Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4591260, 11 pages
https://doi.org/10.1155/2020/4591260

Research Article

Potential Trend for Online Shopping Data Based on the Linear Regression and Sentiment Analysis

Jian Dong, Yu Chen, Aihua Gu, Jingwei Chen, Lili Li, Qinling Chen, Shujun Li, and Qifeng Xun

School of Information Engineering, Yancheng Teachers University, Yancheng 224002, China

Correspondence should be addressed to Aihua Gu; guaihua1978@163.com

Received 6 May 2020; Revised 8 August 2020; Accepted 12 August 2020; Published 26 August 2020

Academic Editor: Yumin Cheng

How to reduce the cost of competition in the industry, identify effective customers, and understand the emotional needs and consumer preferences of customers, so as to carry out fast and accurate commercial marketing, is an important research topic. In this paper, we discussed the method for the analysis of three product data which represent the customer-supplied ratings and reviews for microwave ovens, baby pacifiers, and hair dryers sold in the Amazon marketplace over the time period. The sentiment analysis, linear regression analysis, and descriptive statistics were implemented to analyze the three datasets. Based on the sentiment analysis given by the naive Bayesian classification algorithm, we found that the star rating is positively correlated with the reviews, while the helpfulness ratings have no specific relationship with the star rating and reviews. We use multiple regression analysis and clustering algorithm analysis to get the relationship between the 4 indexes such as time, star rating, reviews, and helpfulness rating. We find that there is a positive correlation between the 4 indexes, and the reputation of the product online market is improving as time grows. Based on the analysis of the positive reviews and star ratings, we suggested indicating a potentially successful or failing product by the positive reviews. We also discussed the relations between the star ratings and number of reviews. Finally, we selected the words from the Amazon sentiment dictionary as candidate words. By counting the candidate words' appearance in the review, the keywords that can reflect the star rating were found.

1. Introduction

1.1. Background. Through the vigorous development of e-commerce in these years, the Internet traffic dividend [1] of the Internet ceased to exist, and the profit cost of merchants has become higher and higher. How to reduce the cost of competition in the industry, identify effective customers, and understand the emotional needs and consumer preferences of customers, so as to carry out fast and accurate commercial marketing, is an important research topic.

The data such as customers' online comments, ratings, and sentiment are important sources of information on this issue. For example, Amazon provides customers with an opportunity to rate and review purchases. Individual ratings—called "star ratings"—allow purchasers to express their level of satisfaction with a product using a scale of 1 (low rated, low satisfaction) to 5 (highly rated, high satisfaction). Additionally, customers can submit text-based messages—called "reviews"—that express further opinions and information about the product. Other customers can submit ratings on these reviews as being helpful or not—called a "helpfulness rating"—towards assisting their own product purchasing decision. Companies use these data to gain insights into the markets in which they participate, the timing of that participation, and the potential success of product design feature choices.

The e-commerce marketing department can use the technology of data analysis to figure out marketing strategies that can maximize profits, thereby saving marketing costs for enterprises. In this paper, we discussed the method for the analysis of three product data which represent the customer-supplied ratings and reviews for microwave ovens, baby pacifiers, and hair dryers sold in the Amazon marketplace over the time period from March 02, 2003, to August 31,

2015. The three products are commonly used household products of which the prices are located in three price ranges, respectively. This will make the analysis more comprehensive and convictive. Our datasets come from open source Amazon Review Dataset [2]. Our datasets have 32022 data points in total. Each example includes the type, name of the product as well as the text review, the star_-rating, and votes.

1.2. Our Works. We are asked to supply the identified key patterns, relationships, measures, and parameters in past customer-supplied ratings and reviews associated with other competing products, so as to inform their online sales strategy and identify potentially important design features that would enhance product desirability. We mainly focus on the tasks in the following.

At first, the sentiment analysis is performed. This is a crucial step to begin our analysis. Online reviews and comments are important information resources and contain a lot of intuitive data. Both online reviews and comments are types of text data which account for over 50% of business data. Text-based sentiment analysis is mainly focused on two technologies: machine learning (naive Bayes [3–6], SVM [7–10], and ME [11, 12]) and emotion analysis based on the emotional lexicon (dictionary-based method [13–15] and corpus-based method [16, 17]). The machine learning-based method is more accurate than emotional lexicon-based method and has been widely used in sentiment analysis. Now, the deep learning technology is also applied in sentiment analysis; however, this technology needs huge amount data for model training and is not suitable for the data analysis of small e-commerce company. In this paper, the naive Bayes classifier is applied to extract the sentiment orientation (positive or negative) from the reviews. We quantified the sentiment orientation into 11 levels ranging from −5 to 5. A smaller level of quantification means more negative sentiments, and a bigger level means more positive sentiments. Quantitative level changes from small to large which means the process of emotional change from negative to positive.

Based on the linear regression analysis of the three datasets, we found that there is a significant linear relationship between the reviews and star rating. With the rise of star rating, the reviews of the products are more positive. Then, we build a multiple linear regression [18] model on the three datasets and found that there is a significant linear relationship between the three datasets.

We implement descriptive statistics [19] to analyze the relationship between the three datasets and their variation trend and then normalize these indexes. Based on the sentiment analysis given by the naive Bayesian classification algorithm, we found that the star rating is positively correlated with the reviews, while the helpfulness ratings have no specific relationship with the star rating and reviews. Moreover, there is no obvious boundary between the reviews, so we regard star rating as the most valuable index.

