Sentiment Analysis of Chinese Words Using Word Embedding and Sentiment Morpheme Matching

Jianwei Niu(✉), Mingsheng Sun, and Shasha Mo

State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University,
Beijing 100191, China
niujianwei@buaa.edu.cn

Abstract. Sentiment analysis has become significantly important with the increasing demand of Natural Language Processing (NLP). A novel Chinese Sentiment Words Polarity (CSWP) analyzing method, which is based on sentiment morpheme matching method and word embedding method, is proposed in this paper. In the CSWP, the sentiment morpheme matching method is creatively combined with existing word embedding method, it not only successfully retained the advantages of flexibility and timeliness of the unsupervised methods, but also improved the performance of the original word embedding method. Firstly, the CSWP uses word embedding method to calculate the polarity score for candidate sentiment words, then the sentiment morpheme matching method is applied to make further analysis for the polarity of words. Finally, to deal with the low recognition ratio in the sentiment morpheme matching method, a synonym expanding step is added into the morpheme matching method, which can significantly improve the recognition ratio of the sentiment morpheme matching method. The performance of CSWP is evaluated through extensive experiments on 20000 users' comments. Experimental results show that the proposed CSWP method has achieved a desirable performance when compared with other two baseline methods.

Keywords: Sentiment polarity analysis · Word formation rule
Sentiment morpheme matching · Word embedding
Synonym expanding

1 Introduction

Recent years has witnessed the rapid development of the Internet. With the explosive growth in the online comment number, the value of massive information has emerged gradually. Mining the valuable information from comments often relies on extracting the sentimental polarity of texts. In much previous work, the sentiment analysis is done at document level. However, the sentiment

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 3–12, 2018.
https://doi.org/10.1007/978-3-030-00916-8_1

analyzing of comments often requires sentence-level or even word-level sentiment analysis. What's more, the word-level sentiment analysis is the basics of all other sentiment analyzing works.

One promising sentiment analysis method is to combine the manual tagged corpora with the supervised learning method. In this sentiment analysis method, a supervised learning model will be trained with tagged dataset, and then use the trained model to execute analyzing work. For training the machine learning models, [1] manually collected and tagged 177 negative tweets and 182 positive tweets and, [2] manually tagged 900 sentences and 2519 words. As can be seen that before the analysis work, lots of corpora must be manually collected and tagged for training the supervised machine learning models. However, collecting and tagging the massive corpora is time costly and boring. Additionally, the training set in [1] is collected in period time, and the training set in [2] only content limited sentiences and words. The limited size will lead these datasets cannot include enough sentiment expression form, thereby leading the bad adaptive capacity in different situations.

The other promising sentiment analysis method is to use massive untagged corpora to train unsupervised method. A typical practice of these methods is to use the statistical methods on massive corpora to calculate the probability of a word's sentiment polarity. [3] exploited the PMI to calculate the polarity score for the candidate sentiment words. However, the PMI is sensitive to frequency. It means that two low frequency words can also have a high PMI value. Additionally, [4] utilized word embedding method to convert words into word vectors and use word vectors to calculate the similarity between candidate words and sentiment seed sets. Word2vec tool can generate the real vectors for words according to the context of them. However, two words which have opposite sentiment polarity also can have similar context in sometimes.

In this paper, A novel Chinese Sentiment Words Polarity (CSWP) analyzing method is proposed. The design of the CSWP is to exploit the word embedding method and the sentiment morpheme matching method. The CSWP retains the advantages about flexibility and timeliness of the unsupervised methods. These features can lead to better practicality in real life. The performance of CSWP is evaluated through extensive experiments on 20000 online comments. The experimental results showed that the proposed CSWP method achieved a desirable performance compared with other methods.

The rest paper is organized as follows. Section 2 reviews related research about word-level sentiment analysis. Section 3 overviews the architecture of CSWP and describes the implement of CSWP in detail. Section 4 reports a series of experiments to evaluate the performance of CSWP, and the results are shown in the end of this section. Finally, conclusions are driven in Sect. 5.

2 Related Work

This section will review the related work in the field of word-level sentiment analysis. Up to now, there are two kinds of methods have been proposed for sentiment words analysis: the supervised based method and the unsupervised based method.

