Classification of Customer Reviews on E-commerce Platforms Based on NaiveBayesian Algorithm and Support Vector Machine

Wenyuan Zhao1, *
1School of Information Science and Engineering, Southeast University, Nanjing, 211189, China
*Corresponding author’s e-mail: 213170530@seu.edu.cn

Abstract. Customer reviews on e-commerce platforms contain valuable information, while sifting through them manually tends to dismay people because of the huge amount of data. Ideally, the identification classifier would analyze the emotional disposition of product reviews (positive, negative) and aggregate opinions about each of them. Previous literature has demonstrated extensive research conclusions on opinion extraction and semantic classification of product reviews on websites. However, analysis on the universality of machine learning algorithms in identifying the emotional tendency of different Chinese e-commerce reviews is not yet studied. This study uses a method, which is based on general machine learning algorithms, to classify feedbacks. Our classifier extracts Chinese word segmentation and text frequency for feature extraction and scoring, and implements the classification with methods of Naive Bayesian and Support Vector Machines. Experimental results on the Alibaba product review sentiment datasets show that our model based on two machine learning algorithms, though results in different performances, can provide suggestions on the selection of the identification classifier using a trade-off strategy and helps obtain fast and accurate classification on reviews of different categories.

1. Introduction
Nowadays, e-commerce has become a major shopping platform for customers. The Amazon in the US and the Alibaba in China provide numerous services for people around the world. E-commerce brings much convenience to customers due to its huge number of products, quick payment and user-friendly negotiation systems. During shopping process, former customers’ reviews on different categories of goods are decisive for current customers to buy the product or not. These comments and feedbacks are also essential for online stores to acknowledge the advantages and disadvantages of their products so as to provide better services. However, the amount of reviews on online products is too huge to make analysis manually. Therefore, it is necessary to provide an intelligent way to classify the new reviews and make a prediction on whether the product is worth buying.

Online reviews belong to text information so that the classification of Chinese e-commerce reviews is based on the technology of text classification. Text classification technology is an effective way to aggregate valuable information in massive text data [1]. With the continuous development of natural language processing in Artificial Intelligence (AI), text classification has entered the stage of machine learning. Thorsten Joachims proposed that Support Vector Machine can be used to classify documents in text classification [2]. Also, Yiming Yang proposed a method that can apply Decision Trees to text classification. And Yujia Wu proposed a framework based on words in pairs neural networks (WPNN)
for text classification [3].

In the application of the classifier based on machine learning algorithms, there have been a large number of methods and contents on semantic classification of product reviews on English webs. For example, Kushal Dave has identified unique properties of product reviews on websites that are in English language and developed a method that can automatically distinguish between positive and negative reviews [4]. Also, Feng Xu presented a continuous naive Bayes learning framework adopted on Amazon product and movie review sentiment data sets to deal with continuously updated reviews [5]. Sajjad Haider successfully identifies the emotional tendencies of Twitter users by analyzing the impact of adverb forms in users’ reviews [6]. Since Chinese e-commerce has grown rapidly and the characteristics of Chinese differ from those of English, it is necessary to study analytical methods of Chinese product reviews separately. Liming Rong proposed that text mining can be used to evaluate the value of products by processing product reviews and made an experiment on water heats in Tmall [7].

However, most research still focuses on the classification of English web reviews. The related works on Chinese text are still scarce (even though the limited studies only focus on a particular category). In order to explore the universality of machine learning algorithms in the commodity application of Chinese e-commerce platforms, we study the feasibility and efficiency of classifying Chinese e-commerce reviews based on Naive Bayes (NBC) and Support Vector Machine (SVM).

In this paper, we implement the identification classifier based on Naive Bayes and Support Vector Machine and obtain a fast and accurate classification on different Chinese e-commerce reviews by adjusting the corresponding parameters. By comparing the performance of two classifiers, we also give an appropriate trade-off on the selection of identification classifiers when making an analysis on Chinese reviews and point out the challenges to be resolved.

2. Method

The purpose of this paper is to propose a model that can identify the e-commerce reviews, which need an identification classifier on Chinese text data. Since all the reviews contain deleting punctuation, special symbols and meaningless vocabulary (stopwords), it is necessary to pre-process raw review data before training and classifying. The process of classifying e-commerce reviews consists of text pre-processing and text training and classifying.