We use multiple regression analysis and clustering algorithm analysis to get the relationship between the 4 indexes such as time, star rating, reviews, and helpfulness rating. Based on the given data, we use random sampling consistency (RANSAC) algorithm to randomly select the 4 indexes of each year and then calculate the coefficient of multiple regression function by SPSS data processor to get the variation relationships between time and star rating, product rating, and helpfulness rating. We find that there is a positive correlation between time and star rating, product rating, and useful information, and the reputation of the product online market is improving as time grows.

Based on the analysis of the positive reviews and star ratings, we suggested indicating a potentially successful or failing product by the positive reviews. We also discussed the relations between the star ratings and number of reviews.

Finally, we selected the words from the Amazon sentiment dictionary [20] as candidate words. By counting the candidate words' appearance in the review, the keywords that can reflect the star rating were found.

2. Sentiment Analysis

2.1. Data Preprocessing. The sentiment analysis of the reviews is a crucial step of our analysis. To this end, the naive Bayes classifier is used in this section. However, before we get start our tasks, the reviews need to be preprocessed because the reviews always contain a large amount of nonalphabetic characters (e.g., %, $, #, and @) which will cause errors of the word segmentation. Therefore, we get rid of the nonalphabetic characters from the reviews of the datasets, and then, the word segmentation is implemented.

2.2. Naive Bayes Classifier-Based Sentiment Analysis. The naive Bayes classifier is based on the bag-of-words model [21]. With the bag-of-words model, we check which word of the review appears in a positive word set or a negative word set. If the word appears in a positive word set, the total score of the review is updated with +1 and vice versa. If at the end the total score is positive, the review is classified as positive, and if it is negative, the text is classified as negative. The probability a review belongs to a class C_k is given by the class probability $P(C)$ multiplied by the products of the conditional probabilities of each word for that class:

$$P_k = P(C_k) \cdot \prod_i P(d_i|C_k) = P(C_k) \cdot \prod_i^n \frac{\mathrm{count}(d_i, C_k)}{\sum_i \mathrm{count}(d_i, C_k)} = P(C_k) \cdot \prod_i \frac{\mathrm{count}(d_i, C_k)}{V_{C_k}}, \tag{1}$$

where k is the class label, $\text{count}(d_i, C)$ is the number of occurrences of word d_i in class C, V_C is the total number of words in class C, and n is the number of words in the review we are currently classifying.

In our tasks, we classified the sentiments into 11 levels labeled by −5, −4, −3, −2, −1, 0, 1, 2, 3, 4, and 5. A smaller label denotes more negative sentiment, while bigger one denotes more positive. Thus, the model can describe the sentiments more accurately.

2.3. Model Establishment

Step 1: data preprocessing as what we mentioned in Section 3.1

Step 2: segment the word and build the training set based on the dictionary called Amazon Product Review Data which is downloaded from the website "https://github.com/uiuc-cs498/amazon-product-sentiment-analysis"

Step 3: train the NB classifier

Step 4: analyze the reviews by NB classifier

2.4. Experiments and Discussion. Table 1 lists some sentiment analysis of the reviews (row number ranges from 4748 to 4767) in data file hair_dryer.tsv.

The first column in Table 1 is the row number of the test reviews; the second column shows the star rating of the reviews. The 11 columns in the right part of Table 1 show the probabilities of each class. Generally, the analysis is acceptable referring to the star ratings. However, some reviews are quite different from the star ratings, e.g., row numbers 4760 and 4765. This is mainly because of the fact that the positive words in the reviews are not contained in the dictionary. The word number in the dictionary is limited. It cannot completely cover all of the English words. Therefore, we must filter the words in the review that does not belong to the dictionary. The words which are filtered out may have an important effect on the sentiment analysis. The second reason is that some customers' reviews have little correlation with their star ratings. They may give a positive review followed by a low star rating, or vice versa. However, it can be seen that star rating is proportional to the class label.

3. Task I

3.1. The Foundation of Model. To investigate the relationship between two or more variables, regression model [22–24] is a powerful tool. There are innumerable forms of regression models which can be performed, for example, linear regression, multiple linear regression, ridge regression, elasticNet regression, and multinomial logit regression.

In this paper, we filter out some unrepresentative and invalid data, for example, the records without reviews. And then, the unitary and multiple linear regression models are established.

The general form of multiple linear regression model is written as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon. \tag{2}$$

In formula (2), $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are coefficients to be determined by regression. Y is the response variable. $x_1, x_2, \ldots, x_p$ are explanatory variables that can be measured or controlled, called independent variable. ε is model's error term (also known as the residuals). The multiple linear regression model is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

3.2. Model Establishment

Step 1: implement unitary linear regression analysis on each pair of the three datasets. As a result, only the group of star ratings and reviews exhibits a significant linear relationship.

Step 2: implement multiple linear regression analysis of star ratings, helpfulness ratings, and reviews.