The methods based on supervised method always use some trained mathematical models and some manual tagged corpora to predict the sentiment polarity of words. For example, [5] considered the polarity judging question as a binary classification problem, and used SVM models to analyze the sentiment polarity. Moreover, [6,7] used conjunctions as the training features of machine learning models to judge the sentiment polarity of sentiment words. In addition, morphology information and syntactic information also have been exploited in sentiment polarity analysis [8]. What's more, with the huge growth in the use of microblogs [9,10], [11,12] implemented SVM model to extract the emoticons and as these emoticons as the judging gists for analyzing the sentiment polarity of text. They considered that emoticons can accurately reflect the real emotion of microblogs.

As for methods based on the unsupervised method, they usually utilize massive untagged corpora and statistical methods. [13] directly used the similarity of words to judge their sentiment polarity. [14] also used the value of similarity, but the authors considered the similarity as the features of clustering method. Additionally, the statistical variables such as Point Mutual Information, Coincident Entropy and PageRank also have been used in sentiment analysis [3,15]. These statistical variables are exploited to calculate the sentiment polarity score for candidate sentiment words. Because the unsupervised methods are based on untagged corpora and statistical methods, they generally have relative low precision ratio. However, the advance of do not need manual tagged corpora leads these methods always have strong timeliness and can adapt variable situations.

3 Methodology

This section will describe the detailed flow of the CSWP and the rationale behind the design of the CSWP. Firstly, the overview of the CSWP will be introduced, and then the detail of the CSWP will be described step by step. Finally, the pseudo code will be shown in the end of this section.

3.1 System Overview

The architecture of the proposed CSWP method is shown in Fig. 1. The CSWP method mainly contains three phases: the preprocessing phase, the initial judge phase and the further judge phase. In the preprocessing phase, a large collection of online comments is segmented into independent Chinese words. The segment step is implemented by using the HanLP Chinese segment tool[1] —a useful Chinese text processing tool. After finished the segment step, all the adjectives are picked out to build up candidate sentiment words set. In the initial judge phase, the candidate sentiment words are given. The initial judge phase is responsible to calculate polarity score for every candidate sentiment words. In the end, candidate words are sent into further judge phase. The further judge phase exploits

[1] https://github.com/hankcs/HanLP.

sentiment morpheme matching method to make further judge for the polarity of the candidate sentiment words, and the final judging results will be output by this phase.

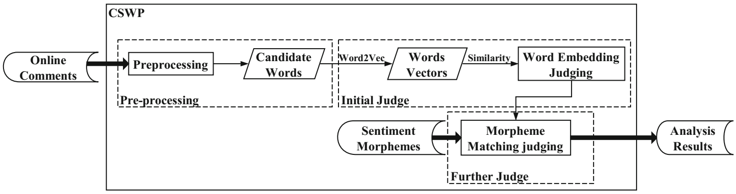

Fig. 1. Architecture of CSWP method

3.2 Preprocessing Phase

Unlike English, there are not separators between Chinese words, so the word segmentation is a necessary step in Chinese natural language processing. The preprocessing phase is responsible to segment the dataset and pick out all the candidate sentiment words by using the HanLP Chinese segment tool. During segmenting the dataset into independent words, the HanLP also tags the part-of-speech for every candidate sentiment words. The HanLP Chinese segment tool utilize “ad” to tag the adjectives. Then in the step of picking out the candidate sentiment words, the preprocessing phase picks out all the adjectives as the candidate words. The examples of Chinese words segmentation results are shown in Table 1.

Table 1. The example of word segmentation results

Original Sentence	Segment Result
今天天气很好 (The weather today is good)	今天/t 天气/n 很好/ad
欢迎来到中国! (Welcome to China!)	欢迎/v 来到/v 中国/ns ！ /w

3.3 Initial Judge Phase

The initial judge phase is used to calculate the polarity score for every candidate sentiment words. The specific process of this phase can be divided into the following two steps:

- **Step 1. Generate Word Vectors**
  Firstly, all the words are sorted according to their frequency in the dataset, and then twenty obvious positive candidate words and twenty obvious negative candidate words are picked out to build up positive negative seed sets. These two seed sets will be used in step 2 to assist with calculating the polarity

score. Next, all the candidate sentiment words are converted into 200 dimensional vectors by exploiting word2vec tool. The word2vec tool can convert the words into continuous space vectors and while retain the context features for the words.

