Sentiment Analysis Of Movie Reviews Based On Improved Word2vec And Ensemble Learning

Xiaoan baoa, Shasha Lin, Ruilin Zhang, Zechuan Yu, Na Zhang

School of information science and technology, Zhejiang Sci-Tech University, Hangzhou, 310018, China
Email: 294249711@qq.com
abaoxiaohan@zstu.edu.cn

Abstract. Taking the field of movie review as an example, this paper proposes a sentiment analysis method based on improved word2vec and ensemble learning. The basic design idea is: firstly build the corresponding corpus through new word discovery and use the TF-IDF algorithm to exponentially weight the word2vec word vector, which is used to integrate the semantic relationship between words and the importance of vocabulary information into the model; secondly, to avoid the cumbersome problems of data labeling, the existing algorithms of automatic labeling reviews are improved to increase the adoption rate of data; finally, Stacking algorithm is used to train and classify the emotional data. The proposed model can simplify the domain text representation and improve the classification performance of the model. The experimental results show that compared with existing methods, the accuracy, precision and recall rate of the algorithm proposed in this paper have been improved on film review data.

1. Introduction

Text sentiment analysis is also called opinion mining and tendentiousness analysis. Simply speaking, it is a process of analyzing, processing, inducing and reasoning the subjective text with emotional color.

At present, due to the universality of data, the test results of text sentiment analysis models trained for ordinary texts are not very ideal when faced with texts in different fields. When targeting specific areas of text, it is necessary to study text features deeply, but the process is complicated and can not be generalized. At the same time, some researchers often only consider the effect of classifier in a specific situation when selecting classifiers, which has certain limitations in practical application. Therefore, sentiment analysis algorithms that can not only simplify the expression of domain text characteristics but also improve the accuracy of the model have been a hot research topic.

Rui Li et al.[1] used TF-IDF as the weight of word2vec word vector to incorporate domain text features, but the experimental results were not very significant. In 2019, Fan Zhen [2] proposed a method to automatically annotate review text using user ratings as weak tagging information. However, this paper only studies the reviews of a single film and the ratings were mostly positive, which is not universal. Hence, this paper proposes a sentiment analysis method based on improved word2vec and ensemble learning. The main work is as follows:

1) Based on the integration of the basic affective words, the new words are found by the three indexes of word frequency, agglomeration degree and degree of freedom, and the emotional tendency
is calculated by the PMI Algorithm, so as to construct an emotional dictionary suitable for the field of movie reviews, and combine user rating information to label the data.

2) Split the text into words based on the new words in 1), and the corpus of the film field is constructed to calculate the IDF value.

3) Construct an exponential function with $e$ as the base and TF-IDF as the exponent. On the basis of the word2vec word vector model, this function is used to weight the sentence vector of each review, thereby converting the text into a numerical form that can be input into the model.

4) Taking into account the idea of "integrating the strengths of different models" in ensemble learning and its excellent performance in various fields, this article uses the Stacking method to analyze the emotion of data.

2. Related Work

At present, there are two kinds of sentiment analysis techniques. One is the method based on the emotional dictionary. It is mainly based on the sentiment words marked by domain experts in advance to match text strings, and then get the general sentiment tendency of the text through certain rules, thereby mining positive and negative information. The other is the method based on machine learning. The primary idea is to transform sentiment analysis into a classification problem. The current mainstream direction is the binary classification, which means that comments are divided into positive and negative emotions.

Pang et al. [3] expressed the text as different feature combinations, and conducted comparative experiments under different classification algorithms. The results show that SVM is the best in the case of using Unigrams features. Based on the basic emotion dictionary, Yu Li [4] has integrated the emotion dictionary of Emoji and the emotion dictionary of online words, and used the maximum entropy and the support vector machine classification model to verify the effectiveness of the dictionary. Zhitao Wang [5] and others used statistical information for new word mining and sentiment recognition based on mutual information between points, constructed a new sentiment word dictionary, and proposed a Weibo sentiment analysis method based on rule sets and dictionaries. PKalaivani et al. [6] studied sentiment analysis methods for online movie reviews and compared three machine learning methods. The results show that the SVM method is better than the Naive Bayes and KNN methods. Yan Wang [7] and others used an unsupervised word2vec model to convert movie review words into word vectors and used word frequency statistics to extract public concerns.