3.3. Experiments and Discussion. Tables 2 and 3 show the results of unitary and multiple linear regression. As to multiple linear regression analysis, microwave, hair dryer, and pacifier are all passing the significance test because the P values are all less than 0.05. It shows that when star ratings and reviews increase, the helpfulness ratings also increase. The multiple R-squared value of microwave is 0.1458. The residual standard error of microwave is 1.521. So the standard error is 1.521. The multiple R-squared value of hair_dryer is 0.1133. The residual standard error of hair_dryer is 1.225. So the standard error is 1.225. The multiple R-squared value of pacifier is 0.03709. The residual standard error of pacifier is 1.559. So the standard error is 1.559. Multiple R-squared values of the three datasets are all less than 0.8, which means the multicollinearity between the variables is weak.

R square (r square value) is the coefficient of determination, which means the percentage of changes in dependent variables that can be explained by the model you fit. In Tables 2 and 3, the value of R-squared is very low. This means that the predicted points are quite different from the actual points, and there is a lack of fitting. Because of the low r square value, the residual will be high. There are three reasons. At first, because of the influence of other independent variables, if some variables have been proved to be related to the independent variables in this paper, they must be introduced as control variables, even if they have nothing to do with the research hypothesis. The second is the influence of system error. In fact, it is also an independent variable with a specific meaning. The third is the influence of random error. Therefore, the size of R square mainly affects the accuracy of the model rather than its correctness.

In multiple regression analysis, the regression coefficient means that when other forecast variables keep unchanged and a certain forecast variable increases, the value of the dependent variable increases. Since there have some

Table 1: Sentiment analysis of the reviews (row number ranges from 4748 to 4767) in data file hair_dryer.tsv.

Row number	Star ratings	Class labels										
		−5	−4	−3	−2	−1	0	1	2	3	4	5
4748	5	0	0	0	0	0	0	0	0	1	0	0
4749	4	0	0	0	0	0.2	0	0.2	0.2	0.4	0	0
4750	4	0	0	0	0	0.33	0	0.33	0	0.33	0	0
4751	5	0	0	0	0	0	0	0.33	0.33	0.33	0	0
4752	4	0	0	0	0	0	0	0	0	1	0	0
4753	4	0	0	0	0	0.14	0	0.28	0.43	0.14	0	0
4754	5	0	0	0	0	0.25	0	0	0.5	0.25	0	0
4755	5	0	0	0	0	0	0	0	0	1	0	0
4756	3	0	0	0	0	0	0	0.5	0.5	0	0	0
4757	5	0	0	0	0	0	0	0	0.5	0.5	0	0
4758	5	0	0	0	0	0	0	0	0	1	0	0
4759	5	0	0	0	0	0	0	0	0	1	0	0
4760	5	0	0	0	0	1	0	0	0	0	0	0
4761	5	0	0	0	0	0	0	0	0	0	0	0
4762	5	0	0	0	0	0.33	0	0	0	0.67	0	0
4763	5	0	0	0	0	0	0	0	0.5	0.5	0	0
4764	5	0	0	0	0	0	0	0.6	0	0.4	0	0
4765	4	0	0	0	0	1	0	0	0	0	0	0
4766	4	0	0	0	0	0.33	0	0	0	0.67	0	0
4767	5	0	0	0	0.4	0	0	0	0.4	0.2	0	0

Table 2: Unitary linear regression analysis on star_ratings and reviews of the three datasets.

Datasets	P values	R-squared	Adjusted R-squared	Residual standard error	Univariate linear regression equation
Oven	$<2.2e-16$	0.1454	0.1449	0.9787	$y = 0.24532x_1 + 1.97062$
Hair dryer	$<2.2e-16$	0.112	0.1119	1.225	$y = 0.40317x_1 + 2.88024$
Pacifier	$<2.2e-16$	0.02993	0.02988	1.565	$y = 0.24185x_1 + 357530$

Table 3: Multiple linear regression analysis on star_ratings and reviews of the three datasets.

Datasets	P values	R-squared	Adjusted R-squared	Residual standard error	Multiple linear regression equation
Oven	$<2.2e-16$	0.1458	0.1448	1.521	$y = 0.0007589x_1 + 0.5929x_2 + 1.7724$
Hair dryer	$<2.2e-16$	0.1133	0.1132	1.225	$y = -0.003337x_1 + 0.4021x_2 + 2.8906$
Pacifier	$<2.2e-16$	0.03709	0.03698	1.559	$y = 0.02338x_1 + 0.2445x_2 + 3.5479$

unrepresentative and invalid data, we deleted them for preprocessing. Based on the unitary linear regression analysis on each pair of the three datasets, we found that "star ratings" and "reviews" have a significant linear relationship. This means the higher the "star_ratings" is, the more the positive sentiment. The e-commerce should pay more attention to the products of high star ratings.

Table 4: The list of descriptive statistics.

Index	Descriptive statistics
1	Distribution of each product
2	Useful evaluation number (UER)
3	Average customer sentiments on certain product over year
4	Percentage pie chart of star ratings for certain product

4. Task II

4.1. Most Informative Data Measures. We make descriptive statistics on the three datasets and analyze the relationship between the three datasets. We calculated the star distribution of star ratings, then analyzed the basic characteristics of these data with Excel and SPSS software, such as range, standard deviation, and variance, and finally obtained the ratio of "helpful votes" and "total votes" (RHT). The descriptive statistics of the data are listed in Table 4.

