Sentiment Analysis of Chinese Online Reviews
Based on Word2vec and DBN

Sai-hong ZENG* and Chao-fan DAI
Science and Technology on Information Systems Engineering Laboratory,
National University of Defense Technology, Changsha, China

*Corresponding author

Keywords: Word2vec, Deep belief network, Sentiment analysis.

Abstract. Sentiment Analysis of Online Reviews has a broad application prospect in the commercial field as it truly reflects the user satisfaction with the product. However, the online reviews belong to short text and the feature words are sparse. The traditional classification algorithm makes a very little effect. And the traditional method based on the bag-of-words model always ignore the relationship between words and lose the context information, resulting in poor classification results. In this paper, we have proposed a method combining Word2vec and DBN towards the sentiment classification for Chinese online reviews. Experimental results show that our method has good performance.

Introduction

With the development of Internet and e-commerce, the user can post comments of their purchase on the Internet. Through the analysis of such information, businesses can know their own strengths and weaknesses, so as to make the right choice. Sentiment analysis is a method that finds the consumer's opinion by automatically analyzing the textual content of a commodity review [1].

At present, there are two main methods of sentiment analysis: unsupervised classification based on emotion dictionary and supervised classification based on machine learning. Due to its better classification effect, Sentiment analysis based on machine learning method has become more and more popular. In 2002, Pang et al. first applied the method of machine learning to field of Sentiment analysis [2]. They try to use N-grams model to extract the features, then use three kinds of classification model in the field of machine learning (SVM, NB and ME) to test. They found that selecting unigrams as feature set and using the SVM classification get good effect. Yang used SVM in emotion recognition of a sentence [3] by extracting and analyzing the characteristic of the related emotional words; Aiming at the shortcomings of the supervised learning classification, Li Suke et al. proposed a semi-supervised classification algorithm [4] for clustering of the emotional feature.

However, the online reviews belong to short text and the feature words are sparse [5]. The traditional classification algorithm makes a very little effect. Also, the traditional classification algorithm only focuses on the lexical and syntactic features between the words in the sentence while the semantic features are rarely studied.

Therefore, the main idea of this paper is to model this problem and to seek for a resolution when dealing with the short-text sentiment analysis of online reviews by combining semantic features and machine learning techniques. This paper combined the word2vec and DBN model to classify the comment text in order to enhance the classification effect.

Sentiment Classification Model

The sentiment classification model combining Word2vec, feature selection and DBN for Chinese online reviews is shown in Figure 1. The process of sentiment classification is the following:

1) Pretreatment and Comment Vectorization. Before carrying out specific work, we did a simple pre-processing work about data of comments, getting corresponding segment files. Then Word2vec trained out each word of the corpus to word vectors.
(2) Feature Extraction. Based on the theory, the paper puts forward two kinds of feature selection methods: feature selection based POS (part of speech) tagging and feature selection based POS structure. In result, the text vector \( V(s) \) based on POS tagging and POS structure vector \( V(e) \) are merged to form the final emotional feature vector \( V \).

(3) Deep Belief Network. Finally, we get the results of online emotional classification by DBN’s unsupervised learning.

What's special about the model is: (1) That our method focuses on the original meaning of the word and context information while the semantic features are rarely studied in the traditional classification algorithm. (2) That the method of the feature sets construction is original. (3) And that DBN’s unsupervised learning based on depth feature selection effectively solves the problem of high dimensional vector generated by feature extraction.

![Figure 1. Framework of sentiment classification.](image)

**Pretreatment**

This paper does not use the public experimental data set on the network, but use our own crawled data. Therefore, before carrying out specific work, the data need to a simple pre-processing work to improve the efficiency of the subsequent algorithm analysis.

(1) Eliminating data duplications, this article mainly removes a large number of duplicate comments.

(2) Noise elimination, this article mainly removes the comment that contains only emotions, English characters, punctuation and other non-text information.

(3) Word segmentation, this paper uses the NLPIR Chinese word segmentation system developed by Dr. Zhang Hua-ping of the Institute of Computing Technology of the Chinese Academy of Sciences for segment and POS tagging in the corpus.

(4) Removal of stop words, this article mainly removes the English characters, mathematical characters, numbers, punctuation in comments.

