.

[4] Huang Xiuli, Wang Yu.SVM in unbalanced data set [J], computer technology and development, 2009, volume nineteenth, issue sixth, 190-193.

[5] Jian-Yun Nie, Unknown Word Detection and Segmentation of Chinese using Statistical and heuristic Knowledge. Communications of COLIPS,2008,5(I&2),47-57.

[6] Han Xiulong. Research on Weibo New Word Discovery Based on SVM and Feature Correlation [J], Computer Knowledge and Technology, 2018,14,66-69.

[7] Qian Qiuyin, Zhang Zhenglan. A method based on multiple SVM classification method of relevance feedback image retrieval [J], computer technology and development, 2009, volume nineteenth, issue eighth, 66-69.

[8] Fu Lina, Xiao He, Ji Donghong. New Emotional Word Recognition Based on OC-SVM [J], Computer Application Research, 2015,71946-1048.