Chinese Movie Comment Sentiment Analysis Based on HowNet and User Likes

Yanqiu Liu\textsuperscript{1,2} and Zekun Duan\textsuperscript{1}

\textsuperscript{1} Computer Science Department of Inner Mongolian University, Hohhot, China

\textsuperscript{2} E-mail: 651220506@qq.com

Abstract. Emotional information in movie comments is critical to sentiment analysis. Sentiment analysis, which focuses on classify the comments into positive class and negative class according to sentiment lexicon, is one of the studies. Most of the existing researches are centered on sentiment words and user rating, while the user’s attitude towards comments are ignored. And, considering that Chinese is the second largest language in the world. In this paper, in order to get this point to be considered, we propose a method for Chinese movie comments sentiment analysis based on HowNet and user likes which we called HAL. Our research consists of four parts. First, we use HowNet sentiment lexicon to get a new lexicon in the field of movies. Second, we use the new lexicon and word segmentation tool named Jieba to segment the movie comments. Third, we use the user likes and sentiment words to get the positive feature and negative feature. Finally, we train the movie comment data using three models (SVC, LinearSVC, Logistic Regression). The experimental results show that our method performs better than HowNet-based method in Chinese movie comment sentiment analysis.

1. Introduction

With the rapid development of the Internet, watch movie and comment on the Internet is becoming more and more popular. In order to improve user’s experience, many movie websites allow them to comment on movies. Therefore, there are a lot of movie comments on the website. It is difficult for users to distinguish between good and bad movies, while sentiment analysis can divide movie comments into positive and negative. Different from the traditional topic-based classification method, it can only be identified by keywords, while sentiment analysis needs to extract more implicit information\cite{1}. Therefore, lexicon-based method\cite{2} has been proposed, and many basic lexicons have emerged, such as WordNet and HowNet.

At present, in China, there are relatively few studies on movie comments. Such as research based on machine learning and semantic orientation (Chaovalit P et al.\cite{3}) and method of extracting feature word pairs (Zhuang L et al.\cite{4}). Some researchers believe that the emotional polarity of the comment is mainly judged by the adjectives in the commentary. It is proposed to establish an sentiment lexicon based on the adjectives, and then calculate the emotional polarity of the comments according to certain rules, but this method relies too much on the quality of sentiment lexicons and rules, requires a certain amount of empirical knowledge, and has poor generalization ability (Hu et al.\cite{5}). In this paper, we propose a Chinese movie comment sentiment analysis method based on HowNet and user likes (HAL). The experimental results show the performance of our method is better than HowNet-based method.

The rest of this paper is organized as follows. Section 2 reviews related work on sentiment...
classification and custom lexicon. Methodology is described in Section 3. The experimental results are illustrated and analyzed in Section 4. Finally, conclusion are described in Section 5.

2. Related work
In this section, we give a brief introduction to the custom lexicon and the previous work on the sentiment lexicon-based methods and machine learning-based methods for sentiment classification.

2.1 Sentiment classification based on sentiment lexicon
This method is mainly based on the emotional lexicon to match the emotional polarity of the calculated comments. (Liu et al.[6]) considered the influence of the distance between the subject words and the emotional words on the emotional sentiment of the comments. The farther the distance is, the weaker the impact on the emotions. However, there is no way for the sentiment lexicon-based approach to identify the ideas implicit in a comment. In addition, in movie reviews, there are often cases that have totally different sentiment between first half sentence and second half sentence, sometimes the same emotional words have different polarities in different environments.

2.2 Sentiment classification based on supervised machine learning
The research method based on machine learning mainly transforms sentiment analysis of comments into a classification problem, including supervised learning and unsupervised learning (Pandarachalil R et al. [7]). At present, the mainstream is the dichotomy, which divides comments into positive and negative emotions. Text can be classified by manually designing features that represent the text of the review, then extracting the feature features and representing them as text vectors. For example, On the basis of previous work, (Li et al. [8]) propose to use parts of speech, emotional words, negatives and degree adverbs to construct features, and select different combinations of features to carry out a series of experiments. Their experimental results show that SVM works best when using part of speech tagging, negative words and emotional words, while conditional random field (CRF) model works best when using emotional words, degree adverbs, negative words and special symbols.

