An Improved SVM for Book Review Sentiment Polarity Analysis

Xinxin Guan [a], Yeli Li [b], Hechen Gong [c], Huayan Sun [d] and Chufeng Zhou [e]

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China.

[a] 1145512971@qq.com, [b] liyl@bigc.edu.cn, [c] gonghechen123@163.com, [d] 603757282@qq.com, [e] 576213376@qq.com

Abstract. In the internet age, whether a book has the value of reading, online comments play an important role. The data set in this paper is 4,000 comments obtained by the web crawler in Douban Reading. Based on the improved support vector machine (SVM) algorithm, a sentiment analysis has been given to these comments. The experimental results show that the improved SVM algorithm has a good effect on the rate and accuracy of sentiment polarity analysis of book reviews.

Keywords: Support vector machine, Sentiment polarity analysis, Book review.

1. Introduction

With the rapid development of e-commerce, people's lives have become more convenient, a lot of information about book reviews has been produced as well. The sentiment polarity analysis of such information can help form a win-win relationship for both businessmen and users. Merchants can find out the quality of books through the comments on them and improve the quality of their stores. Users can judge the contents of the books according to the comments. Therefore, the sentiment analysis has an important academic research value and practical application value.

Text sentiment analysis refers to the process of analyzing and collating texts with personal subjective opinions, which involves text categorization and information extraction etc. [1]. As a new field of data mining, it has an important academic research value and practical application value, and Web user comment mining is an important research field of sentiment analysis.

In the aspect of sentiment analysis, both English and Chinese texts sentiment analysis have been greatly developed. The text information, especially the analysis of sentiment tendency on online comment texts has been made a significant breakthrough. At present the accuracy of sentiment tendency analysis based on the online comments has reached more than 90%.

In the text sentiment analysis based on dictionary, Hu and Liu judge the sentiment tendency by comparing the synonyms and synonyms between different words in the WordNet dictionary, and use it to tell the sentiments of t viewpoint holders [2]. Turney and others use a simple unsupervised learning algorithm to classify the comment texts [3]. Kim and Hovy combine the sentimental attributes of sentimental l words in a separate sentimental dictionary, the emotions are divided into three categories, and three models are used to identify the three types of sentiments [4]. This document has achieved a system of sentiment classification, which can automatically query the user's sentimental attitudes and suggestions on a given topic.

There are also relevant scholars who build the sentiment dictionaries in different fields, such as Zhou Xingmei and other scholars use the graphic sequence model for reference and coeme up with a method of building news comment sentiment dictionary [5]. Based on the improved Hevner sentiment ring model, Jiang Shengyi and others use the semantic resources provided by hownet and the lyrics text corpus crawled from the webto build a Chinese sentiment dictionary in the music field [6]. There are other sentiment dictionaries related to different fields, such as hotel reviews sentiment dictionary [7], Weibo product reviews sentiment dictionary [8], film comments sentiment dictionary [9] and so on.

In the above-mentioned algorithms and methods, the applicability and accuracy of certain foreign research methods on the Chinese text sentiment analysis are not so good, And Chinese text dictionaries and corpora have not been so perfect and unified, the text sentiment judgement accuracy rate is not very high as well. In this paper, the author uses the improved SVM classification algorithm to analyze and study the readers ' comments, combining Bayes to excavate the features of the book comments on Douban Reading. These results have important reference value for merchants. In this paper, the sentiment analysis can understand the questions of text classification. The data set uses 2 classification, that is, only positive and negative evaluation.

2. The Construction of the Emotional Dictionary of Book Reviews

2.1 Algorithmic Framework.

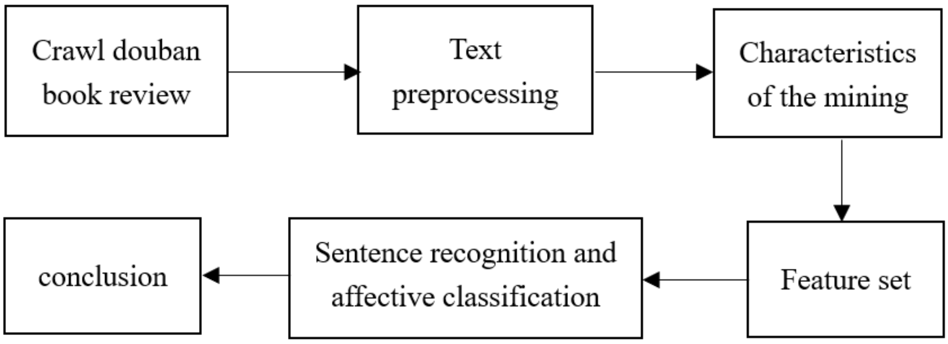

Crawl douban book review
Text preprocessing
Characteristics of the mining
Feature set
Sentence recognition and affective classification
conclusion