2.1 Text Pre-processing

2.1.1 Word Segmentation.

Before processing Chinese text data, it is necessary to separate the text into several word segments. The precision of word segmentation determines the efficiency of Chinese text feature selection and training, since the meaning of a Chinese paragraph is determined by a phrase rather than a single word, which is different from English that every word has its own meaning. Therefore, it is important to choose a fast and precise word segmentation algorithm. There have been several common tools of Chinese word segmentation, such as Poading Analyzer, IK Analyzer and Jieba [8].

2.1.2 Text Feature Selection.

There are some meaningless vocabulary and symbols (punctuation, stop-words) that need to be eliminated for that they do no contribution to the classification. After deleting the stop-words and obtaining the word segments, features could be extracted from the text, which is represented by more representative and distinctive words, which can reduce the number of Text Feature Spaces and improve the speed and accuracy of text classification.

In addition, the feature weights should be calculated to indicate the importance of features in the document after the selection of text features. At present, the widely used weight calculation method is Text Frequency and Inverse Document Frequency (TF-IDF). The tools used in this paper are provided by Scikit-learn Library.
2.2 Text Training and Classifier

In the current study, Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) are used as the classifier of Chinese review texts. NBC is used to calculate the probability of which category a short text belongs to, while SVM is used to divide the positive and negative of the text meaning.

2.2.1 Naive Bayes.

NBC is used to describe the probabilistic relationship between text and the category to which the text belongs in text classifications [9]. \( d = \{ w_1, w_2, \cdots, w_i \} \) denotes a sample of the document set to be sorted, where \( w_i \) is the \( i_{th} \) feature of \( d \). According to Bayes Equation,

\[
p(c_n | d) = \frac{p(d | c_n) p(c_n)}{p(d)},
\]

where \( c_n \in C, C = \{ c_1, c_2, \cdots, c_j \} \) denotes the set of categories and \( p(d) \) is a constant. Then we can calculate \( p(c_1 | d), p(c_2 | d), \cdots, p(c_j | d) \). In order to reduce the computational overhead, we assume that features are conditionally independent of each other. Therefore, we have

\[
p(d | c_n) = \prod_{k=1}^{i} p(w_k | c_n).
\]

With the maximum value of \( p(c_m | d) \), \( c_m \) is the category the document \( d \) belongs to.

When constructing the Bayesian classifier, the category of text is judged by calculating the conditional probability of feature items in different categories. The efficiency of the classifier presented in this paper is higher when the contribution of the feature weights to the classification is not that different from each other.

2.2.2 Support Vector Machine

![Figure 1](image)

Figure 1. A simple illustration of SVM where hollow and solid circles denote samples of two different categories, the solid line denotes the hyperplane, and the dotted line denotes the margin of each class.

Support vector machines (Vapnik, 1982) have solid theoretical foundations and have been proved empirical successful. They have been applied to tasks such as handwritten digit recognition, object recognition, and text classification [10]. We shall consider SVM in the binary classification setting as is illustrated in Figure 1. The algorithm produces a hyperplane that separates the training data with the maximal margin. All vectors lying on one side of the hyperplane are labeled as \(-1\), and all vectors lying on the other side are labeled as \(1\). The equation of the hyperplane is given by two parameters, a real-
valued vector \( \mathbf{w} \) of the same dimension as the input feature vector \( \mathbf{x} \), and a real number \( b \):
\[
\mathbf{w} \mathbf{x} - b = 0, \tag{3}
\]
where \( \mathbf{w} \mathbf{x} \) equals \( w^{(1)} x^{(1)} + w^{(2)} x^{(2)} + \cdots + w^{(D)} x^{(D)} \), and \( D \) is the number of dimensions of the feature vector \( \mathbf{x} \). The predicted label of some input feature vector \( \mathbf{x} \) is given by:
\[
y = \text{sign} (\mathbf{w} \mathbf{x} - b), \tag{4}
\]
where the sign is a mathematical operator that takes any value as input and returns +1 if the input is a positive number or -1 if the input is a negative number. Once the optimal parameters \((\mathbf{w}^* \text{ and } b^*)\) have been determined, the category of \( \mathbf{x} \) can be identified by
\[
f(\mathbf{x}) = \text{sign} (\mathbf{w}^* \mathbf{x} - b^*) \tag{5}
\]