In Table 5, the left column is the total votes of each product_ID given by the data file, and the right is the summation of reviews of each product_ID. Figure 1 gives the useful evaluation number (UER) which is defined as the number of helpful votes in certain RHT region. For example, in the RHT region (0.8, 1], we got 2500 helpful votes. Table 5 and Figure 1 show that most of the products are not voted. Nearly half of the product reviews have total votes between 0 and 100, and few products have more than 100 votes. This means most of the customers tend to give up voting. Therefore, the evaluation data of useful information are incomplete and one-sided.

Table 5: Distribution of the total number of product reviews.

Total_votes	Sum
0	7141
0–100	4293
100–400	34
400–600	2
>600	0

Figure 2 shows the average customer sentiments on certain product over year. We can find out the sentiments can reflect the quality of the product as a whole, but the boundary of text evaluation attitude is not very clear, and the differences between them are very small, which is not conducive to quantitative evaluation.

Figure 3 gives the percentage pie chart of star ratings. The proportion of five-star rating products is 58%, followed by four-star products which account for 18%. The proportion of three-star, two-star, and one-star products is less than 10%. The differences between different star ratings are very clear. It can be seen that consumers have a high evaluation of commodities, and star rating can clearly and quantitatively show the value of commodities.

Every product will have star ratings and commodity evaluation, but only some products have useful commodity evaluation information. Star rating is positively related to the trend of commodity evaluation. However, commodity evaluation is subjective. Sometimes customers make fake comments for some purpose. Amazon uses machine learning models instead of raw data averages to calculate star ratings for products. This means commodity evaluation is not suitable for qualitative analysis, so star rating is the most valuable index. This conclusion is consistent with what we conclude in Section 3.

4.2. Time-Based Measures and Patterns within Each Dataset. We cluster the total sales of commodities (as shown in Figure 4) in each year and get the icicle chart (as shown in Figure 5). In Figure 5, index Num denotes the total sales of each year. According to the hierarchical clustering analysis, we can divide the online sales of products from 2002 to 2015 into four categories: 2015, 2002–2005, 2006–2009, and 2010–2014. We found that the total number of goods purchased in 2002–2005 was very small, which increased year by year in 2006–2014 and decreased in 2015 compared with 2014. Based on the investigation, we found that the rapid popularization of the Internet and the rapid development of online shopping malls in 2006–2014 and a series of macroeconomic and financial policies and reform measures launched by the international community may be the cause for the increase of the purchase volume of commodities in 2006–2014.

Figure 6 shows the summation of helpful votes and total_votes over year. In Figure 6, the horizontal and vertical coordinates present the year and summation of helpful votes, respectively. Figure 6 shows that the proportion of five-star products is 58%, followed by four-star products which account for 18%, and the proportion of three-star, two-star, and one-star products is less than 10%.

Figure 7 presents the star rating variation over year. From 2002 to 2007, the proportion of stars (the vertical coordinates) has changed a lot. From 2007 to 2015, the proportion of stars has changed steadily, among which the proportion of five stars has increased gradually.

The general trend of the total number of votes and helpful votes is positively related to the general sales volume. With the increase in the number of purchased commodities over time, consumers' commodity evaluation attitude is becoming more and more positive. Among them, the five-star ratio has been higher than other stars over the years. After 2007, the proportion of four-star products is higher than that of one-star products, higher than that of three-star and higher than that of two-star products. It can be seen that the reputation of commodities is improving.

4.3. Relationship between Star Ratings and Reviews. For the data of the three charts given, vine (string): customers are invited to become Amazon vine voices, which is based on the trust they have won in the Amazon community, because they have written accurate and insightful comments. It can be inferred that under the condition of vine $= y$, the selected comments are those written by the "specific star" mentioned in the title. Through the analysis of the three charts, it can be concluded that the level of star rating is corresponding to the positive and negative aspects of the review. The lower the star rating is, the more negative the review is. On the contrary, the higher the star rating is, the more positive the review is. Therefore, in the process of research, we can use star rating to replace the positive and negative aspects of the review. We can use helpful_votes to divide total_votes to calculate the value and compare it with 50%; if it is higher than 50%, it is a positive impact, and if it is lower than 50%, it is a negative impact. When the positive impact is higher than the negative impact, the comments written by these specific stars are a positive impact.

The datasets are analyzed as following steps:

Step 1: vine (string): customers are invited to be Amazon vine voices based on the trust they have earned in the Amazon community because they have written accurate and insightful comments. It can be inferred that under the condition of vine $= y$, the selected comments are those written by the "specific star" mentioned in the title.

Step 2: calculate the proportion of star rated reviews for different types of products when vine is n and Y, respectively.

Step 3: we can use helpful_votes to divide total_votes to calculate the value and compare it with 50%.

Tables 6 and 7 present the ratio of star rating reviews of hair dryer, oven, and pacifier, respectively, with different vine values.