3. Emotion analysis and calculation

3.1 Data annotation

The sentiment classification problem requires consideration of both positive and negative discrimination. However, the method provided in reference [2] is too weak to label the negative comments, resulting in an unbalanced distribution of positive and negative samples. The research found that the main reason is that the lexicon is for general reviews, which does not contain network catchwords and new words in the field of film reviews, especially with negative emotions, such as "gou xue" which means melodramatic plots in films or novels, "cao dian" which means things complained about, etc. It is necessary to identify the unique new words in this field, and calculate the sentiment tendency to build a sentiment dictionary belonging to the film review field.

The discovery of new words is mainly divided into four parts:

1. Extract all possible candidate words in reviews and calculate the word frequency.
2. Set the longest word length to 5.
3. Calculate the agglomeration degree of each word.
4. Divide each candidate word into different combination pairs, and then calculate the agglomeration degree of them by formula, which is defined as

\[
D(s_1, s_2) = \frac{P(s_1s_2)}{P(s_1)p(s_2)}
\]  

(1)
where \( P(s_1) \) is the probability of \( s_1 \) appearing in the comments, \( p(s_1s_2) \) is the probability of \( s_1s_2 \) appearing in the comments. The minimum value of these combinations is selected as the agglomeration degree of the whole word combination.

3. Calculate the degree of freedom for each word.

The left neighbor information entropy and the right neighbor information entropy are used to measure the degree of freedom of the candidate words. The information entropy is defined as

\[
H(x) = - \sum_{i=1}^{n} p(x_i) \log(p(x_i))
\]

Similarly, the smaller value of the left neighbor information entropy and the right neighbor information entropy is selected as the degree of freedom of the candidate word.

4. Filter candidate words by comparing word frequency, agglomeration degree and degree of freedom, and output new words that meet the requirements.

After analysis, the three indicators of word frequency, agglomeration degree, and degree of freedom are all proportional to the possibility of forming a word. Considering the above indicators comprehensively, select candidate words with word frequency greater than 30, agglomeration degree greater than 100, and degree of freedom greater than 1.5. Finally, compare with the original lexicon to select new words.

Point-wise Mutual Information (PMI) can calculate the similarity between words. The similarity formula of two words \( w_1 \) and \( w_2 \) is:

\[
PMI(w_1, w_2) = \ln \frac{p(w_1, w_2)}{p(w_1)p(w_2)}
\]

Based on the above formula, the PMI of new words and emotional words can be calculated, and the emotional tendency of new words can be obtained. The formula is as follows:

\[
senti(w) = \sum_{w_p \in W_p} PMI(w, w_p) - \sum_{w_n \in W_n} PMI(w, w_n)
\]

where \( senti(w) \) represents the emotional tendency value of \( w \), \( W_p \) represents the positive emotional word set, and \( W_n \) represents the negative emotional word set. If \( senti(w) \geq 1 \), then \( w \) is a positive emotional word; If \( senti(w) \leq -1 \), \( w \) is a negative emotional word.

3.2 Feature selection

The word vector word2vec model maps each word to a fixed-size vector, which converts the text nicely into the input value of the model. The principle is to use the context of the current word to construct a word vector, which can measure the similarity between words. The model contains two different models: CBOW and Skip-Gram. In this paper, Skip-Gram is used for feature selection.

For text data, what we need to analyze is the entire text content, such as the entire comment data and the tendency of the overall content. This requires us to obtain the vector representation of the whole text based on words, namely sentence vectors.

