2015 International Conference on Software Engineering and Information System (SEIS 2015)
ISBN: 978-1-60595-297-0

A Novel Feature-based Method for Opinion Mining in Chinese Product Reviews

Han-Qian WU[1,a,*], Li-Feng ZHOU[2,b], Xie JUE[3,c] and Yao LI[4,d]

[1,2,4] School of Computer Science and Engineering, Southeast University, China

[3] Southeast University-Monash University Joint Graduate School, China

[a]hanqian@seu.edu.cn, [b]492282524@qq.com, [c]jue.xie@monash.edu, [d]yao.li@seu.edu.cn

*Corresponding author

Keywords: Opinion Mining, Supervised Learning, Feature Model, Opinion Polarity.

Abstract. Due to the rapid growth of web resources in the digital era, user comments have become ubiquitous in most websites. In China, the most prevalent websites include user-generated content site such as DOUBAN, review site like TAOBAO and JD, and various Blog sites. Existing research focuses on the opinion mining area using two main methods, including the supervised learning method and the semantic-based method, both of which have been proven as effective approaches. This paper reports our efforts in improving the work done by previous research. First, a few dependency relations are constructed to identify features and the opinion set. In order to extract features effectively, various combinations of digits and letters are also taken as features. Next, a novel distance method is utilized to identify features and opinion tuples. Finally, we build a new feature model using supervised learning method to detect the overall opinion polarity of a review. With the comparison to traditional methods, the experimental results indicate that our approach is more effective than the others.

Introduction

Opinion mining or sentiment analysis refers to identifying and extracting subjective information or opinion from reviews. As a result of the explosion of big data produced by online shopping websites, micro-blogs or other newly web media, the research of opinion mining has received significant attention in recent years both in industry and academia [1, 2, 3]. The target of opinion mining is to predict whether the polarity of opinion in one review might be positive or negative**.** The tasks of opinion mining include the following: 1) extracting product features; 2) mining opinion expressions or opinion words related with a feature; 3) predicting sentiment polarity; There are normally two ways for the solution of opinion mining, one is the machine learning method using various algorithms (SVM, Bayesian and maximum entropy [4, 5]**,** etc.) and the other is semantic-based method [6, 7, 8]**.** However, both of the two methods have their own drawbacks. The machine learning method needs plenty of manual labeled samples to train the classifier, and its efficiency and accuracy could be very low. Moreover, most researchers use Bow and N-gram as the feature model, which is regarded as time-consuming, large dimensional, and data sparsity. Similarly, the semantic-based method depends on the seed selection and corpus's scale, and it cannot be applied to different domains [9].

This work is supported by National High-tech R&D Program of China (863 Program) (Grant No. 2015AA015904) and the Open Project Program of the State Key Lab of CAD&CG(Grant No. A1429), Zhejiang University.

In order to improve the accuracy of sentiment analysis, 6 mostly used dependency relations are firstly selected to get a refined sentence structure. Then opinion word, interjection, interpunction and semantic information are extracted to build a novel feature model for applying the machine learning algorithms. In order to obtain deep analysis for the feature model, we select different attributes from the feature model to design the experiments. Finally, 5 most related attributes are obtained, which reduces half the computing time. The experiments shows that our proposed new method yields better results than the two baseline feature models.

The reminder of this paper is organized as follows. Section 2 performs the feature and opinion tuple extraction, which is the key step of this study. That includes an overview of the used data set, and the selection of dependency relations by analyzing the corpus. Section 3 focuses on review sentiment polarity assignment, which mainly includes the task of building a classification feature model by using the collected corpus as training and testing data. The experimental results and future work are summarized in Section 4.

Feature AND Opinion Tuple Extraction

A pair of feature and opinion means the match of a feature word to an opinion word [10]. The extraction of feature and opinion pairs can help customers or vendors to realize the features of the product more clearly. A Chinese review normally includes one or more sentences, and a sentence may refer to more than one feature word and opinion word. A feature and opinion pair consists of a feature word and a corresponding opinion word in a review. For example, the review " (it feels good, the price is cheap, we use it every day)" contains a feature and opinion pair [(price), (cheap)], where the feature word is (price), and the related opinion word is (cheap). Particularly, this paper extracts explicit feature and opinion tuples, which extends feature and opinion pairs. The followings are the main steps to extract the tuples of the feature and the opinion.