- **Step 2. Calculate the Polarity Score**
  Firstly, the value of similarity between candidate sentiment words and sentiment seed sets (including the positive seed set and the negative seed set) is calculated. Then the similarity between candidate word and the positive seed set is defined as the word's positive similarity, and the similarity between a candidate word and negative seed set is defined as the word's negative seed set. After getting the similarity, the polarity score of a word is calculated by utilizing the word's positive similarity to minus it's negative similarity. The positive and negative of the score implies the positive and negative of the sentiment polarity of the word. The formula of computing similarity and polarity score are shown in 1 and 2, respectively.

$$sim(A_i|B)=\sum_{j=1}^{m}\frac{\sum_{l=1}^{k}A_{il}\times B_{jl}}{\sqrt{\sum_{l=1}^{k}(A_{il})^2}\times\sqrt{\sum_{l=1}^{k}(B_{jl})^2}\times m} \tag{1}$$

$$score(A_i) = sim(A_i|C) - sim(A_i|D) \tag{2}$$

  where A donates the set of candidate words, and B donates the set of sentiment seeds (positive seeds or negative seeds). $A_i(i = 1, 2, \ldots, n)$ and $B_j(j = 1, 2, \ldots, m)$ donate the vector of the word. Vector A_i is composed by $A_{i1}, A_{i2}, \ldots, A_{ik}$, and vector B_j is composed by $B_{j1}, B_{j2}, \ldots, B_{jk}$. The subscript k is the dimension of the word vector. What's more, C donates the positive sentiment seed set, and D donates the negative sentiment seed set. C_j and D_j donate the vector of word, they have same structure as B_j.

3.4 Further Judge Phase

Although word embedding method can judge the polarity of sentiment words independently, the frequency of words and the context of words can influence the analyzing results. Moreover, the size and the content of text also will influence the results. All above factors may lead to a reduction in performance. To mitigate the influence of these issues, a sentiment word formation rule matching method is proposed to assist with the word embedding method. Through utilizing the formation rule matching method, the CSWP can correct some errors caused by previous factors. But if you want to use formation rules to judge the polarity of sentiment words independently, you may also get a bad result, because formation rules always hard to be summarized, and it is impossible to collect all rules. Therefore, an independent rule matching method cannot adapt complex real situations.

The sentiment word formation rules which the CSWP exploits is based on sentiment morpheme matching method and synonym expanding method. Existing work mentioned that some sentiment words express their sentiment through their interior morphemes, so exploiting some sentiment morpheme matching rules can also analyze the sentiment polarity of Chinese words. For example, “很好 |very good” only contains the positive morpheme “好 |good”, so this word will be judged as positive sentiment word. But “不好 |not good” contains the privative morpheme “不 |not” and positive “好 |good”, furthermore, the privative morpheme has more front position than positive morpheme in “不好|not good”, so “不好 |not good” will be judged as negative sentiment word. The specific steps of this phase can be divided into following 3 steps:

- **Step1. Sentiment Morpheme Matching**
  After using cosine similarity to calculate the polarity score, CSWP scans every candidate word for checking whether it contains positive morpheme, negative morpheme and privative morpheme. In course of scanning morphemes, CSWP will utilize the morpheme matching rules to judge the sentiment polarity of the candidate words. If a candidate word cannot match any rule, then CSWP will jump to step 2. If all the candidate words have finished the sentiment morpheme matching, then CSWP will jump to step 3. The sentiment morpheme matching problem can be described as follows: given a candidate sentiment Chinese word $W = w_1 w_2 \ldots w_n$, where w_i is the morpheme in the Chinese word W, and subscript i is the position indicator of w_i. The goal of the method is to determine the position of w_i in W. Before giving the specific rules description, we have several conventions about symbol expression, they are shown as follows:
  1. if w_i is a positive morpheme, then use p_i to replace it.
  2. if w_i is a negative morpheme, then use e_i to replace it.
  3. if w_i is a privative morpheme, then use r_i to replace it.