This article uses the weighted vector of words to generate the sentence vector for each comment:

\[
sen_vec = \frac{\sum_{i=1}^{m} word_vec_i \cdot e^{w_i}}{m}
\]

where \( word_vec_i \) is the ith word vector of the current review, \( w_i \) is the corresponding weight, and \( m \) represents the number of word vectors contained in the current review. This paper uses TF-IDF as the value of \( w_i \).

The TF-IDF algorithm is completely based on word frequency statistics to calculate the weight of words, which is defined as
TF = IDF = TF * IDF

TF represents the frequency of words in the text, which is defined as

$$TF_{w_i} = \frac{f(w_i)}{\sum_{i \in I} f(w_i)}$$

where $f(w_i)$ is the number of times the word $w_i$ appears in the comment set, and $I$ represents the subscript set of all words in the comment set.

IDF is the frequency of reverse documents. The fewer comments contain the word $w$, the better the classification ability $w$ has. The formula is

$$IDF_{w_i} = \log\left(\frac{\sum_{i \in I} F(w_i)}{F(w_i) + 1}\right)$$

where $F(w_i)$ represents the number of comments containing word $w_i$ in the corpus, and $I$ represents the subscript set of all words in the corpus.

Each field contains unique language characteristics, so the words used to express emotions will be different, and the frequency of each word will be different. For example, words such as "box office", "new year film" can be said to be unique and frequent words in the field of movie reviews. Therefore, to better reflect the characteristics of this field, the corpus used by IDF in this article is constructed based on film review data.

### 3.3 Model Training

Ensemble learning is an essential and popular branch of machine learning. Its core idea is to train multiple weakly-supervised models in order to obtain a more comprehensive strong-supervised model. General weak classifiers can be composed of decision trees, neural networks, Bayesian classifiers, K-nearest neighbors, etc. Common ensemble learning strategies mainly include: Bagging, Boosting, Stacking and Blending.

This article chooses the Stacking method in the ensemble model. Stacking first trains the primary learner from the initial data set, and then combines the output of the primary learner into a new data set for training the secondary learner.

![Figure 1 Relationship between word vector and sentence vector](image1.png)
4. Experimental results and analysis

4.1 Experimental Data
This paper uses python to capture reviews and user ratings of 18 movies on Douban, including six movies with high, medium and low ratings. The number of reviews for each movie is kept at about 200. After cleaning and filtering, 3482 pieces of data were obtained as experimental data, including 1,749 reviews with user ratings of three stars and above, and 1,733 reviews with two stars and below. Integrating the National Taiwan University Simplified Chinese Sentiment Dictionary (NTUSD) and HowNet Chinese Sentiment Dictionary (HowNet) to form a basic dictionary, and adding the calculated new sentiment words to obtain a unique sentiment dictionary in the field of Douban film reviews. This dictionary contains 9,040 positive emotion words and 7,756 negative emotional words. After that, automatically annotate the above data with the dictionary and user rating information. Comparing the method without adding new words, the results are as follows:

| Dictionary                           | positive comments | negative comments | Adoption rate |
|--------------------------------------|-------------------|-------------------|---------------|
| Basic emotion dictionary             | 1627              | 636               | 65%           |
| Basic emotion dictionary + new       | 1723              | 751               | 71%           |
| sentiment words                      |                   |                   |               |

It can be seen from Table 1 that after adding new emotional words, the adoption rate of automatic labeling has increased significantly.

4.2 Experimental Results and Analysis
Experiment on the above 2474 pieces of valid data, taking 80% as the training set and the remaining 20% as the test set. To evaluate the performance of the algorithm in this paper, different weighting methods and classifiers are selected for three sets of comparative experiments. The evaluation indicators are accuracy, precision, recall, and F-score.