3. Experiment
We first extract users' reviews from Alibaba, the biggest e-commerce platform in China, for training and testing, including ten categories: “book”, “iPad”, “iPhone”, “fruit”, “shampoo”, “water heater”, “milk”, “cloth”, “computer”, “hotel”. The total amount of data is 62774. Each review has a label, where “-1” denotes negative reviews and “1” denotes positive reviews. The size of reviews in each category is presented in Table 1. The experiment is performed on Python 3.7 which is installed in Windows 10 operating system.

| Category         | Total number | Positive number | Negative number |
|------------------|--------------|-----------------|-----------------|
| “book”           | 3851         | 2100            | 1751            |
| “iPad”           | 10000        | 5000            | 5000            |
| “iPhone”         | 2323         | 1165            | 1158            |
| “fruit”          | 10000        | 5000            | 5000            |
| “shampoo”        | 10000        | 5000            | 5000            |
| “water heater”   | 575          | 475             | 100             |
| “milk”           | 2033         | 992             | 1041            |
| “cloth”          | 10000        | 5000            | 5000            |
| “computer”       | 3992         | 1996            | 1996            |
| “hotel”          | 10000        | 5000            | 5000            |

First of all, we test the accuracy of classification with different sizes of datasets, and we compare the classification accuracy between NBC and SVM. In addition, the accuracy of classification in a multi-category model is difficult to be illustrated because the training data is unbalanced (there are many data in some categories, and few data in others). We calculate the Precision, Recall [6], and F1-score [11] as the evaluation metrics of our experiment.

- **Precision**: Precision is the fraction of relevant reviews (true-positive) against the total reviews (true-positive + false-positive) that are classified. Precision informs us of the performance on correct results. What we want to know is whether the received results are correct rather than the number of correct results we received.
- **Recall**: The recall is the fraction of relevant reviews (true-positive) against the total amount of reviews that are predicted as positive. We hope that all the reviews are classified correctly so that users could make important decisions based on the classifier.
- **F1-score**: F1-score is the mean of Precision and Recall.

On the other hand, the segmentation ratio of the training data set and test data set also has a certain influence on the performance of classification. We calculate the average accuracy under different proportions of the test set by changing the parameter \( test \_size \). The result and the discussion of the experiment are presented in Section 4.
4. Results and Discussion
The accuracy of the on-line review classification based on SVM and NBC is presented respectively in Table 2 and Table 3. The mean recall rate of the SVM is 86.95%, the accuracy is 86.7%, and the mean F1-score is 86.69%. The mean recall rate of NBC is 79.90%, the accuracy is 77.4%, and the mean F1-score is 78.67%. In addition, we compare the performance of SVM and NBC in Table 4. It can be seen that the training speed of NBC is up to 37914.4 reviews per second, while the speed of SVM is 2166.9 reviews per second.

| Category       | Precision | Recall | F1-score | Support |
|----------------|-----------|--------|----------|---------|
| “book”         | 95%       | 91%    | 93%      | 1271    |
| “iPad”         | 74%       | 79%    | 77%      | 3300    |
| “iPhone”       | 92%       | 78%    | 85%      | 767     |
| “fruit”        | 90%       | 84%    | 87%      | 3300    |
| “shampoo”      | 78%       | 83%    | 81%      | 3300    |
| “water heater” | 87%       | 53%    | 66%      | 190     |
| “milk”         | 100%      | 99%    | 99%      | 671     |
| “cloth”        | 85%       | 88%    | 86%      | 3300    |
| “computer”     | 94%       | 86%    | 90%      | 1317    |
| “hotel”        | 98%       | 98%    | 98%      | 3300    |
| Mean           | 86.95%    | 86.43% | 86.69%   | 20716   |

*The precision denotes the proportion of positive samples that are classified as positive; The recall denotes the proportion of the total positive cases that are predicted to be positive; The support denotes the size of samples.

| Category       | Precision | Recall | F1-score | Support |
|----------------|-----------|--------|----------|---------|
| “book”         | 100%      | 56%    | 71%      | 1271    |
| “iPad”         | 58%       | 79%    | 67%      | 3300    |
| “iPhone”       | 100%      | 13%    | 23%      | 767     |
| “fruit”        | 83%       | 87%    | 85%      | 3300    |
| “shampoo”      | 74%       | 84%    | 79%      | 3300    |
| “water heater” | 0%        | 0%     | 0%       | 190     |
| “milk”         | 100%      | 35%    | 52%      | 671     |
| “cloth”        | 82%       | 88%    | 85%      | 3300    |
| “computer”     | 99%       | 44%    | 61%      | 1317    |
| “hotel”        | 83%       | 99%    | 90%      | 3300    |
| Mean           | 79.90%    | 77.44% | 78.67%   | 20716   |

