Asia-pacific Journal of Convergent Research Interchange
Vol.6, No.6, June 30 (2020), pp.25-33
http://dx.doi.org/10.21742/apjcri.2020.06.03

Customer Sentiment Analysis: Take Restaurant Online Reviews as an Example

Xu Gang[1)]

Abstract

The customer's sentiment analysis can fully grasp the customer's consumption trends and the popularity of the product. This article takes the restaurant's online reviews as an example and uses machine learning algorithms to analyze the customer's sentiment tendency. The research describes the overall process of sentiment analysis, and discusses the implementation methods of corpus acquisition, feature extraction, feature selection, expanded sentiment dictionary construction, and determination of class labels. The training method of machine learning sentiment analysis model is analyzed. After the experimental analysis of the obtained corpus, the experimental results are given through the chart. The C4.5, Bagging, Multinomial naïve bayes algorithm has achieved good results.

Keywords: Schematic Diagram, Product Reviews, Machine Learning, Evaluation

1. Introduction

With the development of the Internet, more and more users are accustomed to online reviews of products. It is not only an effective way for users to provide feedback after using products, but also has irreplaceable references for merchants and browsers. However, at present, the vast majority of comments have not been effectively used. In the vast number of comments, people cannot intuitively draw conclusions, so how to effectively excavate and process these comments is a very meaningful thing. At the same time, because Internet reviews are very different from traditional texts, sentiment analysis for online reviews will be more difficult than traditional texts. Research on sentiment analysis has significant academic value and intuitive broad commercial value. In the computer field of various countries in the world, sentiment analysis has become a hot direction of scientific research[1][2].

Sentiment analysis fully reflects the interdisciplinary task. Researchers in this field have introduced more and more research methods in other professional fields into the field of text sentiment analysis, innovating many excellent methods and promoting the development of the

Received(December 31, 2019), Review Result(1st: February 03, 2020, 2nd: March 26, 2020), Accepted(May 27, 2020)

1) (Professor) Chengdu University of Technology, China
email: xugang@cdut.edu.cn

ISSN: 2508-9080 APJCRI

field, and the application of Vector Space Model (VSM) in text processing, makes it possible for many algorithms in other fields to be applied in the field of sentiment analysis. This makes various machine learning algorithms can be applied to text sentiment analysis easily. Among the commonly used analysis algorithms are Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (support vector machine, SVM), K-nearest neighbor algorithm, LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis, etc[3-6].

This paper deeply discusses the role of sentiment dictionaries in sentiment analysis of online product reviews, compares the effects of typical machine learning methods on sentiment analysis of restaurant reviews. An effective machine learning sentiment analysis model based on the extended sentiment dictionary method is proposed focus on the hidden attributes discovery, product inter-relationship discovery and customer sentiment tendency judgment of online product review sentiment analysis.

2. Product Review Sentiment Analysis System

In order to get a comprehensive sentiment analysis of restaurant reviews based on website reviews, this article uses natural language processing methods, combined with machine learning methods to mine online product reviews for preprocessing and sentiment analysis.

First, select the restaurant on the public review to obtain the corpus, and then preprocess the corpus, use the improved sentiment dictionary construction method to complete the sentiment dictionary required for restaurant sentiment analysis, and use machine learning algorithms to realize sentiment analysis of restaurant merchant reviews.

The core of study is to comprehensively mine product reviews and obtain an overall description of restaurant review sentiment analysis based on the analysis and processing of user reviews. The overall process of product review sentiment analysis is shown in [Fig. 1].

Step 1. Obtain corpus, select words and phrases through preprocessing and extract features;

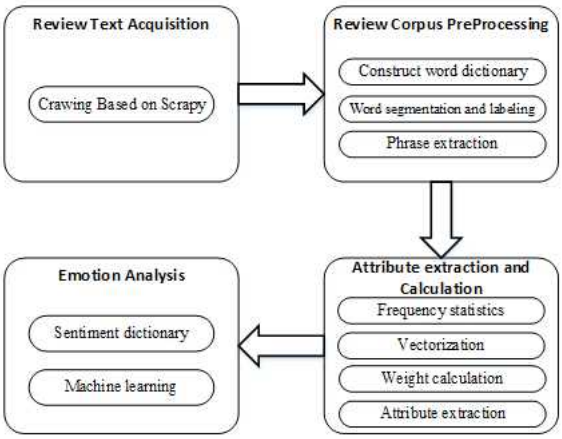
Step 2. Filter features, filter related words and discover hidden attributes, construct an extended sentiment dictionary;

Step 3. Text characterization, determine the characterized text class labels;

Step 4. Use machine learning algorithms for training to build a sentiment analysis model;

Step 5. Obtain the machine learning method accuracy rate, error rate, accuracy rate, recall rate, and ROC curve (receiver operating characteristic curve) and other judgment indicators through testing. Analyze and determine a machine learning method that meets sentiment

analysis of restaurant reviews.