1) In order to find a better weighting method, use the unweighted word2vec method to classify the data, then use the word2vec method based on TF-IDF direct weighting or weighting according to
formula (5) to classify and compare the effect of sentiment classification. At the same time, considering the essential similarity between TextRank and TF-IDF, it is also added for comparison.

| weight | Accuracy | Precision | Recall | F-score |
|--------|----------|-----------|--------|---------|
| 1      | 0.7798   | 0.8367    | 0.8488 | 0.8427  |
| TF-IDF | 0.7899   | 0.8680    | 0.8227 | 0.8478  |
| TF-IDF | 0.8323   | 0.8518    | 0.9186 | 0.8839  |
| TextRank | 0.7980   | 0.8720    | 0.8314 | 0.8512  |

By comparing the performance of the four weighting methods, we can see that the weighting method proposed in this paper, that is, using TF-IDF as the value of the $w$ and $e^w$ as the weighting method, achieves a better classification effect. Its accuracy, precision, recall and F-score have been significantly improved.

2) Calculating based on different corpora will result in different TF-IDF, which will affect the final classification effect. In order to test that the improved corpus has improved classification performance, the following comparative experiments were done under the same weighting method and classifier:

| Corpus            | Accuracy | Precision | Recall | F-score |
|-------------------|----------|-----------|--------|---------|
| No New Word       | 0.8283   | 0.8776    | 0.875  | 0.8763  |
| Add New Words     | 0.8323   | 0.8518    | 0.9186 | 0.8839  |

It can be seen that the TF-IDF value calculated based on the improved corpus does improve the classification effect, and the accuracy, precision, recall and F-score have been improved.

3) In order to compare the effects of classifiers, the following experiments were done under the same weighting method and corpus:

| Classifier | Accuracy | Precision | Recall | F-score |
|------------|----------|-----------|--------|---------|
| Stacking   | 0.8323   | 0.8518    | 0.9186 | 0.8839  |
| SVM        | 0.8      | 0.8165    | 0.9186 | 0.8646  |

It can be seen from the experimental results that compared to the SVM classifier, the classification effect of Stacking is better.

5. Conclusion
This paper calculates new words and new emotional words in the film review field, and builds a corpus in this field, which improves the adoption rate of automatic annotation of review data. At the same time, TF-IDF is used as the value of the variable $w$ and $e^w$ is used as the weight of word2vec to integrate text features, and using Stacking for classification has achieved good experimental results. Of course, this method can also be extended to sentiment analysis in different fields, such as book reviews and drama reviews.

Moreover, we found that the captured data include not only the comments themselves, but also the id of the user, the time of the comment, and the number of likes. Through the research of this information, it may be possible to find the inner relationship and achieve a better classification effect, which needs further research and experiment in the future.
References

[1] Rui Li, Qian Zhang, Jia Yong Liu. Microblog sentiment analysis based on weighted word2vec[J], Communications Technology, 2017, 50(3): 502-506.

[2] Zhen Fan, Yi Guo. Sentiment analysis of movie reviews based on dictionary and weak tagging information[J], Journal of Computer Applications, 2018, 38(11): 3084-3088.

[3] Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? sentiment classification using machine learning techniques[J]. Empirical Methods in Natural Language Processing, 2002, 79-86.

[4] Yu Li. Microblog Emotional Dictionary Built and Application on sentiment[D], Zhengzhou University, Zhengzhou, 2014.

[5] Zhitao Wang, Zhiwen Yu, Bin Guo, et al. Sentiment analysis of Chinese micro blog based on lexicon and rule set[J], Computer Engineering and Applications, 2015, 51(8): 218-225.

[6] Kalaivani P, Shunmuganathan K L. Sentiment classification of movie review by supervised machine learning approaches[J]. Indian Journal of Computer Science & Engineering, 2013, 4(4).

[7] Yan Wang, Xuteng Wang, et al. Sentiment Analysis of Weibo Movie Reviews Based on Support Vector Machine[J]. Advanced Motion Picture Technology, 2019, (8): 41-45.

[8] Lei Zhang, Bing Liu. Identifying noun product features that imply opinions[C], Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Cambridge, MA: MIT Press, 2011: 575-580.

[9] Minqing Hu, Bing Liu. Mining and summarizing customer reviews[C], Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168-177.