Data Sets Preparation

The reviews of four products from jd.com are crawled, which includes a MP3 player, a cell-phone, a notebook and an earphone. For the original dataset includes a large amount of noise data, it has to be filtered to obtain experimental data with high quality. The followings are the filtering conditions for the raw dataset:

1) If the length of a review is less than 4 words or longer than 50 words, then omit it.

2) If one review is only composed of special character, then omit it.

3) If the duplication degree of content in a review is bigger than 0.5, then omit it.

4) Because the implicit feature would not be considered here, therefore if there has no noun in review, the review would be omitted.

5) The Jarcard algorithm is adopted to compute the similarity of two reviews, if the similarity value of two reviews exceed the thresholds (here is set as 0.8), then the reviews would be omitted.

We write the above conditions as java program, all raw reviews were deal with by the filtering program one by one. At last, the total number of mp3's reviews is 6070 (Data1), the number of cell-phone's reviews is 6910 (Data2), the number of notebook's reviews is 5210 (Data3), the number of earphone's reviews is 7092 (Data4).

Dependency Parsing

Dependency parsing has been widely employed in review mining because of its high performance and domain independency [11, 12, 16]. The syntax structures of language units could be obtained by utilizing LTP' grammar analyzer published by HIT. The main elements of dependency structures are dependency arcs. Each arc links a core word and subordinate word. The arc presents the relations type, such as SBV (subject-verb), VOB (verb-object), COO (coordinate) etc. For example, the parsing result of the review (the outlook of the phone is very pretty) is shown in Figure 1.

Figure 1.

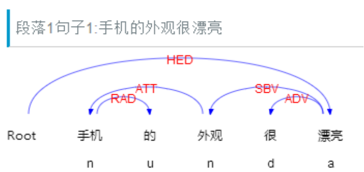


Figure 2.


```
<sent id="0" cont="手机的外观很漂亮">
    <word id="0" cont="手机" pos="n" ne="O" parent="2" relate="ATT" semparent="2" semrelate="Poss"/>
    <word id="1" cont="的" pos="u" ne="O" parent="0" relate="RAD" semparent="0" semrelate="mAux"/>
    <word id="2" cont="外观" pos="n" ne="O" parent="4" relate="SBV" semparent="4" semrelate="Exp"/>
    <word id="3" cont="很" pos="d" ne="O" parent="4" relate="ADV" semparent="4" semrelate="mDegr"/>
    <word id="4" cont="漂亮" pos="a" ne="O" parent="-1" relate="HED" semparent="-1" semrelate="Root">
        <arg id="0" type="A0" beg="0" end="2"/>
        <arg id="1" type="ADV" beg="3" end="3"/>
    </word>
</sent>
```


The parse results is a directed graph, the node presents the words extracted from the review, while the arc between nodes represents the dependency relations. The direction of an arc (SBV) indicates the subordinate word (pretty) that depends on the core word (outlook). Although the visualization result is very intuitive, it is difficult for the program to realized. For this reason, the results are stored in XML format, as shown in Figure 2. The review is interpreted to structured text, which contains one or more sentences. Each sentence is composed of multiple words. Each word has its own attributes, which only need "id", "cont", "pos","parent", "relate" etc. The "cont" is the word itself, while the "pos" is the POS tagging of the word, For instance, 'a' represents adjective like (pretty); 'c' represents conjunction, like (and); 'd' represents adverb like (too); 'n' represents nouns like (outlook), 'v' represents verb like (like); 'e' presents exclamation like (sigh). The "relate" presents the semantic relation between word itself and its parent word. The parent word and the current word are the start and the end point of the relation arc respectively.

Dependency Relations

The result of dependency parsing can produce fourteen different relations, but only a few of them is useful to this research. Although parsing dependency relations is an effective and common way in opinion mining, its performance is not ideal when using it alone [17]. Therefore, combined with POS, some patterns represented in table 1 are constructed to extract features, opinion words and adverb alternately according to Chinese expression habits.

Table 1.

