Comments Prediction Model on Emotional Analysis Based on Bayes Classification

Xuegang Chen, Sheng Duan* and Luda Wang

College of software and communication engineering, XiangNan University, Chenzhou, China
Corresponding Author:Sheng Duan

Abstract. The comments of e-commerce platform are the information feedback of users on the purchase of goods, which shows the concerns and emotional tendencies of the commodity. The research has important research value for enterprises. In this paper, we constructed the analysis model of e-commerce comments. Firstly, reptile obtained data to segment word segmentation, and then use the characteristics of the initial selection model to review data and transfer matrix conversion, using Bayes probability model to prediction work, finally through the prediction error to select the optimal characteristics of the model. Through the empirical study of the commentary data of BOE, the results show the method is effective, and has better performance and higher classification accuracy.

1. Introduction
With the rising of the Internet era, e-commerce is also evolving, a lot of e-commerce platform accumulate a large number of data. Although there are a lot of data, it is difficult for customers to choose good products when choosing goods, and e-commerce enterprises are also not easy to locate the customer's interest. However, contain the emotional trend of the customer in the comments data of goods, How to analyze the information accurately and extract the key topics, has become an important part of our research. On the one hand, it helps enterprises to design the user's marketing model, and provide selling quantity and product services for enterprise. On the other hand, users can analyze the subjective emotional content of the comments, understand and compare their interest in products or services, and make the appropriate purchase decision. Therefore, how to extract the emotional information from the commentary has a practical application value, it can help users make decision, not only get feedback on the product, but also to predict the sales of the product.

Emotional computing has become a hot field of human-computer interaction, and its goal is to analyze emotional trends and categories by calculating. The most convenient and rich text resource on the Internet is increasingly becoming an important resource for emotional analysis, and the emotional information has broad application prospects in many ways. Fox example, public opinion analysis, product reviews and personalized recommendations and so on.

There are three aspects of the research methods of emotional analysis at home and abroad, and they are descriptive statistical analysis, feature correlation analysis, and machine learning analysis respectively[1]. The tendency of emotional is analyzed mainly using the linear analysis, vector transformation in descriptive statistical analysis, and the semantics of the comments are identified and classified for commendatory and derogatory. Machine learning analysis uses Bayesian and support vector machine to predict and evaluate. There are some researches in feature correlation analysis, for example, study lexical for the microblog data, and analyze the feature relevance establishing the corpus[2]. The correlation calculation model is constructed in order to analyze comments in an...
automobile[3]. There are some researches in machine learning. For example, analyze the text of appraising by support vector machine[4]. However, there are still some shortcomings in the current research technology, for example, emotional dictionary is not perfect, vocabulary is not compatible professional field, the identify the tendency algorithm is not accurate and scalability is poor, etc.

In view of the above problems and research current situation, product evaluation and rating of Jingdong e-commerce platform are taken as a case, and study the model of user’s reviews emotional trends and predictive, and the validity of the model is verified by the comments data of iphone7.

2. Related Theory and Technology

2.1. Text Processing Technology

We can be able to extract the effective content using the regular expression from the page, and divide the document into sentences, and divide the sentence into words, remove punctuation and stop words, and get no noise phrases, sentences. The objects are analyzed from an emotional point of view. String is parsed using JSON, and its code is as follows.

```java
public static String getJsonOfContent(String url) throws IOException {
    StringBuffer sb = new StringBuffer();
    FileInputStream fis = new FileInputStream(url);
    byte[] bbuf = new byte[3050];
    int hasRead = 0;
    while ((hasRead = fis.read(bbuf)) > 0) {
        sb.append(new String(bbuf, 0, hasRead));
    }
    fis.close();
    return sb.toString();
}
```

Mechanical participle method is used[5], and analyzed Chinese character string, and match with a "full big" machine dictionary according to a certain strategy, if a string is found in the dictionary, the match is successful. According to the scanning direction, the string matching word segmentation method can be divided into forward matching and reverse matching. These can be divided into the largest match and the smallest match by the prior matching of the different length[6-9].

The word segmentation system is a means for the mechanical word, and improves the accuracy of segmentation by the integrated approach.

2.2. TF-IDF Word Frequency Statistics

Word frequency statistics is a vocabulary analysis method, the word frequency is counted, analyzed for the text of a certain length, and find out the word rules.

