Advances in Economics, Business and Management Research, volume 133
3rd International Conference on Advances in Management Science and Engineering (IC-AMSE 2020)

Sentiment Analysis Based on Product Review Data of Chinese Commerce Website of JD

Wenhua Song*, Aiming Qin and Tiansheng Xu
Capital University of Economics and Business, Beijing, China
***Corresponding author**

Keywords: sentiment analysis, machine learning, web spider

Abstract. With the popularity of the Internet and the development of e-commerce, online shopping has become more and more popular. Based on the commodity text comments of an e-commerce website, this paper uses machine learning method to analyze and mine the emotional direction of commodity comments. Finally, combining JIEBA and SNOWNLP, K-Folding Cross-Validation was used to obtain the final emotional score of the product review. The review scores displayed on e-commerce sites are unreliable. Moreover, the positive comment rate of the e-commerce platform is relatively high.

Introduction

Nowadays, China's online shopping market is growing steadily, and its scale is expanding [1]. The report on the market consumption of B2C e-commerce platforms shows that from January to August 2019, China's online retail sales reached 6439.3 billion yuan, up 16.8%. Sentiment analysis is a process of calculating and processing the views, emotions and subjective opinions contained in the text [2].

In foreign countries, text sentiment analysis has begun relatively early in China. Blog Track of TREC mainly looks for information related to ideas contained in the English text [3]. The technology of sentiment analysis in China started relatively late compared with that in foreign countries, but developed rapidly. By comparing the SVM method and the statistical method, Wei jingjing et al. classified the e-commerce comments and reached the conclusion that the SVM effect was better [4]. Xu Peng studied intuitionistic fuzzy reasoning and classified online comments on web pages with high accuracy [5].The above studies did not compare commodity text comments with those displayed on e-commerce platforms, nor did they use the method of combining SNOWNLP corpus with JIEBA segmentation. Therefore, this study has great significance.

Research Method

Data Acquisition and Preprocessing

This study takes JD commodity review as the research object. This paper selects the product reviews of Huawei and Xiaomi, the most popular mobile phones among Chinese, for research.

In this paper, python language is used to write a web spider to get the name of the commodity to publish the comment time commodity comment text, the data is stored in the MySQL database. There are still many problems with the product review data obtained by the crawler, such as invalid comments, emoticons such as “&hellip” entered by the mobile phone user, or invalid information, etc.

Data Cleaning

There is a lot of invalid data in the initial data. Use SQL statements to delete invalid comments directly. The data composition after deleting invalid comments is shown in Table 1.

Table 1. New data sheet

Product	Number	Types
Xiaomi6	981	Positive
Xiaomi6	997	Negative
Xiaomi note3	969	Positive
Honor9 Lite	950	Positive

User-defined Dictionary

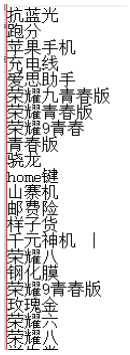

Figure 1. User-defined dictionary

To build a user-defined dictionary is to add some special nouns contained in the Chinese text data to be processed in the dictionary. When JIEBA imports a user-defined dictionary for word segmentation, the word in the dictionary is considered to be a special noun and will not be segmented. As shown in figure 1, JIEBA continues to improve the segmentation results after word segmentation. Part of the user-defined dictionary is built, which mainly includes the name of the phone model and special nouns in the field of mobile phone reviews.

Segmentation

This study uses jieba.cut() for precise modal segmentation of each comment.

Stop Words Dictionary

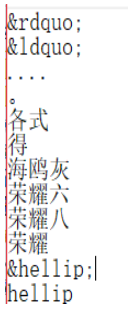

Figure 2. Stop words dictionary

The preprocessing process of this Chinese text is to divide the words first and then remove the stop words (including space return punctuation marks and so on are counted as the stop words). However, the JIEBA package does not have a built-in stop word list for word segmentation. Therefore, based on the comprehensive collection of stop words in Harbin Institute of Technology's extended stop words list, this paper constructs the stop words list by comparing the segmentation results with the stop words list. When constructing the stop words list, this paper uses the first 100 data as the training set, compares the word segmentation results before and after the use of stop words, removes some unnecessary stop words and adds some needed stop words in a manual way. The deleted words are mainly positive emotion words and some degree adverbs. As shown in figure 2, where “&hellip”; It is the emoticon entered by the mobile phone user during the product review. It belongs to nonsense words, so it needs to be deleted.

Sentiment Analysis

K Fold Cross Validation

This study is based on the SNOWNLP library for cross-validation, First, prepare the data. The second step is to conduct 5 fold cross-validation on the segmented data. One is used as the test set and the other four are used as the training set for repeated training and testing. The third step, based on the corpus generated after the training, determines the emotional tendency of commodity comments in the test set and obtains the final data analysis result。

Data Preparation

In this experiment, 4000 pieces of data were divided into 5 pieces. Four pieces were selected as the training set and the remaining one as the test set. The four data sets were divided into two types of training sets: one is the positive emotion corpus training set, that is, all text comments in the negative comment part of mi 6 were removed and stored in the text document of pos.txt; the other is the negative emotion corpus training.