2.3 Custom lexicon
Because, in movie comments, some sentiment words with emotional tendencies are not recorded in the basic dictionary. For example, words like "Nao3can2fen3", "Gao3xiao4" and "Gei3li4" in Chinese are all positive emotions. It is not enough to identify the emotional tendencies of movie comments only by using basic lexicons (HowNet). Therefore, a lexicon of this field is also needed.

2.3.1 Constructing a lexicon of movie domain. In this paper, we use Point-wise Mutual Information (PMI)[9] to construct a lexicon of movie domain. PMI can calculate the similarity between \( w_1 \) and \( w_2 \). The formula(1) is as follows:

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

(1)

Where \( p(w_1, w_2) \) is the probability of \( w_1 \) and \( w_2 \) appear in same sentence; \( p(w_1) \) is the probability of \( w_1 \), \( p(w_2) \) is the probability of \( w_2 \). \( w_1 \) is an emotional word segmented from the movie comments, \( w_2 \) is the core emotional word in HowNet. Computing the similarity of two words by PMI. If \( p(w_1, w_2) \) is high, the two words are considered to have the same emotional polarity. If \( p(w_1, w_2) \) is low, the two words are considered to have the different emotional polarity.

After calculating the PMI between the new words and the positive and negative core sentiment words, we can get the emotional tendency of the new words. The formula(2) is as follows[10]:

\[
senti(w) = \sum_{w \in W_P} P(w, w_P) - \sum_{w \in W_N} P(w, w_N)
\]  

(2)

Where \( senti(w) \) is the emotional tendency value of new sentiment word \( w \), \( W_P \) is the set of
positive sentiment words and \( W^- \) is the set of negative sentiment words. if \( \text{senti}(w) > 0 \), \( w \) is positive sentiment word; if \( \text{senti}(w) < 0 \), \( w \) is negative sentiment word. In this way, we get a lexicon of sentiment in movie domain.

2.3.2 Negative word set and degree word set. Negative words in a comment can directly change the emotional polarity, while degree words can strengthen emotional polarity to a certain extent. Therefore, we re-judges the emotional polarity of comments based on the position and frequency of negative words before and after sentiment words, and weights the emotional polarity to a certain extent according to the intensity of the degree words. The common negative words as shown in Table 1 and degree words as shown in Table 2:

| Table 1. Negative words |
|-------------------------|
| Negative words          | Weight |
| Bu4, Mei2, Fei1, Wei4, Fou3, Wu4, Bu2Shi4 et al. | -1      |

| Table 2. Degree words and weights |
|----------------------------------|
| Level   | Degree words                  | Weight \((w^1)\) |
|---------|-------------------------------|------------------|
| Zui4    | Bai3 Fen1 Zhi1 Bai3, Fei1 Chang2, Ji2 | 1.5              |
| Chao1   | Guo4 Du4, Chao1 E2, He2 Zhi3   | 1.2              |
| Hen3    | Duo1 Jia1, Duo1 Me, Fen4 Wai4, Ge2 Wai4 | 1.1              |
| Jiao4   | Geng4 Wei2, Hai2, Hai2 Yao4, Jiao4 | 0.75             |
| Shao1   | Shao1 Xu3, Yi4 Dian3, Lve4 Wei1 | 0.5              |
| Qian4   | Ruo4, Si1 Hao2, Bu4 Ding1 Dian3 | -0.5             |

3. Methodology
This paper proposes a method of sentiment analysis of movie comments based on HowNet sentiment lexicon and user likes. On the basis of user likes, comments are categorized by lexicon of movie domain, and selects training data models with the same score and classification results. The general framework of our work is shown in Figure 1.

![Figure 1. The general framework of our work](image)

3.1 Data label based on custom lexicon
The emotional classification problem based on machine learning is mainly supervised classification
problem, which requires manual tagging of data, time-consuming and labor-consuming. The data collected in this paper contains user ratings, so it can be initially labeled according to user ratings. This is called weak tagging information [11]. But there are deviations in user ratings (user ratings do not match the emotions of the comments). So, in this paper, we strengthens this kind of weak tagging information by user likes, and carries out emotional analysis of movie comments with a custom lexicon.