From the above data, it can be seen that online customer reviews transfer other information to consumers except for commodity attribute information, so that consumers can better understand the quality and performance of products and avoid decision-making errors caused by information asymmetry, which has become the most important information for online shopping users before making purchase decisions. Online reviews not only have an important

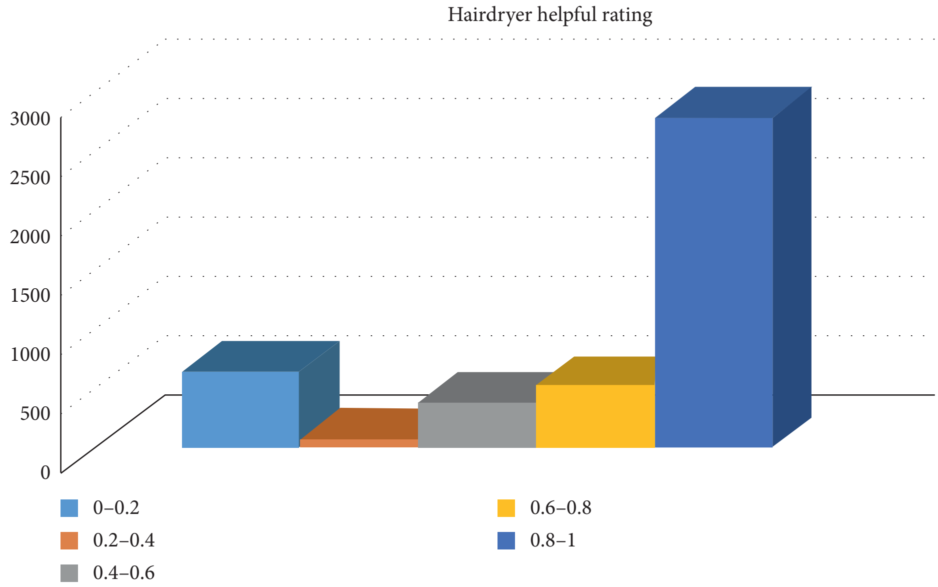
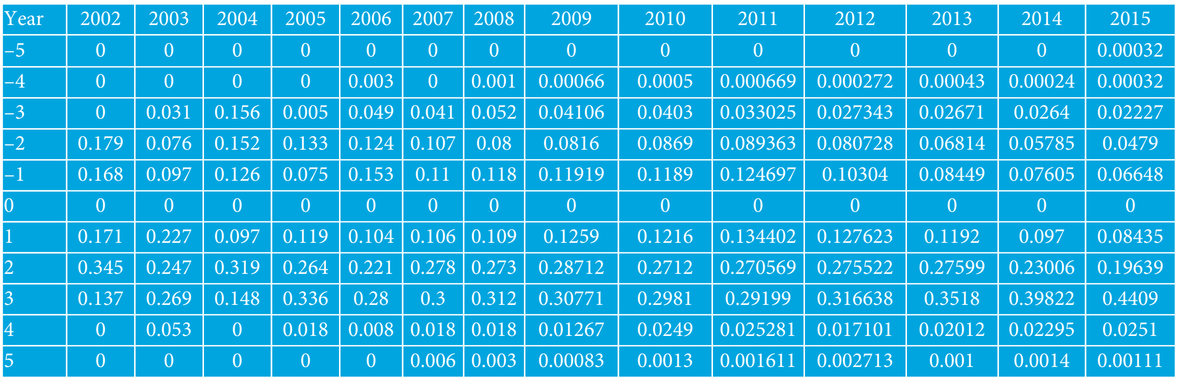


Figure 1: Useful evaluation number.



Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
–5	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00032
–4	0	0	0	0	0.003	0	0.001	0.00066	0.0005	0.000669	0.000272	0.00043	0.00024	0.00032
–3	0	0.031	0.156	0.005	0.049	0.041	0.052	0.04106	0.0403	0.033025	0.027343	0.02671	0.0264	0.02227
–2	0.179	0.076	0.152	0.133	0.124	0.107	0.08	0.0816	0.0869	0.089363	0.080728	0.06814	0.05785	0.0479
–1	0.168	0.097	0.126	0.075	0.153	0.11	0.118	0.11919	0.1189	0.124697	0.10304	0.08449	0.07605	0.06648
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0.171	0.227	0.097	0.119	0.104	0.106	0.109	0.1259	0.1216	0.134402	0.127623	0.1192	0.097	0.08435
2	0.345	0.247	0.319	0.264	0.221	0.278	0.273	0.28712	0.2712	0.270569	0.275522	0.27599	0.23006	0.19639
3	0.137	0.269	0.148	0.336	0.28	0.3	0.312	0.30771	0.2981	0.29199	0.316638	0.3518	0.39822	0.4409
4	0	0.053	0	0.018	0.008	0.018	0.018	0.01267	0.0249	0.025281	0.017101	0.02012	0.02295	0.0251
5	0	0	0	0	0	0.006	0.003	0.00083	0.0013	0.001611	0.002713	0.001	0.0014	0.00111

Figure 2: Average sentiments of consumers over year.

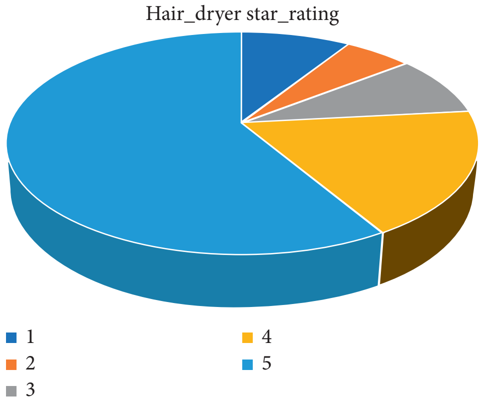


Figure 3: Percentage pie chart of star ratings for hair dryer.