The main principle of TF-IDF is as follows. If a word or phrase appears in a document and the frequency TF is the higher, and rarely appeared in other articles, and it is considered that the word or phrase has good classification ability, and it is suitable for classification. TF-IDF is actually: TF * IDF, where, Term Frequency (TF) shows that the term appears the number in the document, and Inverse Document Frequency (IDF) is the bigger if documents that contain the term t is the fewer, and it means that t has a good class distinction.

If the number of containing the term t in the documents C is m, and the total number of containing t documents in the other classes is k, obviously all documents containing t are n = m + k. get the IDF value will be a smaller when m is the larger, n is also the larger, according to IDF formula, and it means that the term t class distinction is not strong. TF-IDF often uses to filter out common words and keep important words. The calculation steps are described as follows.

Step 1: Calculate term frequency, Term Frequency(TF)=total number of times that a word appears in a document.
Step 2: Calculate the inverse document frequency, IDF = log(N/(M+1)),where, N is the total number of documents, and M is the number of documents that contain the term.
Step 3: Calculate the TF-IDF value, TF-IDF = TF * IDF.
Step 4: Find the keyword, The TF-IDF values of each word are sorted, and select the keyword of the former n item.

Calculate the keywords of each document, select the same number of keywords, merge into a set, and generate word frequency vector.

2.3 Bayesian Classifier

Bayesian classification algorithm is a statistical classification method, and it is a classification algorithm using probability and statistics knowledge, and the algorithm can be applied to large data, and the method is simple, classification accuracy is high and fast. The principle is to obtain the prior probability of an object, calculate its posterior probability using the Bayesian formula. The algorithm is described as follows.

**STEP1:** We describe the value of n attributes by an n-dimensional feature vector in each data sample, that is, \( X = \{x_1, x_2, \ldots, x_n\} \), and suppose that there are m classes, express with \( C_1, C_2, \ldots, C_m \), respectively. Given an unknown data sample A without class labels, assign the unknown sample A to the class \( C_j \) using the naive Bayesian classification, has \( P(C_j | X) > P(C_i | X) \) for all \( 1 \leq j \leq m, j \neq i \).

**STEP2:** If assign the unknown sample A to the class \( C_j \) using the naive Bayesian classification, has \( P(C_j | X) \) can be obtained from the training data set.

**STEP3:** According to the Bayesian theorem, \( P(A) \) is constant for all classes, and maximizing posterior probability \( P(C_j | A) \) can be transformed into maximizing the prior probabilities \( P(A | C_j) P(C_j) \). If the training data set has many attributes and tuples, time spending of calculating \( P(A | C_j) \) may be very large. Therefore, it is common to assume that the value of each attribute is independent each other, and prior probability \( P(A_1 | C_j) P(A_2 | C_j) \ldots P(A_n | C_j) \) can be obtained from the training data set.

**STEP4:** A sample X of an unknown class, and we calculate the probability \( P(X | C_j) P(C_j) \) that X belongs to each class \( C_j \) respectively, and sample X is assigned to the highest posterior probability class.

Bayesian theorem is described as follows.

\[
P(C_j | A) = \frac{P(A | C_j) P(C_j)}{P(A)}
\]

Where, \( P(C_j) \) is the prior probability, \( P(A | C_j) \) is the joint probability, \( P(C_j | A) \) is the posterior probability.

The joint probability is the probability that the sample A appears when the known class is \( C_j \). Suppose that \( A = a_1, a_2, \ldots, a_m \), and \( a_1, a_2, \ldots, a_m \), is independent each other, then

\[
P(A | C_j) = P(a_1, a_2, \ldots, a_m | C_j) = \prod_{i=1}^{m} P(a_i | C_j)
\]

Given the data sample A, posterior probability is the probability when \( C_j \) is established, and this is very useful in classification and prediction.

2.4 Evaluation Index

We select error rate and ROC(Receiver Operating Characteristic) as evaluation indexes in this paper, which originated the theory of statistical decision-making in the 1950s, and is applied to evaluating radar signal receiving capacity. Error rate is described as follows.

\[
V_e = P_e / N
\]

Where \( V_e \) is error rate, \( P_e \) is the number of error sample in the classification, and \( N \) is the total number of samples.

ROC evaluation indicators are mainly loss rate \( P \), recall rate \( R \) and AUC(Area Under the ROC Curve) value. The ROC evaluation model is described as follows. We can divide the sample into four parts for the relationship between the prediction and the true values in the dichotomous problem.