[Fig. 1] Schematic Diagram of Sentiment Analysis Process of Product Reviews

3. Feature Processing Technology

To build the machine learning model, the first step is to calculate weights, extract attribute words, and filter features. This study uses TF-IDF (termfrequency-inverse document frequency) for calculating weights. TF-IDF is a text word weighting method based on statistical methods, used to evaluate the importance of text vectors in the entire document set.

The core idea of TF-IDF lies in two factors that affect the weight, namely the word frequency TF and the inverse document frequency IDF:

$$W(t,d) = tf_{ij} \times idf_i = tf_{ij} \times lb\left(\frac{N}{idf_i}\right) \tag{1}$$

Where is the weight of feature t in text d. N is the total number of texts. Tfis the number of times feature t appears in text d. is quotient of the total number of files divided by the number of files containing the word. Then take the logarithm of the quotient. To avoid some words not appearing in the text, the following formula is used to calculate the weight:

$$W(t,d) = \frac{(1+lb(tf_{ij})) \times lb\left(\frac{N}{n_t}\right)}{\sqrt{\sum_{t \in d}\left[(1+lb(tf_{ij})) \times lb\left(\frac{N}{n_t}\right)\right]^2}} \tag{2}$$

Where is the number of text with t features in N.

Compared with latent Dirichlet distribution and latent semantic analysis, the semantics it carries become richer because Word2vec considers the context. Therefore, this study selects Word2vec to filter related words, and adjusts Word2Vec to obtain an accurate and effective word vector model. Hidden attributes can be discovered through the weights of phrase combinations, and filter out indirect dishes.

The establishment and expansion of sentiment dictionary is an important process of sentiment. In terms of the construction of the polar dictionary, all the words are used in the obtained corpus as a dictionary to filter the words in the ordinary polar dictionary and filter the words that do not appear in the review document, thereby improving the efficiency of dictionary use. Word2vec is used to expand similar words in the existing degree dictionary to achieve the purpose of expanding the degree dictionary and improve the utilization efficiency of the degree dictionary. According to TF-IDF, words with emotional tendencies are marked, and Word2vec is also used to filter related synonyms and expand the domain dictionary.

4. Machine Learning Methods

For the problem of sentiment analysis of product reviews, this article builds and tests the following typical machine learning algorithms, including Ada Boosting, Bagging, Bayes Network, Decision Tree, C4.5 classification tree, Naive Bayes classifier, Multinomial Naive Bayes and Ripper and other algorithms. The typical Naive Bayes classification algorithm and C4.5 classification tree that have achieved good results in the test.

As a typical probability model algorithm, Naive Bayesian classification algorithm is used to obtain the probability value of the text to be classified according to the Bayesian formula. Maximum of these probabilities is used to complete the classification. Assume that each text sample in the text set can be represented by an n-dimensional feature vector.

$$(c|d)= \frac{p(c)p(d|c)}{p(d)} \tag{3}$$

Among them, c represents the category vector. When using the Naive Bayes method, each feature is regarded as completely independent of each other, so:

$$p(d|c)= \frac{p(d)}{p(c)} \prod p(t_k|c) \tag{4}$$

In the polynomial model, let a text, is a word that has appeared in the text, and it can appear repeatedly, then the prior probability is the total number of words under class c/the total number of words in entire training text.

Class conditional probability = (the sum of the number of times the word of class appears in each text +1)/(class c; the total number of words in class +), v is the word list vector of training text.

The C4.5 algorithm is a typical decision tree, and the CA.5 algorithm also uses a top-down approach when constructing the decision tree, choosing the best feature to split at each step.

The best feature here is usually to make the training set in the child nodes as pure as possible. The most commonly used index for measuring the purity of the training set in the C4.5 algorithm is the information gain rate.

The information gain rate is calculated as follows:

$$GainRatio(S,A)= \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (5)$$

Where represents the information gain, and the division information represent the breadth and balance of the sample set S divided by the feature A.

The larger the information gain rate, the less impure the training set, so in the process of building the C4.5 classification tree, the information gain rate of each feature division needs to be calculated.

C4.5 The classification tree calculates the information gain rate of each segmentation and combination of each variable, finds the optimal value combination and segmentation point of the variable, and then compares the optimal value combination and segmentation point of each variable. Finally, the best variable and the best value combination and cut point of the variable are obtained.