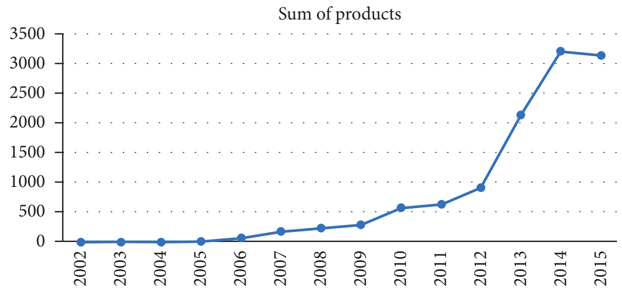


Figure 4: Total sales over year.

reference value for consumers but also have an impact on their subsequent comments. Based on the fine processing possibility model, it is concluded that high-quality reviews have a positive impact on Website Trust, and Website Trust has a positive impact on consumer reviews. Therefore, e-commerce websites can improve the comment system, rank high-quality comments first, and invite specific stars to write positive impact to guide consumers and encourage them to publish high-quality comments (in the pacifier table, some star ratings are beyond the normal range of star rating 1–5, and there are blank star ratings. These data will affect the research on the correct results, which is invalid).

```
Case        0         5         10        15        20        25↵
Label  Num  +---------+---------+---------+---------+---------+↵
  7         -+↵
  8         -+---+↵
  9         -+    +-----------------------------------+↵
  1         ----+                                     |↵
  2         -+---------+                              |↵
  6         -+          +-----------------------------+
  3         -+          |↵
  4         -+---------+↵
  5         -+↵
```

Figure 5: Icicle tree view.

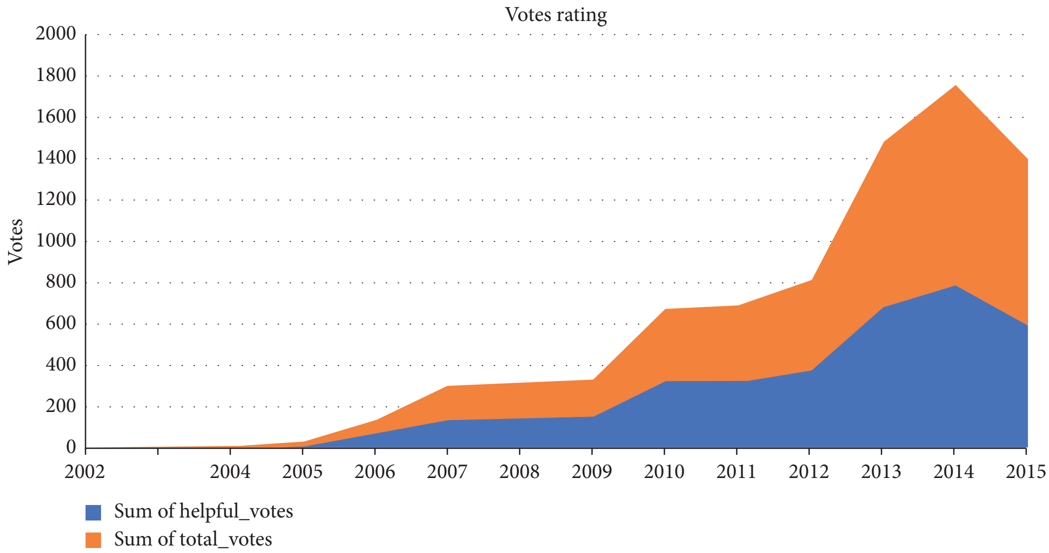


Figure 6: Summation of helpful votes and total_votes over year.

4.4. Relationship between Quality Descriptors of Text-Based Reviews and Rating Levels. We selected the words from the dictionary as candidate words and count the candidate words' appearance in the review. We define the top 50 words with the highest word frequency as set A, while set B is defined as the top 50 words with the highest word frequency in the reviews which have star ratings equal to 1. Similarly, sets C, D, E, and F composed of the top 50 words with the highest word frequency in reviews which have star ratings of 2, 3, 4, and 5, respectively. Finally, the intersection of A, B, C, D, E, and F will reflect the relations between quality descriptors and rating levels (25 words are contained in the intersection). Table 8 shows the 24 words of the intersection. Figure 8 gives the curve of the word frequency versus star ratings.

Table 8 indicates an obvious positive correlation between rating levels and quality descriptors such as "great," "like," and "love". As shown in Figure 8, we present the word frequency (vertical coordinates) of 24 words (horizontal coordinates) in the intersection of sets A, B, C, D, E, and F. The word frequency in the 6 sets is displayed in different colors. Figure 8 indicates that a more positive quality descriptor will lead to a higher rating level. With the decrease of the positivity of quality descriptors, costumers tend to select the star rating in the 5 choices with similar possibilities.

5. Strength and Weakness

5.1. Strength. Perhaps the biggest strength of our method is the NB classifier which is applied in this paper to extract the sentiment orientation (positive or negative) from the reviews. This method enables our model to use the online text data and makes the follow-up model more simple.

5.2. Weakness. The accuracy of our model is influenced by the vocabulary and classification accuracy of the sentiment dictionary.