3.2 User likes and weight
The quality of a comment is often visualized by the likes of other users. If a comment receives a certain amount of praise, it will strengthen the emotions and views of the comment to a certain extent. Therefore, in this paper, we weight the user likes of each comment. It can be divided into the following Table 3:

| User likes(n) | Weight(\(w_2\)) |
|--------------|-----------------|
| \(n \geq 1000\) | 2               |
| \(1000 > n \geq 100\) | 1.5             |
| \(100 > n \geq 0\) | 1.1             |

Sentiment analysis method in this paper:
- Comment preprocessing: Word segmentation and part-of-speech tagging of comments Using the Jieba word Segmentation Tool.
- Calculate the emotional tendencies of comment\((j)\): According to the location of the emotional word \((a)\), the number of negative words \((m)\), the weight of degree word \((w_1(b))\) and the weight of user likes \((w_2(c))\) in comment\((j)\). We can calculate the emotional value \((sentiment(j))\) of a single comment by the formula (3) as follows:

\[
sentiment(j) = \sum_{a \in \text{dict}} (-1)^m \text{senti}(a) w_1(b) w_2(c)
\]

- If \(sentiment(j) > 0\), the sentiment is positive; if \(sentiment(j) < 0\), the sentiment is negative.

3.3 Feature selection
One point of sentiment analysis based on machine learning is feature selection, which is related to the accuracy of the emotional classification. In this paper, we choose degree adverbs, negatives, user likes, positive emotional word and negative emotional word as representative features. One comment is composed of many words and their parts of speech, in which part of speech plays a very important role; emotional words are the key core of emotional classification of a text, while negative words often reverse the emotional polarity of a text; at the same time, degree adverbs can change the intensity of emotional words. When there are both positive and negative emotional words in a text, it is difficult to judge the emotional tendency of the text by relying solely on the polarity of the emotional words, and negative words and the adverb of degree can help to make a choice. And the user likes can reflect the effectiveness of comments. The algorithm is shown in Table 4.

| Algorithm: Sentiment analysis based on custom lexicon and user likes |
|---------------------------------------------------------------|
| 1: \text{dict\_set} \leftarrow \text{sentiment words in custom lexicon} |
| 2: \text{neg\_set} \leftarrow \text{negative words} |
| 3: \text{degree\_set} \leftarrow \text{degree words} |
| 4: \textbf{for} a is in \text{dict\_set} \textbf{do} |
| 5: \textbf{if} b is in \text{neg\_set} \textbf{or} c is in \text{degree\_set} \textbf{then} |
| 6: sent = senti(a) w_1(b) w_2(c) |


7:     else
8:         senti(a)
9:     end if
10: sum(sent)
11: end for

4. Experiment and analysis

4.1 Data Sets
More than 100,000 Chinese comments on movies are selected from Douban. After the removal of duplicates and meaningless data, we get 80,000 comments.

According to the research needs of this paper, the data set can be divided into positive comments and negative comments. After sorting out, 10,000 positive comments and 10,000 negative comments were obtained respectively. Among them, 1000 comments including user likes were selected for the experiment. From the remaining comments, we take the equal number of comment used as comparative experiment. For the experiment, the data set is divided into two parts. One part has 70% positive and negative comments for training and the other 30% for testing. A brief introduction of data sets is shown in Table 5.

| User likes(Y/N) | Positive | Negative |
|----------------|----------|----------|
| Total          | 10000    | 10000    |
| Y              | 1000     | 1000     |
| N              | 1000     | 1000     |

4.2 Experimental results
We conduct the experiment to evaluate our work of sentiment classification. In this section, the experimental results are shown as followed.

4.2.1 Method. In this paper, we use SVC, LinearSVC and LogisticRegression training model. SVM is a new classification method developed in recent years, which can be used for classification and regression problems[12]. LogisticRegression can also be used for classification problem[13].