C4.5 the steps to build a tree are as follows:

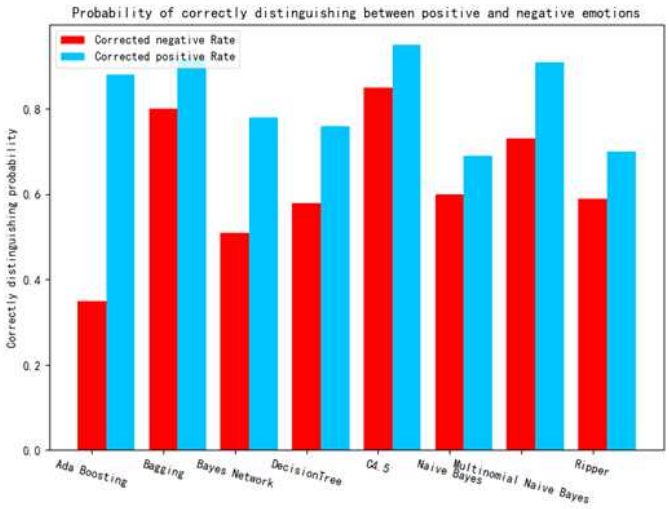
1) Determine whether the current sample set meets the termination condition, if not, calculate the information gain rate of each feature of the current sample set.

2) Select the feature with the largest information gain rate as the divided feature.

3) The two parts of the sample set divided by the features are used as new sample sets, and the steps are repeated from Step 1 to 3.

5. Experiment Result

Aiming at the problem of restaurant review sentiment analysis, using the 1900 features screened, a variety of machine learning algorithms are used to build a classifier, and the algorithm performance indicators are verified by the results of the 10-fold cross-validation to determine the effective machine learning algorithms. Firstly, the algorithm correctly distinguishes the probability of positive and negative emotions, as shown in [Fig. 2].



[Fig. 2] The Probability of Correctly Distinguishing between Positive and Negative Emotions

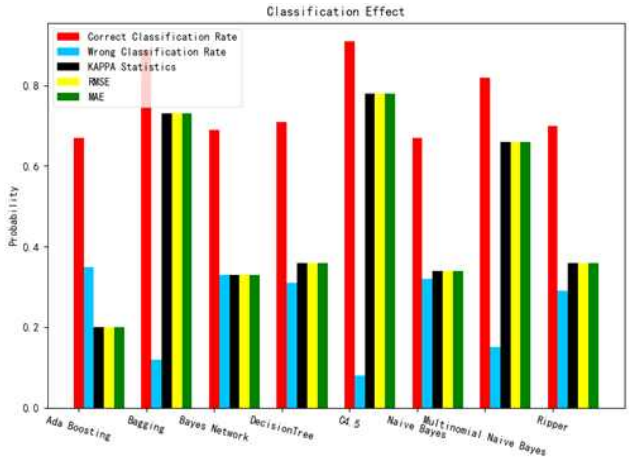
[Fig. 2] shows the probability of correctly distinguishing between positive and negative emotions. It can be seen from the test results that the C4.5 algorithm performs well in two aspects. Ada boosting is excellent in distinguishing positive emotions, but the results are very unsatisfactory in distinguishing negative emotions. In general, analysis of negativity in the field of distinguishing emotions is more difficult than analysis of positivity. It can be seen that the better algorithm is C4.5, Bagging, Multinomial Naïve Bayes.

In the classification effect diagram of [Fig. 3], the absolute error and the root mean square error are both the predicted and the actual difference, so the smaller the better, the same, the lower the error rate, the better.

KAPPA statistics is the relative situation of classifier statistics and real classification. Its range is between [-1, 1]. When the value is close to 1, it indicates that it is more similar to the real classification. When the value is close to 0, it indicates that it is more similar to the random distribution. The closer to -1, the more opposite to the true classification.

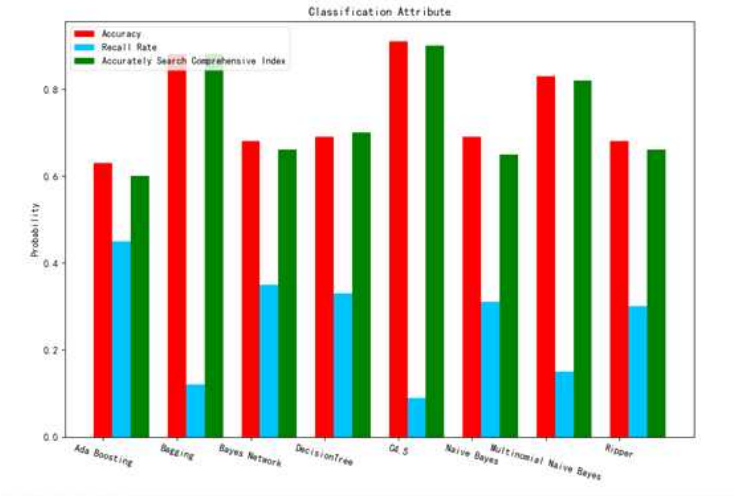
From the performance index results in [Fig. 3], it can be seen that C4.5 Bagging, Multinomial

Naïve Bayes are better algorithms.