The prediction value and the actual value are 1 in True Positive(TP), the prediction value and the actual values are 1,0 in False Positive(FP), respectively. The prediction value and the actual values are 1 in True Negative(TN). The prediction value and the actual values are 0,1 in False Negative(FN), respectively. Loss rate \( P \) is the number of the actual correct samples in the judging correct sample, \( P=TP/(TP+FP) \). Recall rate \( R \) is the proportion of True Positive sample in all positive samples, \( R=TP/(TP+FN) \).
Figure 1. ROC curve

The ROC curve can also be used to evaluate the performance of the classifier, and abscissa is the false positive rate, if FP=0, means that all the counterexamples can be judged correctly. For example, all bad reviews can be identified correctly. Therefore, it is suitable that its value is the smaller. Ordinate is the true positive rate, predicted comments are also well comments in all good reviews, and the higher its value is, the better. It is better in the upper left corner. Not only filters out all the negative feedback, but also distinguishes good reviews.

AUC (Area Under Curve) is the area under the ROC curve. ROC curves are generally above the straight line of y = x, so the AUC value ranges from 0.5 to 1. AUC gives the average performance of the classifier, a perfect classifier AUC is 1, and random guess is 0.5, which are described as shown in Figure 1.

3. An Emotional Predictions Analysis of E-commerce

The emotional analysis model is divided into four parts, which are data acquisition and processing, calculation and feature selection, modeling and calculation, prediction model optimization, respectively. The optimal eigenvector is selected by iteration and evaluation, and different feature vector has significant differences in the review data. The emotional analysis model is shown in Figure 2.
3.1. Data Acquisition and Processing
We select the comment data for a product taking Jingdong platform as an example by crawling, and word is cut, filter out some significant noise data, and data preprocessing methods are used by the text repeat, mechanical compression, phrases deleted. Fox example, suitable, suitable, suitable, are compressed a suitable, remove the proper nouns of the commodity, and get valid comment data[10-15]. Then, the comment data are divided into words or word combinations using the word segmentation technique.

3.2. Calculation and Feature Selection
In order to emotional analysis, it is first necessary to digitize the text, and change into a word vector, and combine all the praise with the bad reviews, and calculate TF-IDF value of the combined text, evaluate the importance of each word in the overall product reviews, the evaluation feature is selected based on the TF-IDF value. For example, the "the phone and very", "very and suitable "," suitable and !", three collocations as a classification feature.

3.3. Vector Transformation
We convert a document into a vector that can be used for mathematical operations, and this is the vector conversion. We set up a training set which contain U documents of comments information, and convert them into corresponding vectors with V dimensions, and constitute a matrix of U * V, that is, constitute a Vector Space Model(VSM), and all operations are based on this.

Use a function $\Gamma(x)$ to represent the value of feature vector, which is described as follows.

$$\Gamma(x) = \begin{cases} 1, & \text{if Object appears} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Fox example, feature vector $V=[\text{clothes, very good, beautiful, fashion, good style}]$, if a comments data $V1=[\text{clothes, buy, beautiful, general ,uncomfortable, next time, wish}]$, According to vector transformation, clothes and beautiful appear in the comments data $V1$, their value of feature vector are 1, the rest doesn’t appear, their value of feature vector are 0, and transformed vector $V=[1,0,1,0,0]$. Also need to categorize each comment data and label it after transforming vector, and add a dimension, we also use a function $f(x)$,which is described as follows.

$$f(x) = \begin{cases} 1, & \text{if Comment data is praise} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

If comment data $V1$ is bad review, and transformed vector $V=[1,0,1,0,0]$. 

3.4. Bayesian Classification and Feature Selection
We predict the emotional trend of user using Bayesian Probability Prediction Model after getting feature vector, and calculate the prediction accuracy and ROC evaluation values. Different prediction results are compared and analyzed, and the number of feature vectors with the smallest prediction error is chosen as the optimal feature vector model.

3.5. Model Evaluation
We get the optimal model parameters using different model evaluation indicators after getting the number of different feature. The model is used to predict the random sampling comment data, and solve the evaluation index problem of corresponding ROC curve, verify the validity of the model.