Fig 1. Algorithm Framework diagram

Preprocessing: The blank spaces in the original user comment, such as those that do not contain emotional information, must be filtered out before the word breaker is used for word segmentation and speech tagging of user comments.

Fig. 1 is the algorithm framework diagram for this article. Feature mining module digs out the characteristics of readers ' attention, and makes feature filtering by finding and merging synonyms. The view sentence recognition and Emotion classification module identifies the viewpoint sentence according to the feature set of mining, and classifies the emotional polarity by the emotion dictionary matching method according to the reader's comment feature.

2.2 Key Algorithms.

2.2.1 SVM Algorithm

Support vector machine (SVM), as a supervised learning method, is often used for two classification problems [10]. Its main idea is to solve the problem of two classification, look for a super plane as a two-class training sample point segmentation, to ensure the minimum classification error rate. In the case of linear separable, there is one or more hyper-planes that make the training samples completely separate. The goal of SVM is to find the optimal super plane, the optimal hyper-plane is the plane with the largest distance between the nearest vector and the super plane of each type of data, and for linear irreducible conditions. By using the kernel function, the linear non-fractal sample of the low-dimensional input space is transformed into a high-dimensional feature space, which is linearly divided.

Basic model of SVM [11]. Set the input mode set $\{x[i]\} \in R^n$ consists of 2 types of points, if $x[i]$ belongs to class 1th, the $y[i] = 1$; if $x[i]$ belongs to class 2nd, the $y[i] = -1$, Sample with training set $\{x\ [i], y\ [i]\}\ (i = 1,2,3\ldots, n)$ finding the optimal classification surface $wx - b\ = 0$, Meet: $y[i]\ (w \cdot x\ [i] - b) \geq 1$ and $2xh = 2/||w||$ is maximum. So min $||w|| \times ||w||/2$. According to the duality theory, the optimal solution can be obtained by solving the duality problem of the problem.

The duality problem is:

$$max \sum \alpha[i] - 1/2 \sum \alpha[i] * \alpha[j] * y[i] * y[j] * x[i] * x[j]. \quad (1)$$

Constraint conditions:

$$0 \leq \alpha[i] \leq C \ \ and \sum \alpha[i] * y[i] = 0 \tag{2}$$

Where $x[i] * x[j]$ denotes the inner product of these two vectors, when for a linear irreducible condition, Replace $x[i] * x[j]$ with kernel product $x[i] * x[j]$ (the kernel function maps to the inner product of the corresponding vector in the high dimensional space). According to the solution of the duality problem α . Get w, b and the optimal classification surface.

2.2.2 Bayes Theorem

The probability of event A in the condition where event B occurs is not the same as the probability of event B occurring under event A, however, there is a definite relationship between the two, and the Bayes theorem is the statement of the relationship. The formula is as follows:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{3}$$

The purpose of Bayes is to speculate on a fourth probability by knowing three probabilities. Its content is: On the premise that B occurs, the probability of the occurrence of A is equal to the probability when B occurs on the premise that A also occurs, multiplied by the probability of the occurrence of A and then divided by the probability of the occurrence of B. By contacting A and B to calculate the probability of another event occurring from an event, that is, from the result to the source (also known as inverse probability).

In layman's words, when you cannot determine the probability of an event occurring, you can rely on the probability of an event that is related to the nature of the event to speculate on the probability of the event occurring. The expression in mathematical language is that the more events that support an attribute, the greater the likelihood that the event will occur. This reasoning process is sometimes called Bayes reasoning.

Naive Bayes method is a classification algorithm based on Bayes theorem and independent hypothesis of characteristic condition. Give you the data set, and then assume that each event is relatively independent, the algorithm learns the joint probability distribution between the input/output of this dataset, using this model, given an x, to output the maximum probability. Play a role in forecasting.