*The precision denotes the proportion of positive samples that are classified as positive; The recall denotes the proportion of the total positive cases that are predicted to be positive; The support denotes the size of samples.

| Alg. | Training set | Test set | Time  | Recall | Accuracy | Speed  |
|------|--------------|----------|-------|--------|----------|--------|
| NBC  | 42058        | 20716    | 0.546s| 77.44% | 77.4%    | 37941.4 doc/s |
| SVM  | 42058        | 20716    | 9.56s | 86.43% | 86.7%    | 2166.9 doc/s  |

Based on the comparison results of all experiments, SVM displays more advantages in recall rate and accuracy, but NBC has obvious advantages in classification speed. One possible reason for the slow speed of SVM is that SVM classification method can only classify positive and negative classes at one time, so it needs to classify the 10 categories in the experiment requirements one by one. Therefore, if
the number of reviews is massive and the accuracy requirement of classification is not strict, NBC is more applicable. Whereas if the number of reviews is small and the precision requirement of classification is strict, SVM is more applicable. In addition, if the classification task is to separate into two classes, the SVM algorithm is more convenient. On the contrary, NBC is more convenient under the division of multiple tasks.

NBC assumes that attributes are independent of each other. However, the number of attributes in the work is large and there could be some correlation between features in the TD-IDF method, which explains why the accuracy of NBC is lower than SVM. However, considering the high speed and convenience of classification in multiple categories, NBC still has a place in the classification of customer reviews on e-commerce platforms.

The classification accuracy under different test \textit{size} values is shown in Figure 2. It can be seen that the optimal proportion of the test set is about 25% to 35%, which has a high accuracy rate of text classification. Although the highest value of the accuracy appears to be 5%, we do not consider it as the optimal proportion of the test set since the size of samples in the test set is too small.

![Accuracy under different test size values](image)

**Figure 2.** Accuracy under different sizes of test set

5. Conclusion
For the current prosperous e-commerce services and urgent needs for the analysis of on-line reviews, we proposed an identification classifier of e-commerce reviews based on NBC and SVM. We implement the classifier based on NBC and SVM and obtain a fast and accurate classification of e-commerce reviews by adjusting the corresponding parameters. Comparing the performance of two classifiers, we make conclusions and therefore give suggestions to further research based on trade-off decision making: NBC is more suitable to classify the on-line reviews into different categories in the least time, while SVM is more suitable to analyze emotional orientation of on-line customers’ reviews. There are also some challenges to be resolved: although the training speed of NBC is very fast, the possible correlation between features in the TD-IDF method conflicts with the assumption of NBC that all the attributes are independent. It is necessary to improve the TD-IDF algorithm to eliminate the correlation between Chinese text features in the future. In short, the identification classifier in this paper can improve the efficiency of analyzing e-commerce reviews for shoppers and customers.

References
[1] Wang, Z., and J. Liu. "PU Chinese text classifier based on support vector machine construction." J. Nanjing Univ. Posts Telecommun.(Natural Sci. Ed.) 35.6 (2015): 100-105.
[2] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." European conference on machine learning. Springer, Berlin, Heidelberg, 1998.
[3] Yujia, W. U., et al. "Words in Pairs Neural Networks for Text Classification." Chinese Journal of Electronics 29.3 (2020): 491-500.

[4] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." Proceedings of the 12th international conference on World Wide Web. 2003.

[5] Xu, Feng, Zhenchun Pan, and Rui Xia. "E-commerce product review sentiment classification based on a naive Bayes continuous learning framework." Information Processing & Management (2020): 102221.

[6] Haider, Sajjad, et al. "Impact analysis of adverbs for sentiment classification on Twitter product reviews." Concurrency and Computation: Practice and Experience (2018): e4956.

[7] Liming Rong. "Application of Text Mining in E-commerce Review Analysis" MS thesis. Huazhong University of Science and Technology. 2016

[8] Sun, J. "Jieba chinese word segmentation tool." 2018-01-21[2018-06-25]. https://github.com/fxsjy/jieba (2012).

[9] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[10] Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." Journal of machine learning research 2.Nov (2001): 45-66.

[11] Fujino, Akinori, Hideki Isozaki, and Jun Suzuki. "Multi-label text categorization with model combination based on f1-score maximization." Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II. 2008.