4. Experiment and Analysis
Each figure should have a brief caption describing it and, if necessary, a key to interpret the various lines and symbols on the figure.
4.1. Data Acquisition and Processing
We use crawler technology to download iphone7 comments from Jingdong, total 42615. Remove some unrelated noise and repeat comments, and save the excel file, utilize mentioned technology to transform VSM. There are 2200 bad reviews, 11000 neutral reviews and 30600 praise reviews after cleaning the data. Because of uncertainty and other noisy data, the remaining valid comment data for the praise of 25,000 and the bad reviews of 1900, and mark the classification label. The comment data is shown in Table 1.

Table 1. Comment data style sheet on mobile phone

| No. | comment content                                                                 | comment classification |
|-----|---------------------------------------------------------------------------------|------------------------|
| 1   | the perfect machine, come to Jingdong, and save a lot of money, happy, very happy, good, good and very good, must be said three times, the appearance of flawless, good call quality | Positive               |
| 2   | no-good, bad                                                                    | Negative               |
| ... | ……                                                                               | ……                    |
| 42615 | bad reviews, furious, quality problems, Mobile phone cannot be opened           | Negative               |

Authors should try to try to make economical use of the space on the page; for example:
- avoid excessively large white space borders around your graphics;
- try to design illustrations that make good use of the available space—avoid unnecessarily large amounts of white space within the graphic;

4.2. TF-IDF Calculating and Feature Selection
Bad reviews and praise are combined together, and add up to 26900, utilize word segmentation technology to transform, and get feature vector \( W = (w_1, w_2, w_3, ..., w_n) \), \( w_i \) is the ith word, calculate TF-IDF values of the comments vocabulary by TF-IDF, get \( R = (r_1, r_2, r_3, ..., r_m) \), where, \( r_i \) is the TF-IDF value corresponding to word \( w_i \). There are some repetitive words after cutting words, and only has 5200 different words after calculating the TF-IDF value. There are some repetitive words after cutting words, and only has 5200 different words after calculating the TF-IDF value. Select separately \( N = \{100, 300, 500, 1000, 1500, 2000, 3000, 4000, 5000\} \), which is regarded as the number of features. And sort after calculating the TF-IDF value, and select the TF-IDF value of the top \( N \) as the feature vector, and predict using Bayesian classifier.

We transform vector for each comment using the selected feature vector. Finally, each comment is converted to \( V = [1, 0, 0, ..., 0, 1] \) by formula(3), and add classification label by formula(4), get \( V = [1, 0, 0, ..., 0, 11] \), and it means to praise.

Through the features selection and the vector transformation, we get the vector matrix with 25000 reviews and 1900 negative reviews. Since the number of good reviews is more than 80% of the sample space, we randomly select 1000 negative reviews and 1000 praises as model training samples.

4.3. Optimal Feature Selection
The following table shows the number of features from 100 to 5000, calculate error and AUC values using the Bayesian classifier in emotional prediction.
Experimental results are illustrated in Figure 3. At the mark of the triangle, when the number of features is 4000, the error value (difference value) is minimal, and the auc value is the largest, and its value is 0.82, the error value is 0.10. So we can choose the number of features for 4000, and it is looked as the best parameters in this experiment.

When constructing the forecasting model, the higher the number of features is, the higher the accuracy is, but when the number of features rose to 4000, the prediction accuracy is reduced. With the increase of features, the division of vocabulary decreases when we calculate TF-IDF. Negative vocabulary may be used to look as positive. So it is important that we select the number of features for the prediction accuracy of the model reasonably.

4.4 Experiment Analysis
The selected parameter is optimal parameters of the Bayesian classification model based on the previous analysis. All comments are classified and predicted, and draw the roc curve. As shown in Figure 4.

The solid line in the figure represents the roc curve, and Y-axis is a true positive, X axis is false positive, when $X \in [0.3, 0.5]$, the solid line is far from the dotted line, which indicates that the model can be well predicted and classified.
From Figure 5, we find that the negative recognition rate of the model is high. This also shows that the sensitivity of the positive of vocabulary is higher than the negative vocabulary.

The following Table 2 shows the comparison of the final forecast and actual values of the 26800 comments in this experiment, where, the number of positive is 25,000, and the number of positive that are predicted correct is 21935, and the number of positive that are predicted incorrect is 21935, the number of negative is 25,000, and the number of negative that are predicted correct is 1745, and the number of negative that are predicted incorrect is 155.