2.3 A Dictionary Matching Technique for Sentiment Analysis.

The key of emotion dictionary matching technique is the construction of sentiment dictionary and the design of matching algorithm. In order to improve the accuracy of classification, it is necessary to add these special words to the sentiment dictionary to construct the Emotion dictionary in the book field. Book Emotion Dictionary should include: Basic Emotion dictionary, Network Hot Word dictionary, emoji Dictionary, negative Word dictionary, degree adverb dictionary.

The basic sentiment dictionary consists of a positive basic emotion dictionary and a negative basic emotion dictionary. The positive sentiment words, the evaluation words and the words in the Chinese emotion lexical ontology library with the polarity of "1" are combined to HowNet, and the words with the less sentiment tendency are composed of the positive Basic sentiment dictionary; The negative emotion words, the evaluation words and the Chinese emotion lexical ontology in the HowNet, are combined with the word "2". and remove the non-significant emotional tendency of the entry to form a negative basic emotional dictionary. The final form of the basic Emotion dictionary contains 5,821 positive emotion words, 10,186 negative emotion words.

Constructing Special Sentiment Analysis dictionary in the field of books The SVM algorithm is used to establish the relationship between the two words and to predict the sentiment points of the target words according to the point mutual information between the target words and the reference words. Bayes algorithm is used to calculate the difference between the positive and negative mutual information of the target Word and the reference word, and the difference is greater than 0 for the positive sentiment word and vice versa.

3. Research Methods

3.1 Experimental Process.

The data set is a 4,000-piece data that has been processed from the watercress, first preprocessing the data and dividing the book review into 12 different fields, such as fiction, history, biography, reasoning, fantasy, etc. Secondly, it constructs the dictionary library in the field of book review. Finally, the improved SVM algorithm is used to classify the viewpoint and emotion, and compare it with the traditional algorithm.

The centralized distribution of datasets is shown in Fig 2, with the horizontal axis representing 12 different fields, and the ordinate represents the number of positive and negative data in each field.

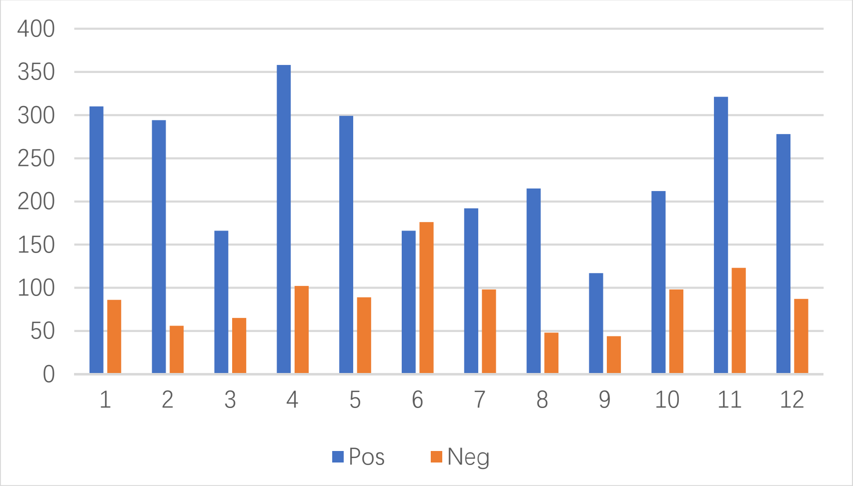
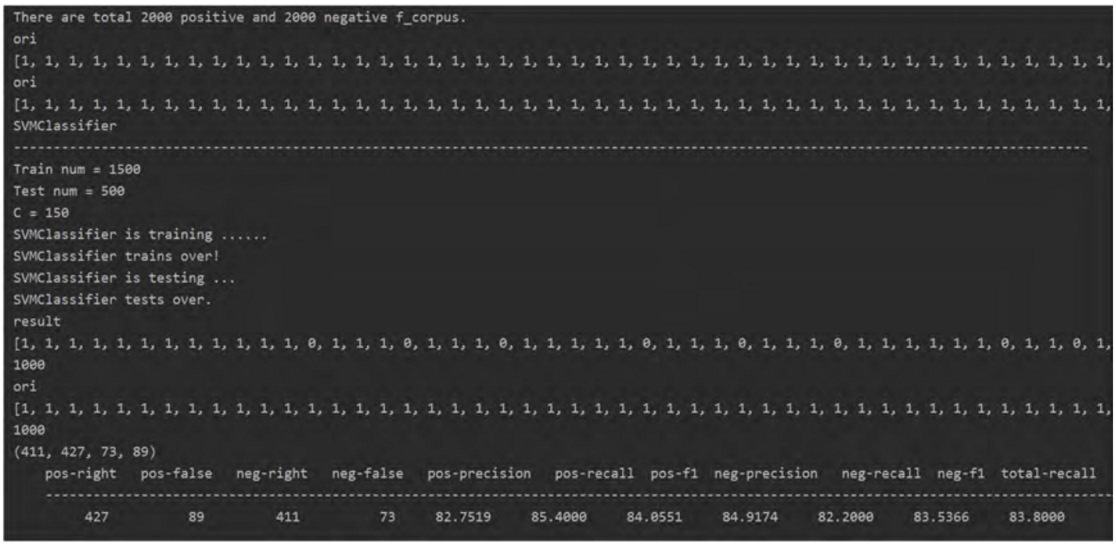