  So, the specific rules can be described as follows:
  - **Rule1:** given a W, if $p_i \in W$, $e_j \notin W$, $r_k \notin W$; then $score = 1$.
  - **Rule2:** given a W, if $p_i \in W$, $e_j \notin W$, $r_k \in W$ and $i < k$; then $score = 0$.
  - **Rule3:** given a W, if $p_i \in W$, $e_j \notin W$, $r_k \in W$ and $k < i$; then $score = -1$.
  - **Rule4:** given a W, if $p_i \notin W$, $e_j \in W$, $r_k \notin W$; then $score = -1$.
  - **Rule5:** given a W, if $p_i \notin W$, $e_j \in W$, $r_k \in W$ and $k < j$; then $score = 1$.
  - **Rule6:** given a W, if $p_i \notin W$, $e_j \in W$, $r_k \in W$ and $j < k$; then $score = 0$.
  - **Rule7:** given a W, if $p_i \in W$, $e_j \in W$, $r_k \notin W$ and $j < i$; then $score = -1$.
  - **Rule8:** given a W, if $p_i \in W$, $e_j \in W$, $r_k \in W$ and $k < j < i$; then $score = 1$.
- **Step 2. Synonyms Expanding**
  If a candidate word can't match any rule, then the process will jump to this step. The candidate word sent to this step will be expanded to some synonyms by searching synonym dictionary. Then, the CSWP will use this word's synonyms continue to execute step 1. And if one of the synonyms can match a rule, the polarity score of this candidate word also will be changed.
- **Step 3. Output Final Results**
  Finally, after step 1 has finished the sentiment morpheme matching process,

the final results have come into being. The candidate words which have positive score will be marked as positive sentiment words, and the candidate words which have negative score will be marked as negative sentiment words. Moreover, the words which have zero score will be given up.

3.5 Pseudo Code of CSWP

Algorithm 1 explains how CSWP works. Line 2–7 corresponds the word embedding judging phase. Line 8–14 corresponds the sentiment morpheme matching phase.

Algorithm 1. Sentiment polarity judging

```
Input: SDP dataset, sentiment morphemes, synonym dictionary.
Output: positive sentiment word set, negative sentiment word set.
 1: Let $A$ donates the candidate word set, $C$ donates the positive seed set, $D$ donate
    the negative seed set.
 2: convert all $A_i$ into word vectors.
 3: build up $C$ and $D$ /*$C$ and $D$ are picked from $A$*/.
 4: for all $A_i$ do
 5:    use equation(1) to calculate the $sim(A_i|C)$ and $sim(A_i|D)$
 6:    $score(A_i) = sim(A_i|C) - sim(A_i|D)$
 7: end for
 8: for all $A_i$ do
 9:    if $A_i$ can match a rule of morpheme matching then
10:       update $score(A_i)$ according to the rule.
11:    else
12:       utilize synonym expanding to expand synonyms and jump back to Line 9.
13:    end if
14: end for
```

4 Experiments and Results

In this section, we firstly introduce the preparing of the experimental dataset, and then we introduced a series of comparative experiments. The purpose of these experiments is to evaluate whether the CSWP method has a desirable performance in word-level Chinese sentiment analysis compared with two baseline method.

4.1 Dataset

To validate the proposed method, 20000 online comments were extracted from Star Data Platform[2], a text data supply platform. For the sake of convenience, the collection of these comments is named as SDP dataset. The source of SDP

[2] https://www.istarshine.com/index.php/Data/dataSurvey#platform.

dataset including microblogs, news and post bar. Then, through summarizing from the NTUSD Chinese sentiment words dictionary[3], three sentiment morpheme sets were build up: a positive sentiment morpheme set, a negative sentiment morpheme set and a privative morpheme set. As for the synonym dictionary utilized in CSWP, the HIT-CIR synonym dictionary[4] was downloaded, this dictionary was developed by HIT Center for Information Retrieval Lab. The sample of sentiment morphemes are shown in Table 2.

Table 2. The sample of the Sentiment Morphemes

Po-Morpheme	善 |kind、好 |good、美 |beauty、优 |good、喜 |happy、妙 |wonderful
Ne-Morpheme	丑 |ugly、妙 |evil、愤 |angry、乱 |disorder、悲 |sad、笨 |stupid
Pr-Morpheme	不 |no、无 |nothing、不曾 |never、没 |no、没有 |no、永不 |never

4.2 Experimental Results

In this subsection, the performance of the proposed CSWP system is evaluated with SDP dataset. To illustrate the advantages of CSWP, we also implement and evaluate an unsupervised baseline method [4] and a supervised method [2] for comparison.