**Table 2.** Result of emotional prediction model with optimal number of attributes

| Actual value | Correct prediction | Wrong judgment | Accuracy |
|--------------|--------------------|----------------|----------|
| Positive     | 25000              | 21935          | 3065     | 87.8%    |
| Negative     | 1900               | 1745           | 155      | 91.9%    |

For all comments, the accuracy rate for forecasting favorable comment in praise is 87.7%, and the accuracy rate for forecasting bad review in bad reviews is 91.6%, the average accuracy rate is about 90%, and it shows that the model has a high accuracy.
5. Conclusions
In big data environment, the massive user's comment information is very important data sources for the user's demand, which is beneficial to the precision marketing and product improvement for the e-commerce company. Traditional data mining methods are more limited when the information is extracted. How do the researchers extract high-quality information from complex, massive, and heterogeneous comments for consumers and decision makers, it has become a common concern of all parties. Based on the general emotional analysis, this paper uses TF-IDF and Bayesian probability model, construct the multi-dimensional emotion analysis and forecasting model, improve the accuracy of feature selection, and the method is also suitable for public opinion analysis. But there are still shortcomings, for example, dictionary is not perfect, the integrity of the comments sentence structure is not considered, the efficiency of word frequency is not high.

6. Acknowledgments
The authors would like to thank for financial support by youth fund project of the humanities and social sciences of Education Ministry (No.15YJC870004), science and technology innovation team of XiangNan University (Recognition and analysis based on big data public opinion), Social Sciences fund project of Hunan Province (No.13YBA302) and education Science in Hunan province in 12th Five-Year planning project (XJK014CGD081, XJK011BXJ004).

7. References
[1] ZHOU Li-zhu, HE Yu-kai, WANG Jian-yong. Survey on research of sentiment analysis, Computer Applications, 2008, 28(11): 2725-2728.
[2] Xiaodong Chen. Research on sentiment dictionary based emotional tendency analysis of Chinese Microblog, Hua zhong University of Science & Technology, 2012.
[3] Jiang Yahua. The extraction of mobile product comments based on the hownet. Harbin Institute of Technology, 2011.
[4] Ma Xiaoling, Jin Biyi, Fan Binshi. An Analysis of Chinese Text Emotional Tendency[J]. Information data work, 2013, (01): 52-56.
[5] PANG Hai-jie. Text sentiment analysis-oriented commodity review detection, Journal of Computer Applications. 2012, 32(7): 2038-2040.
[6] ZHOU Min, LI Rui. Method of review Information Mining Based on Characteristics of Commodity Goods, JISUANJI YU XIANDAIHUA, 2014(6): 98-101.
[7] ZENG Jia-ni, LIU Gong-shen, SU Bo. Research on Sentiment Classification of Chinese Microblog, Information Security and Communication Confidentiality, 2013 (3): 56-58.
[8] Xie Li Xing, ZHOU Ming, Sun maosong. Hierarchical structure based hybrid approach to sentiment analysis of Chinese MiroBlog and its feature extuaction, Journal of Chinese information processing, 2012, 26(1): 73-83.
[9] WEI Wei1, XIA Yang, CHEN Qian, Survey on Chinese text sentiment analysis. Journal of Computer Applications. 2011, 31(12): 3321–3323.
[10] Zhang Xue Wen. RESEARCH ON COLLABORATIVE FILTERING ALGORITHM IN ARTIFICIAL RECOMMENDATION SYSTEM. Shanghai Jiao Tong University, 2008.
[11] WANG Hong-wei, GUO Kai-qiang, DU Zhan-q1, Intention of Using Review Website to Obtain Information: An Empirical Study, Information Science, 2015, 33(12): 27-33.
[12] LIU Li, HE Zhong-shi. Term selection and weighting approach based on key words in text categorization. Computer Engineering and Design, 2006, 27(6): 934-936.
[13] FANG Ran, MIAO Duoqian, ZHANG Zhefei. An emotion-based method of topic detection from Chinese microblogs, CAAI Transactions on Intelligent Systems, 2013(3): 208-213.
[14] Yang liang, Lin yuan, Lin hongfei. Micro-Blog hot events detection based on emotion distribution, Journal of Chinese information processing, 2012(1): 84-90.
[15] LIU Wei-ping, ZHU Yan-hui, LI Chun-liang. Research on building Chinese basic semanticlexic on Journal of Computer Applications, 2009(10): 2875-2877.