Fig 2. Distribution of data sets

3.2 Experimental Results.



```
There are total 2000 positive and 2000 negative f_corpus.
ori
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
ori
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
SVMClassifier
----------------------------------------------------------------------------------------------------------------------------
Train num = 1500
Test num = 500
C = 150
SVMClassifier is training ......
SVMClassifier trains over!
SVMClassifier is testing ...
SVMClassifier tests over.
result
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1,
1000
ori
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1000
(411, 427, 73, 89)
    pos-right   pos-false   neg-right   neg-false   pos-precision   pos-recall  pos-f1  neg-precision   neg-recall  neg-f1  total-recall
    ----------------------------------------------------------------------------------------------------------------------------------
        427          89         411          73      82.7519      85.4000      84.0551     84.9174      82.2000      83.5366     83.8000
```

Fig 3. Experimental results diagram

3.3 Results Analysis.

Traditional SVM has high accuracy but long training time. Bayes algorithm has the advantages of simple, high efficiency, fast operation speed and good expansibility. The accuracy of Bayesian and SVM is improved, and the training time is short.

4. Summary

This paper proposes a method of constructing sentiment dictionaries for Chinese book reviews, which divides user sentiment into positive and negative categories in Chinese book reviews, and proposes an improved SVM algorithm to discriminate the sentiment categories in the Chinese book commentary field. It is proved by contrast experiment that the construction method proposed in this paper has a faster speed, better accuracy and more reliability.

Acknowledgements

Thanks to Beijing science and technology innovation service capability construction project (PXM2017_014223_000055), Beijing Municipal Education Committee project (KM201710015004) and Beijing Institute of Graphic Communication 2018 R&D Project (Ec201802).

References

[1]. Zhao Y, Qin B, Liu T. Text sentiment analysis. Journal of Software Science.Vol.21(2010) No. 8, p. 1834-1848.

[2]. Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, p. 168-177.

[3]. Turney P D. Thumbs up: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, p. 417-424.

[4]. Kim S M, Hovy E. Determining the sentiment of opinions. Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004, p. 1367-1373.

[5]. Zhou Y, Yang A, Yang J. Construction Method of Sentiment Lexicon for New Reviews. Computer Science, Vol. 41(2014) No. 8, p. 67-69, 80.

[6]. Jiang S, Yang Y, Liao J. Research of Building Chinese Musical Emotional Lexicon and Emotional Classification. Computer Engineering and Applications, Vol. 50(2014) No.24, p. 118-121, 163.

[7]. Yang A M, Lin J H, Zhou Y M, et al. Research on Building a Chinese Sentiment Lexicon Based on SO-PMI. Applied Mechanics and Materials, 2013, p. 263-266.

[8]. Yu Z. Research on the Key Technologies of Chinese Online Product Review's Sentiment Analysis. Hangzhou: Hangzhou Dianzi University, 2011.

[9]. Li M. Emotion Classification for Weibo Movie Reviews. Kunming: Yunnan Finance University, 2014.

[10]. Harrington P. Machine Learning in Action. Machine learning in action. Manning Publications Co. 2012.

[11]. Fan G. Research and design of network text emotion classification system based on SVM. Computer Age, 2015, p. 34-37.