**Comment Vectorization**

One of the popular methods about Sentiment Analysis is to use word2vec model to obtain word vector of corpus. The traditional method of using the bag-of-words model ignores the relationship between words and always loses the context information, resulting in bad effect.

In 2013, Google[6] proposed a depth learning model which is called Word2vec. Combining the context information, Word2vec trains the word vector by mapping words into the V-dimensional vector space. By using Word2vec, Vector operations between words can corresponds to semantics[7]. Word2vec model is divided into two kinds: CBOW model (see Figure 2) and Skip-gram model (see Figure 3).

The CBO model predicates the current word by the \( c \) words before and after the word \( w(t) \), while the Skip-gram model uses the word \( w(t) \) to predict the \( c \) \( (c = 2) \) words before and after the current word.
Since the training of the CBOW model is similar to that of the Skip-gram model, we only introduced the training process of the CBOW model here. In the training process of the CBOW model, the input layer is the 2c word vector in the context of the word w(t) while the projection layer vector X_w is the sum of the 2c word vectors. The output layer is a Huffman tree constructed from the words that appear in the corpus as leaf nodes and the number of occurrences of each word in the corpus as weights. In this Huffman tree, the leaf nodes are N (= | D |), corresponding to the words in the dictionary D, and non-leaf nodes are N-1. The result of X_w is predicted by stochastic gradient increasing algorithm, which maximizes p (w | context (w)) where context (w) refers to 2c words in the context of the word.

When the neural network training is complete, all the word vector w can be obtained. It is interesting to note that when a word is represented by a word vector, a rule similar to this can be found: "king" - "man" + "woman" = "queen" [8]. It is obvious that the word vector is very useful for expressing semantic features of words.

The process of the comments vectorization algorithm is described in detail as follows:

First of all, the comment is divided into two classes: positive and negative according to its emotional tendencies. Approximately 3,000 training samples were taken from each class according to a certain percentage.

Secondly, according to the method in Section 1.1, we put the comments for noise elimination, word segmentation (each word separated by space) and removal of stop words, getting two classes of files: positive.txt, negative.txt.

For positive.txt, negative.txt, we execute the word2vec instruction respectively (format: word2vec train word_file_name output vectorization_file name cbow0 size5 window10 negative0 hs1 sample1e 3 threads2 Bi_nary0), getting two classes of files: positive.out negative.out. Each row of positive.out or negative.out is a word vector. For example: accessories 0.034467 -0.052937 0.098494 -0.037 461 -0.024790……

Feature Extraction

Feature Selection Based on the Method of POS (Part of Speech) Tagging

The method selects features according to the method of POS tagging. The different selection directly affects results [9]. For example, the result of selecting only adjectives as feature set is not as good as that of the simultaneous selection of adverbs, verbs, and adjectives.

The word vectors selected in each comment are concatenated or averaged to give a vector representation of the phrase or sentence [10], as shown in Example 1:

Example 1: 这台热水器使用方便。

If only nouns, adjectives, and verbs are selected as features, the feature set consists of three words. Then we extract their corresponding word vector representation from the word vector library W, and we get the text vector of Example 1 V(s) by vector averaging. The process of constructing the sentence vector is shown in Fig 4.

Feature Selection Based on POS (Part of Speech) Structure

Many research results show that, as an important grammatical feature of semantic information, the
POS may play an important role in the emotional classification of implicit emotional statements that may not contain emotional words. In this paper, we construct POS vector to represent the POS feature of text. The POS of each word in the sentence is connected in order to get the representation of the POS of the text. As shown in example 2.

Example 2: 这台热水器使用方便。

The POS structure vector of text is denoted as \( V(e) = \{PNVNA\} \) in Example 2, A ("方便" adjective), V ("使用" verb), N ("热水器" noun), P ("这台" pronoun).

**Feature Vector Fusion**

In order to improve the accuracy of emotion classification, the text vector \( V(s) \) based on POS tagging in Section 2.1 and POS structure vector \( V(e) \) generated in Section 2.2 are merged to form the final emotional feature vector. The construction of the emotional vector is shown in Figure 5.