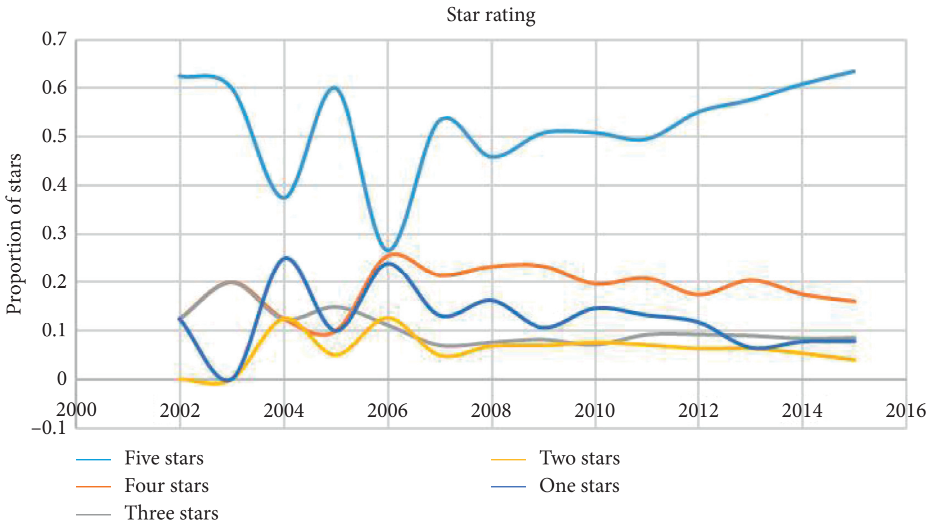


Figure 7: Star rating variation over year.

Table 6: Ratio of star reviews when vine = y.

Product	Total review	Review number (vine = y)	Star_rating	Sum	Proportion
Hair_dryer	1615	19	3	2	0.11
			4	8	0.42
			5	9	0.47
Microwave	11470	179	2	3	0.02
			3	20	0.11
			4	51	0.28
Pacifier	18937	148	1	14	0.1
			2	6	0.04
			3	14	0.09
			4	43	0.29
			5	71	0.48

Table 7: Ratio of star reviews when vine = n.

Product	Total review	Review number (vine = y)	Star_rating	Sum	Proportion
Hair_dryer	1615	1596	1	402	0.25
			2	112	0.07
			3	132	0.08
			4	292	0.19
			5	658	0.41
Microwave	11470	11291	1	1032	0.09
			2	636	0.06
			3	979	0.09
			4	2045	0.18
			5	6599	0.58
Pacifier	18937	18744	1	1182	0.06
			2	939	0.05
			3	1410	0.08
			4	2664	0.14
			5	12549	0.67

6. Further Work

Our future work will mainly focus on big data mining based on deep learning technology. In addition to sentiment analysis, it will also include mining customer needs from text data, trial experience perception, and consumption tendency.

7. Conclusion

In this paper, the naive Bayes (NB) classifier is applied to extract the sentiment orientation (positive or negative) from the Amazon product reviews. The sentiment orientation is quantified into 11 levels. Based on the linear regression and

Table 8: The 24 words of the intersection of sets A, B, C, D, E, and F.

Words	Word frequency					
	Frequency in reviews	Star ratings				
		1 star	2 stars	3 stars	4 stars	5 stars
Great	7878	213	231	377	1369	5688
Like	6695	549	514	813	1516	3303
Love	6191	90	74	206	730	5091
Good	4557	265	237	428	1166	2461
Well	4299	230	222	431	887	2529
Easy	3622	52	84	165	688	2633
No	3461	511	244	312	545	1849
Nice	2277	73	89	207	543	1365
Recommend	2276	155	106	80	322	1613
Cute	2049	53	85	183	294	1434
Happy	1369	64	51	53	243	958
Loved	1315	91	59	71	131	963
Fit	1208	106	84	149	252	617
Want	1205	130	87	137	232	619
Hard	1183	140	111	156	240	536
Pretty	1158	68	64	157	342	527
Worth	1105	123	62	68	150	702
Fine	1075	98	119	194	262	402
Wish	1029	80	47	129	319	454
Problem	1012	180	93	140	213	386
Cool	831	58	51	100	167	455
Top	683	96	61	72	153	301
Liked	655	64	54	92	136	309
Help	604	70	43	42	108	341

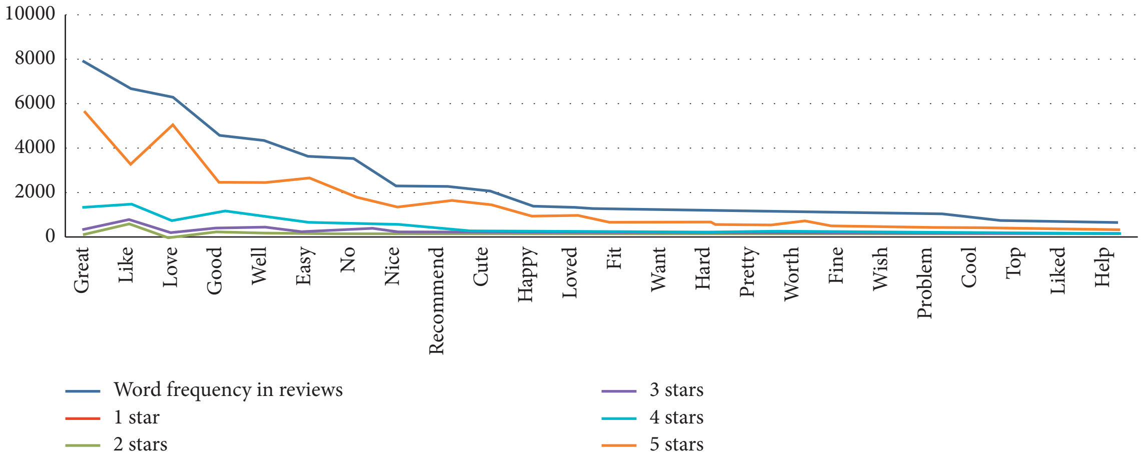


Figure 8: The curve of the word frequency versus star ratings.