For illustrating the performance of the CSWP method. The 10-fold cross-validation is performed to the three methods. In each time of experiments, 2000 comments of SDP dataset are randomly selected as the training data and the rest are the testing data. The training dataset is used to train supervised baseline method, and the testing dataset is used to evaluate the performance of three methods. The final comparative results are divided into positive sentiment words extracting results and negative sentiment words extracting results. The results are shown in Figs. 2 and 3.

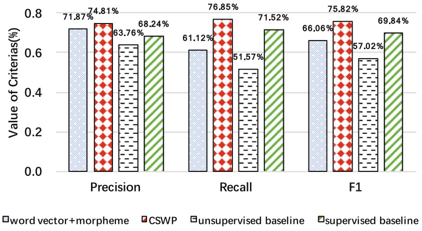

Fig. 2. Positive results

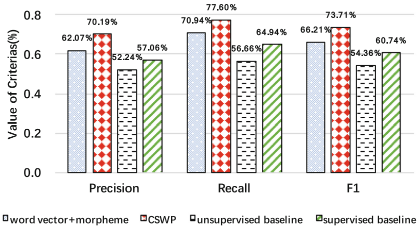

Fig. 3. Negative results

[3] http://www.datatang.com/data/44317.
[4] http://www.ltp-cloud.com/download/.

As can be seen from Figs. 2 and 3, the CSWP achieves the best F1 value, which is average 19.075% higher than unsupervised baseline and average 9.475% higher than supervised baseline. Additionally, we can see through the comparison of the first method and the CSWP that the synonym expanding step has average improved the accuracy of CSWP by 5.53% and average improved the recall ratio by 11.195%. It has brought significant improvement to the CSWP. The reason of the improvement is that the step of synonym expanding can improve the recognition ratio of the sentiment morpheme matching method. Therefore, there are more wrong results can be corrected by this phase, and thereby the performance of the CSWP is improved.

5 Conclusion

This paper proposed a novel Chinese Sentiment Words Polarity (CSWP) analyzing method that exploits the word embedding method and the sentiment morpheme matching method. Compared with the existing word embedding method, the performance of the CSWP method has achieved the significant improvement in analyzing the sentiment polarity of Chinese words. To illustrate the performance of the CSWP, two baseline methods were implemented, and then the three methods were evaluated on SDP dataset. The experimental results show that the CSWP method leads to a desirable performance compared with the unsupervised baseline method which only based on word embedding method, and it also can be seen from results that the performance of the CSWP also better than the supervised baseline method which based on maximum entropy model. The advantages of the CSWP method are that it only needs little manpower in preparing work, and it retains the advantages of flexibility and timeliness which belong to unsupervised method, so it will be more competent than supervised baseline method in different harsh work situation.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (61572060, U1536107, 61472024), and CERNET Innovation Project (NGII20151104, NGII20160316).

References

1. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009):12 (2009)
2. Fei, X., Wang, H., Zhu, J.: Sentiment word identification using the maximum entropy model. In: 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 1–4. IEEE (2010)
3. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association For Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)
4. Fan, X., Li, X., Du, F., Li, X., Wei, M.: Apply word vectors for sentiment analysis of app reviews. In: 2016 3rd International Conference on Systems and Informatics (ICSAI), pp. 1062–1066. IEEE (2016)

5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
6. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 174–181. Association for Computational Linguistics (1997)
7. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 355–363. Association for Computational Linguistics (2006)
8. Ku, L.-W., Liang, Y.-T., Chen, H.-H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI, pp. 100–107 (2006)
9. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. J. Assoc. Inf. Sci. Technol. **60**(11), 2169–2188 (2009)
10. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc, vol. 10 (2010)
11. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: the good the bad and the omg! Icwsm, **11**(538-541), 164 (2011)
12. Huang, S., Han, W., Que, X., Wang, W.: Polarity identification of sentiment words based on emoticons. In: 2013 9th International Conference on Computational Intelligence and Security (CIS), pp. 134–138. IEEE (2013)
13. Gauch, S., Wang, J.: Corpus analysis for TREC 5 query expansion. In: TREC (1996)
14. Wiebe, J.: Learning subjective adjectives from corpora. AAAI/IAAI **20** (2000)
15. Geng, H.T., Cai, Q.S., Kun, Y., Zhao, P.: A kind of automatic text keyphrase extraction method based on word co-occurrence. J. Nanjing Univ. **42**(2), 156–162 (2006)