![Feature Vector Fusion](image)

**Deep Belief Network (DBN)**

Deep Belief Network (DBN)\(^{[11]}\) consists of many simple learning modules, each of which is a Restricted Boltzmann Machines (RBM)\(^{[12],[13]}\), as shown in Figure 6. RBM network consists of a visual layer and a hidden layer. And there are connections between the layers where there is no connection between the layers of cells. The training hidden layer unit obtains the high-order data characteristics from the visualization layer. DBNs is a probability generation model that creates a joint distribution of observed data and labels, at the same time, it calculates the probabilities of both \( P(\text{observation} \mid \text{label}) \) and \( P(\text{label} \mid \text{observation}) \).
The output of the lower RBM in the DBN is taken as input to the upper RBM. Each layer of RBM energy is defined as follows:

\[ E(v,h;\theta) = - v^T W h - b^T v - a^T h - \sum_{i=1}^{D} \sum_{j=1}^{F} W_{ij} v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{E} a_j h_j \]  

(1)

First of all, we use the RBM training algorithm to train the RBM layer. After the training of all the RBM, the top layer of DBN adjusts the weights by using BP algorithm, and obtains the classification surface of a DBN network. DBN performance is better than the pure BP algorithm. Because the BP algorithm of DBNs only needs a local search on the parameter space of weight, the training is faster than that of the forward neural network. Because the input of the 0 element in the first input layer is very large, when selecting the number of neurons in RBM, compression is considered. The specific number of neurons was obtained based on experimental experience.

We extract the training set and test set from the feature vector \( V \) generated in the section 2, and input parameters of DBN to learn and test in the first input layer. The training set is used to train the model, and the test set is used to test the learning effect of the model. Finally, it output the results of online emotional classification.

**Experiment**

**Experimental Corpus**

<table>
<thead>
<tr>
<th>Emotional tendency</th>
<th>Positive</th>
<th>Negative</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>3000</td>
<td>3000</td>
<td>6000</td>
</tr>
</tbody>
</table>

We have collected 6000 reviews about the water heater of a certain brand from the Jingdong with a website crawler. The emotional tendencies of the comments are tagged manually. And this article only analysis positive emotions and negative emotions (ignoring the neutral emotions). A deep belief network is used to classify the candidate feature sets. In this paper, the corpus is divided into training set and test set with ratio of 4: 1.

**Evaluation Criteria**

The main indicators of the evaluation of text classification are: accuracy rate, recall rate and F value\(^{[15]}\), the specific formulas are as follows:

\[ \text{precision} = \frac{a}{a+b} \times 100\% \]  

(2)

\[ \text{recall} = \frac{a}{a+c} \times 100\% \]  

(3)

\[ F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100\% \]  

(4)

In the formulas above, the variable \( a \) represents the number of documents that the classifier correctly classified. The variable \( b \) represents the number of documents that the classifier judged incorrectly which belongs to this class but does not belong to this class originally; The variable \( c \) represents the number of documents which the classifier judged do not belong to this class but actually belongs to.

**Experiments of Sentiment Analysis**

In this paper, we get the accuracy rate, recall rate and F value of each method by comparing the 10
methods as shown in Figure 7. As can be seen from (a), (b) and (c), the precision, recall, F value is relatively high in experiment. Especially, the result of simultaneous selection of adverbs, verbs, adjectives and POS construction has proved to be extremely powerful.

![Comparison diagrams of experimental evaluation.](image)

In summary, this paper combines the context semantic and text depth features. Experimental results show that the proposed method has good performance, indicating the effectiveness of this algorithm.

**Conclusion**

With the socialization of the network in China, the Chinese online reviews have become an
important data source of enterprise competitive intelligence analysis. Because this kind of information has the characteristics of short length, high information density and obvious emotional tendency, this paper proposed a method of sentiment classification combining Word2Vec and DBN for this type of text. The method focuses on the original meaning of the word and context information with the use of word2vec. Through the construction of feature set by vector fusion of multiple features and unsupervised learning method of DBN, this study has achieved good results. Especially, the result of simultaneous selection of adverbs, verbs, adjectives and POS construction has proved to be extremely powerful, indicating the effectiveness of this algorithm.

However, this study has a long way to go from the best results. The two kinds of features selection methods used in this paper are not enough to find all the emotional features in the text. In the future work, we will focus on the extraction of word features.

References