Dependency pattern	relation	feature	Opinions
ATT-Patten	attribute	noun and noun group	
VOB-pattern	verb-object	noun or some verbs	adjective
SBV-pattern	subject-verb	noun	adjective
COO-pattern	coordinate	nouns	adjective
ADV-pattern	adverbial		adjective, some noun,
CMP-pattern	complement	noun or some verb	adverb
Other-feature		letter and digit	adjective

Feature AND Opinion Tuple

Applying the above patterns, the feature set and opinion word set of each sentence could be extracted easily. The task of this part is to extract feature-opinion tuple and to summarize the result. The tuple is a quadruple (feature word, negation, adverb of degree word, opinion word, sentiment value). The initial value of sentiment value is set to zero. Some researchers have used distance and lexicon-based method to extract feature-opinion pair, but it has the following drawbacks [13]. First, a sentence usually includes multiple opinion words, which is a problem that existing methods cannot handle. Second, most of the existing methods only consider whether the opinion is positive or negative, so might lose some information. For instance, many reviewers criticize products by using pretty short text, such as (it's fine, but the quality is not good enough). A new method is proposed – DIMM (Distance with interpunction Match Method presented in Figure 3), which uses the interpunction and distance between feature words and opinion words to decide whether they can match. In fact, if we use distance alone to extract them. In the above case, the pair result is [(quality), (fine)], which is totally opposite to the actual meaning- [(quality), (not so good)]. The DIMM algorithm can be described in Figure 3.

Figure 3.

```
DIMM Algorithm
Input: Reviews, adverbSet, negationSet
Algorithm:
for review in reviews:
    sentences = all sentences in a review
    for sentence in sentences :
        recongnize feature and opinion information with patterns
        flag = whether interpunction exists
                between feature and opinion
        dist = distance between feature and opinion
         adverb ,negation= the related modifier word
         if(flag is false and 0<dist<7 ):
             featureOpinionTuple
                   .add(feature, negation,adverb,opinion,0.0)
         else :sentence++;
     review ++;
Output: featureOpinionTuples
```

Figure 4.

```
Sentiment Strength Algorithm
for each tuple in feature-pinion do
  score = 0.0;
  if tuple.opinion in sentiment dic
      dicValue = get Value in sentiment dic
  else if tuple.opinion in Synon dic:
       obtain its synon from Synon dictionary
       and related strength for dicValue.
  else if tuple.opinion in NTUSD:
       dicValue = 1 or -1
  if tuple.adverb is not null and exist in adverbSet:
      advDegree = the degree of adverb word
  else:  advDegree=0.01
  if tuple.negation is not null and exist in negationSet :
       if the number of it is even
           negValue= 1
    else the number of it is odd:
           negValue = -1
  else: negValue = 1
  computer score according to formular (1)
```

Except the extraction of the feature-opinion pair, a comprehensive analysis over a review, its contained adverbs and negation words should also be considered. With the help of using manually selected negative set (such as (not), (without), (don't have), etc.) and HowNet's adverb dictionary (such as (very), (extremely) or (too) with its degree, etc.), a more refined tuple could be obtained by combining them together. For example, (the appearance is good, the sound effect is good, and the price is very cost-effective, but the packaging is not very good). The corresponding feature-opinion tuple could be extracted like follows: ((appearance, , ,nice,0.0), (sound effect, , ,good, 0.0), (price, , very, cost-effective, 0.0), (packaging, not, very, good, 0.0), "0.0" is the initial sentiment value. After getting these tuples, the sentiment strength could be computed according to the negation, the adverb of degree, and the opinion words through the use of the Sentiment Strength Algorithm presented in Figure 4, and with the help of sentiment dictionary from DuTIR-Emotion word, Synonym dictionary from HIT, and positive/negative sets provided by NTUSD. Below is the formula to compute the sentiment value of a feature [14]:

$$score(f) = negValue\sum_{o_i} advDegree * dicValue(o_i) \tag{1}$$

In the case of score<0, we view the polarity of a feature opinion tuple as negative. In the case of score >0, it indicates the polarity is positive, and score =0 means neutral.

Assignment of Review Sentiment Polarity

The task of this part is to assign the sentiment polarity of reviews. The polarity can be 1 or -1 (1 means positive,-1 means negative), so it could be viewed as a binary classification problem. For a review, it can be classified as a positive or negative. Therefore, some machine learning algorithms such as SVM(Support Vector Machines), ME(Maximum Entropy)、and LG(logistic regression) could be utilized. This paper proposes a novel and simple method to construct a feature model for classification, which is enlightened by Liu Bing's work [15]. We use a novel and simple way to construct a feature model for classification.