Train and Test the Model

In this paper, thefunction in SNOWNLP is used to train both the positive emotion corpus training set and the negative emotion corpus training set.

The process is as follows.The first step is to use K fold cross-validation training for 5 times, and each training will generate a corpus in. Marshal format. The second step is to use the corresponding test set to test the corpus generated by each training, that is, to evaluate the emotional score of each commodity comment in five different test sets. To get a score for each comment, the closer you get to 1, the more negative the comment emotion is, and the closer you get to 0, the more positive the comment emotion is. The third step is the emotional classification of product reviews. The emotion of commodity review is divided into negative, positive and neutral emotion, and the judgment method is shown in table 2..

Table 2. Affective tendency determination table

emotional tendency	Affective scoring interval	lb
Negative emotions	>=0.6	1
Neutral emotional	<0.6&>0.4	0
positive affect	<=0.4	1

Determination of Results

Table 3. K fold cross validation test results and the actual situation of the emotional ratio table

Test set	Measure the proportion of positive emotions	Actual positive emotion ratio	Measure the proportion of negative emotions	Actual negative emotions
Test 1	69.24%	75.00%	30.76%	25.00%
Test 2	73.46%	75.00%	26.54%	25.00%
Test 3	69.59%	75.00%	30.41%	25.00%
Test 4	71.95%	75.00%	28.05%	25.00%

Five cross test result table distribution of the test sets as shown in table 3, you can see that after the corresponding training set training, each test set between the proportion of positive emotions and negative emotions is almost the same proportion, thus it can be seen that the training process is smooth, algorithm stability.

According to table 3, it can be calculated that the average positive emotions accounted for 70.85%, the average negative emotions accounted for 29.14%, and the actual positive and negative emotions accounted for 75% and 25%. Therefore, it can be seen that the error between the real value and fitting value of the training was around 5%, and the training effect was good.

Conclusion

Comparative Analysis of Algorithm Accuracy

Contrast using SNOWNLP default corpus emotion score results for the training and use of research data as corpus, 5 fold cross-validation training emotional score results, the use of research data as the comments resulting from the training corpus emotion classification accuracy is significantly about 5% higher than the default corpus. Among them 4.64% higher than that of positive emotional comments, 13.88% higher than that of negative emotional comments, negative comment on the accuracy of ascension is obvious, the reason is mainly used as the training of 1000 negative emotional comments may be extreme and concise, it reduces the probability that some neutral emotional comments in the positive comments are wrongly classified as negative emotional comments in the judging process.

Table 4. Classification accuracy table based on default corpus

Emotional direction	The original data	Forecast data	accuracy
positive	75%	67.38%	89.84%
negative	25%	32.61%	69.56%

Table 5. Classification accuracy table based on research data corpus

Emotional direction	The original data	The original data	accuracy
accuracy	75%	70.86%	94.48%
negative	25%	29.14%	83.44%

Compare and Analyze the Emotion of Commodity Review with E-commerce Platform

The three mobile phone products analyzed in this paper are all rated above 95% on e-commerce websites, but there is a certain gap between the actual analysis results and the rated rate, with a maximum difference of nearly 10%. Since the product reviews studied in this paper are only analyzed based on the written comments of customers, the analysis results should be closer to the real feelings of customers on the product reviews, so it can be inferred that the actual praise rate of e-commerce product reviews should be lower than that shown in the e-commerce platform.

Through comparative analysis on the accuracy of the algorithm, JIEBA and SNOWNLP were combined to use the research data as the corpus, and the discriminating accuracy of sentiment analysis was higher after 5 fold cross verification. With JD display goods received the analysis found that the rate in JD recommendation of praise, it’s are shown as praise comment count is very high, but with a text comments on customer analysis, evaluating the actual rate of goods should be lower than the JD shows favorable rating, the e-commerce platform goods high praise rate is not accurate and higher overall. The review scores displayed on e-commerce sites are unreliable.

Acknowledgments

This research was financially supported by National Social Science Fund of China (No.19BXW120).

References

[1] Iresrarch.2019 H1 China e-commerce industry data release report[R/OL]. iResearch Consulting, 2019. 10.

[2] Zhi yan research consulting. Research report on the development status and market prospect of China’s online shopping industry from 2019 to 2025 [R/OL]. Zhiyan consulting group,2019. 4.

[3] Craig Macdonald, Iadh Ounis, Ian Soboroff, Overview of the TREC2007 Blog Track[C] TREC 2007.

[4] Wei jingjing, Wu xiaoyin. Research on Multi-level Sentiment Analysis System of E-Commerce Product Review and Implementation [J].SOFTWARE,2013, 34(9):65-67.

[5] Xu Peng. Classifying emotional tendencies of online comments in webpages based on intuitionistic fuzzy reasoning [J]. Computer Applications and Software. 2013, 30(6):40-42,103.

[6] Chen Ying. Research on Sentiment Analysis Method and Optimization of Electronic Commerce Comments [D].Nanchang University, 2017.

[7] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C] International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. 2010.