4.2.2 Performance of sentiment classification. In this subsection, the sentiment classification performance is shown and discussed.

| Model                | F1(%) | Precision(%) | Recall(%) | Accuracy(%) |
|----------------------|-------|--------------|-----------|-------------|
| SVC                  | 84.76 | 78.53        | 92.05     | 82.88       |
| LinearSVC            | 68.35 | 65.45        | 71.52     | 65.75       |
| LogisticRegression   | 68.79 | 66.26        | 72.51     | 66.44       |

| Model                | F1(%) | Precision(%) | Recall(%) | Accuracy(%) |
|----------------------|-------|--------------|-----------|-------------|
| SVC                  | 82.84 | 74.87        | 92.72     | 80.14       |
| LinearSVC            | 72.68 | 62.95        | 85.97     | 63.70       |
| LogisticRegression   | 74.38 | 67.84        | 82.31     | 68.15       |

The performance of HAL-based method in Table 6, the performance of HowNet-based method in Table 7. From the above two tables, we can find that under the same method, the performance of SVC
model is better than other models and the method that we proposed shows the better performance than HowNet-base method. The reason why HAL is better than the HowNet is probably that HowNet features do not contain user likes. The performances of HAL-based method and HowNet-based method are shown more intuitively in Figure 2.

![Figure 2. The performance in three model](image)

4.3 Analysis
From the experiments above, it is illustrated that both HowNet-based method and HAL-based method can get excellent results. But, the proposed method in this paper achieves a better performance. Then, we discuss the advantage of our method.

Compare with the HowNet-based method, our approach has the following advantages. Firstly, in the part of lexicon selection, we use PMI to choose specific sentiment words in the movie field, which is the key to construct the lexicon of movie field. Therefore, we can segment the movie comments more accurate. Secondly, we use the negative words and degree words to get more accurate sentiment value of movie comments. Finally, we use use likes to strengthen the emotional weight of comments. Above all, the method we proposed can perform a little better than HowNet-based method in movie comments sentiment analysis.

5. Conclusion
In this paper, the emotional classification of Douban movie comments is studied, and a method based on HowNet and user likes (HAL) is proposed for emotional classification of movie comments. Three models are used to test our method, and the performance were higher than using HowNet only. But there are still some shortcomings. For example, there is insufficient comment data including user likes; there is room for improvement in setting some weights; whether this method can be applied to other fields. In future research, we will further improve the above-mentioned point, which will make the subject richer and more valuable.

Reference
[1] Zhai Z, Bing L, Xu H, et al. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints.[C]/// The. 2010.
[2] Taboada, Maite, et al. "Lexicon-Based Methods for Sentiment Analysis." Computational Linguistics 37.2(2011):267-307.
[3] Pimwadee Chaovalit, Lina Zhou. "Movie review mining: a comparison between supervised and
unsupervised classification approaches." Hawaii International Conference on System Sciences IEEE, 2005.

[4] Zhuang, Li, F. Jing, and X. Zhu. "Movie review mining and summarization." (2006).

[5] Hu, Minqing, and B. Liu. "Mining and summarizing customer reviews." Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004 ACM, 2004.

[6] Ding, Xiaowen, B. Liu, and P. S. Yu, "[ACM Press the international conference - Palo Alto, California, USA (2008.02.11-2008.02.12)] Proceedings of the international conference on Web search and web data mining, - WSDM \"08 - A holistic lexicon-based approach to opinion mining." (2008):231.

[7] Pandarachalil, Rafeeqe, S. Sendhilkumar, and G. S. Mahalakshmi. "Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach." Cognitive Computation 7.2(2015):254-262.

[8] Ting-Ting, L. I., et al. "Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features." Application Research of Computers (2015).

[9] Khan, Farhan Hassan, U. Qamar, and S. Bashir. "SentiMI: Introducing Point-wise Mutual Information with SentiWordNet to Improve Sentiment Polarity Detection." Applied Soft Computing 39(2016):140-153.

[10] Zhtao W, Zhiwen Y U, Bin G, et al. Sentiment analysis of Chinese micro blog based on lexicon and rule set[J]. Computer Engineering & Applications, 2015.

[11] Yu-Xia and Z. Evaluation Metrics for Recommender Systems[J]. Journal of University of Electronic Science and Technology of China, 2012.

[12] Cortes, Corinna, and V. Vapnik. "Support-vector networks." Machine Learning 20.3(1995):273-297.

[13] Komarek, P. Logistic Regression for Data Mining and High-Dimensional Classification. 2004.