Construction of the Classification Feature Model

A review is considered as an entity, whose strength polarity is 1(positive) or -1(negative), the value 0(neutral) is not included, for it is ambiguous and difficult to deal with. We usually analysis the content of the review to obtain its polarity, thus we focused on some content centric features which are information about the review. We obtained the following features：

1) The length of the review (F1). This feature is selected since longer review tends to get more information and customer attention.

2) The number of sentences of a review (F2). This feature is chosen based on the consideration that the motivation of reviewing tends to be less ambiguous, thus can easily point out the polarity.

3) The number of negation word in a review (F3). It is easy to understand that the polarity of a sentence can't be converted if the number of negation words is even. Otherwise, if the number of negation words is odd, the polarity of a sentence would be an opposite value.

4) The sentiment value of interpunction in a review (F4). It is selected because some interpunction can enhance or weaken the sentiment strength, such as "!" means a strong tone.

5) The positive sentiment (F5), and the negative sentiment (F6) in a review. First, a feature opinion tuple needs to be dug before the sentiment value can be calculated through the Sentiment Strength Algorithm with the help of DuTIR-Emotion word and Synonym dictionary. We choose these features because customers always express one opinion from a feature, and a review maybe contain two kinds of opposite views between two features.

6) Opinion word frequency (F7), which is the number of valid opinion word frequency in a review.

7) Adverb degree (F8) and the number of adverb word in a review (F9). Odd numbers from 1 to 9 are selected to represent the degree. For example, the level of (extreme) is 9 and (mild) is 1. These features are selected since they can strengthen or weaken the utterances of an opinion.

8) The value of positive interjection word (F10) and negative interjection word (F11). These features can express opinions more directly, such as (sigh) means the attitude of a customer who is not satisfied with the product, while (ha-ha) always indicates that one is pleased with the product. Such features could endow the review with an emotional color.

9) The time that reviewer comment on products (F12). This feature is not important for classification, but it can be used to investigate people's evaluation to the product at a different stage.

In this paper, WEKA is used to run SVM and the LR algorithm, and Eclipse is used to run the ME model. WEKA is an open source platform for data mining, which incorporates multiple machine learning algorithms, including classification, regression, association, cluster, etc.

Experiments and Discussions

The performance of the proposed method has been evaluated on four datasets, which are collected from jd.com by running the crawler program we developed. The method proposed in this paper is compared with some state-of-the-art methods including Bigram and Bag of Words (BOW). The reason for selecting Bigram is that Bigram is proven to produce better results than Trigram or High grams in most literatures.

Figure 5.

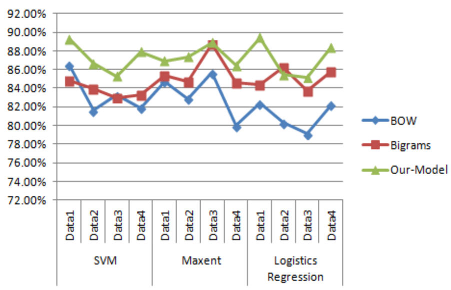
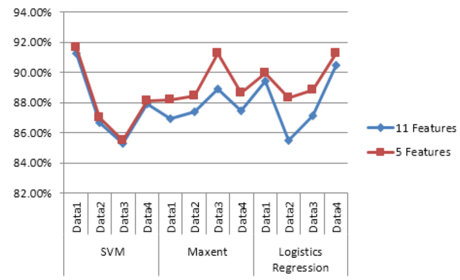


Figure 6.



40 different product reviews have been derived from jd.com. After filtering out stop words, they then are segmented by using word segmentation tool to establish a large corpus. In our experiments, ten cross validation is adopted for generating the classification model, with the use of our proposed feature model. We believe our model will produce better results than BOW and Bigrams. BOW considers the document as a collection of words without considering any of the Lexical, syntactic and grammatical factors between the individual words. The term frequency, which occurred more than twice, is counted. Bigram is a subsequence of 2 words from a given sequence, we count tf-idf weight for each term and the feature vector is limited by setting a threshold for the frequency of the Bigrams. In addition to that, interjection and stop words are filtered for both of them.