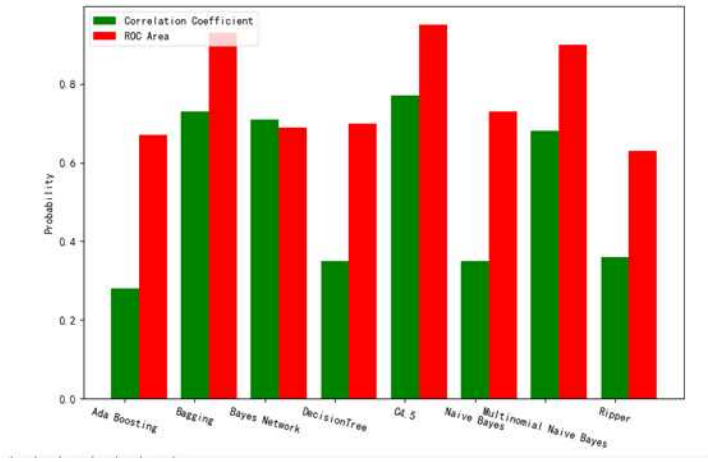


[Fig. 3] Classification Effect

In the experiment results of the machine learning performance indicators shown in [Fig. 4] and [Fig. 5], the accuracy rate indicates how many correctly identified reviews account for the total reviews, and the recall rate indicates how much correctly identified reviews account for all reviews that should be correctly identified. The accurate comprehensive index refers to an index that combines the recall rate and the accuracy rate. The correlation coefficient takes the value between [-1, 1]. 1 represents exactly the same as the actual prediction, 0 represents the same effect as the random prediction, and -1 represents the prediction result is completely opposite to the actual result. Compared with KAPPA, the correlation coefficient index is more suitable for corpora with an unbalanced amount of class label text. The area range of the ROC curve is between [0, 1] and equal to 1 indicates the best effect.



[Fig. 4] Classification Attribute



[Fig. 5] Correlation Coefficient and ROC Area

It can be drawn from the experiment results that the effect of C4.5, Bagging, Mulinomial Naïve Bayes algorithm is better.

6. Conclusion

This study introduces natural language processing and online product reviews, and details the common methods of comment mining and sentiment analysis. In response to customers' online reviews of restaurant products, based on the description of the overall process of sentiment analysis, this article further discusses the design and implementation methods of corpus acquisition, feature extraction, feature screening, expanded sentiment dictionary construction, and class label determination, and analyzes the core The machine learning sentiment analysis model training algorithm implements an online product review sentiment analysis model that is suitable for developing party implied attributes, showing the correlation between products, and judging customers' emotional tendencies. Since there are few related domain dictionaries in the catering field at present, it is very necessary for the expansion of sentiment dictionaries. We use new methods to expand the sentiment dictionaries in the catering domain, and give full play to the effectiveness of the sentiment dictionaries. After testing the obtained corpus, the test results are presented in the form of charts, and on the basis of full analysis and evaluation of machine learning performance indicators, a better machine learning algorithm is obtained.

References

[1] Wang C. H., Han D., Sentiment Analysis of Micro-blog Integrated on Explicit Semantic Analysis Method, Wireless Personal Communications, **(2018)**, Vol.102, No.2, pp.1095-1105.

[2] Mauro Dragoni, Soujanya Poria, Erik Cambria, OntoSenticNet: A Commonsense Ontology for Sentiment Analysis, IEEE Intelligent Systems, **(2018)**, Vol.33, No.3, pp.77-85.

[3] Zhang S., Wei Z., Wang Y., Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary, Future Generation Computer Systems, **(2018)**, Vol.81, April, pp.395-403, DOI: https://doi.org/10.1016/j.future.2017.09.048

[4] Hassan A., Mahmood A., Convolutional Recurrent Deep Learning Model for Sentence Classification, IEEE Access, **(2018)**, Vol.6, pp.13949-13957.

[5] Maria Jimenez-Zafra S., Teresa Martin-Valdivia M., Dolores Molina-Gonzalez M., Alfonso Ureña-López L, How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain, Artificial Intelligence In Medicine, **(2019)**, Vol.93, pp.50-57, DOI: 10.1016/j.artmed.2018.03.007

[6] Kuebler S., Liu C., Sayyed Z. A., To use or not to use: Feature selection for sentiment analysis of highly imbalanced data, Natural Language Engineering, **(2018)**, Vol.24, No.1, pp.3-37.