In order to evaluate whether the proposed feature model has better accuracy than the traditional BOW and Bigram models, we utilize SVM, ME and Logistics as the classifier with 10-folds cross validation in WEKA.

Figure5 demonstrates the results of these experiments. The method this paper proposes yields the best accuracy on Dataset1 when using Logistics. Compared with BOW and Bigram, the accuracy of this method is increased by about 7 points in percentage, which means that this model is more effective for sentiment analysis. Since the model proposed in this paper utilizes the semantic information, the word of the speech, and many other related information to construct the feature model, which altogether achieve the better results.

We further conduct multiple experiments on four datasets. Five most relevant aspects are produced in the dataset according to accuracy. They are the length of review (F1), the number of sentences of a review (F2), positive sentiment (F5), negative sentiment (F6), and adverb degree (F8). Figure 6 shows the results using the five features, the accuracy improved almost 3 points in percentage compared with using 11 features. These features are more efficient because experienced reviewers tend to express critical and objective views in long text after buying a product, and they often use some adverbs to modify their positive or negative opinions.

Conclusions

In this paper, a novel feature-based opinion mining method for opinion mining is demonstrated. In this method, feature opinion tuple can be extracted efficiently. A novel feature model is constructed, which integrates semantics based on sentiment analysis of Chinese product reviews. Four different review data sets of electric products from the website jd.com are obtained to evaluate the effectiveness of the proposed model. As a result, 5 most efficient features are obtained. Through multiple experiments, the results indicate that our proposed feature model is effective for sentiment analysis.

References

[1] Pang Bo, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up: sentiment classification using machine learning techniques. in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.

[2] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, “Sentiment classification of Internet restaurant reviews written in Cantonese”, Expert Systems with Applications (2011).

[3] Liu B. Sentiment analysis and opinion mining [J]. Synthesis Lectures on Human Language Technologies, 2012, 5 (1): 1-167.

[4] Rui Xia, Chengqing Zong, Shoushan Li, “Ensemble of feature sets and classification algorithms for sentiment classification”, Information Sciences 181 (2011) 1138–1152.

[5] Duric A, Song F. Feature selection for sentiment analysis based on content and syntax models [J]. Decision Support Systems, 2012, 53 (4): 704-711.

[6] Taboada M., Brooke J., Tofiloski M., et al. Lexicon-based methods for sentiment analysis [J]. Computational linguistics, 2011, 37 (2): 267-307.

[7] Qiu G., Liu B., Bu J., et al. Opinion word expansion and target extraction through double propagation [J]. Computational linguistics, 2011, 37 (1): 9-27.

[8] Qiu G., Liu B., Bu J., et al. Expanding Domain Sentiment Lexicon through Double Propagation [C]//IJCAI. 2009, 9: 1199-1204.

[9] Kanayama H., Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis [C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006: 355-363.

[10] Ferreira L., Jakob N., Gurevych I. A comparative study of feature extraction algorithms in customer reviews[C]//Semantic Computing, 2008 IEEE International Conference on. IEEE, 2008: 144-151.

[11] Mosha C., Tianfang Y. Combining dependency parsing with shallow semantic analysis for Chinese opinion-element relation identification [C]//Universal Communication Symposium (IUCS), 2010 4th International. IEEE, 2010: 299-305.

[12] Thet T.T., Na J.C., Khoo CSG (2010) Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36 (6): 823-848.

[13] Ding X., Liu B., Yu P.S. A holistic lexicon-based approach to opinion mining [C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 231-240.

[14] Shen Y., Li S., Zheng L., et al. Emotion mining research on micro-blog [C]//Web Society, 2009. SWS'09. 1st IEEE Symposium on. IEEE, 2009: 71-75.

[15] Jindal N., Liu B. Opinion spam and analysis [C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 219-230.

[16] Lizhen L., Wei S., Hanshi W., et al. A novel feature-based method for sentiment analysis of Chinese product reviews [J]. Communications, China, 2014, 11 (3): 154-164.

[17] Mosha C., Tianfang Y. Combining dependency parsing with shallow semantic analysis for Chinese opinion-element relation identification [C]//Universal Communication Symposium (IUCS), 2010 4th International. IEEE, 2010: 299-305.