multiple linear regression models, we analyzed the three product datasets (microwave oven, baby pacifier, and hair dryer) to provide meaningful quantitative relationships between star ratings, reviews, and helpfulness ratings that will help the e-commerce company succeed in their online marketplace product offerings. Descriptive statistics was used to analyze the relationship between the three datasets and their variation trend. Our analysis showed that the star rating is positively correlated with the reviews, while the helpfulness ratings have no specific relationship with the star rating and reviews. This means the star rating is actually the most valuable index. The multiple regression and clustering algorithm analysis give the relationships between the 4 indexes such as time, star rating, reviews, and helpfulness rating. We find that there is a positive correlation between time and star rating, product rating, and useful information, and the reputation of the product online market is improving as time grows. This means people will tend to get used to using new products.

According to our data analysis, “star rating” and “reviews” of the online information are consistent on the whole, and “star rating” can qualitatively and accurately describe product information. The sales volume of hair dryer, microwave oven, and baby pacifier is increasing year by year. Meanwhile, the satisfaction of buyers to the products is also increasing. Comments and helpful comments are also increasing. We suggest that if an e-commerce company wants to further understand the product information accurately, it should take the star rating of the product as the measurement standard. At the same time, the improvement of the three datasets of "star rating," "comment," and "help rating" indicates the improvement of the product online market reputation, and in this case, we can appropriately increase the type and scale of online market products. We also suggest focusing on the 2 important features of your products, such as the convenience of the operation/store and the aesthetics of appearance design. That is mainly because the two most frequent keywords on product experience are "easy" and "cute." The two features gain the highest star ratings.

Our method can be used to deal with any online product reviews not only for Amazon. The online product review analysis can help us to reduce the cost of competition in the industry and understand the emotional needs and consumer preferences of customers, so as to carry out fast and accurate commercial marketing.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was funded by the National Natural Science Foundation of China under grant no. 61602400. This work was also supported by the National Spark Program of China (no. 2015GA690259).

References

[1] R. Kathuria, M. Kedia, R. Sekhani, and U. Krishna, *Growth Dividends of Digital Communications: The Case for India*, Broad India Forum, 2018.

[2] J. Ni, J. Li, and J. McAuley, "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects," *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 188–197, 2019.

[3] A. Goel, J. Gauta, and S. Kumar, "Real time sentiment analysis of tweets using naive bayes," in *Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, IEEE, Dehradun, India, October 2016.

[4] H. Sudira, A. L. Diar, and Y. Ruldeviyani, "Instagram sentiment analysis with naive bayes and KNN: exploring customer satisfaction of digital payment services in Indonesia," in *Proceedings of the 2019 International Workshop on Big Data and Information Security (IWBIS)*, Bali, Indonesia, October 2019.

[5] H. Parveen and S. Pandey, "Sentiment analysis on twitter data-set using naive bayes algorithm," in *Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, IEEE, Bangalore, India, July 2017.

[6] R. S. Surendra, *Simple Text Mining for Sentiment Analysis of Political Figure Using Naive Bayes Classifier Method*, pp. 99–104, Faculty of Information and Communication Technology, Multimedia Nusantara University, Tangerang, Indonesia, 2013.

[7] N. Sumathi and T. Sheela, "An efficient sentiment analysis by using hybrid naive bayes and SVM approach in banking institutions," *International Journal of Civil Engineering and Technology*, vol. 8, no. 12, pp. 373–391, 2017.

[8] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World Wide Web*, vol. 20, no. 2, pp. 135–154, 2017.

[9] R. Moraes, J. F. Valiati, W. P. Gavião Neto et al., "Document-level sentiment classification: an empirical comparison between SVM and ANN," *Expert Systems With Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[10] F. Luo, C. Li, Z. Cao et al., "Affective-feature-based sentiment analysis using SVM classifier," in *Proceedings of the 2016 IEEE 20th International Conference on Computer Supported Co-operative Work in Design (CSCWD)*, pp. 276–281, Nanchang, China, 2016.

[11] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," in *Proceedings of the 2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 1–6, Dhaka, Bangladesh, 2014.

[12] B. Pang, L. Lee L, S. Vaithyanathan et al., "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Empirical Methods in Natural Language Processing*, pp. 79–86, 2002.

[13] J. Jiao and Y. Zhou, "Sentiment polarity analysis based multi-dictionary," *Physics Procedia*, vol. 22, pp. 590–596, 2011.

[14] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[15] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2010.

[16] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, DBLP, Valletta, Malta, May 2010.

[17] A Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, https://arxiv.org/abs/1606.06259.

[18] L. E. Eberly, "Multiple linear regression," *Topics in Biostatistics*, vol. 404, no. 2, pp. 165–187, 2007.

[19] R. C. Taeuber, "Basic data - descriptive statistics," *International Journal of Educational Research*, vol. 11, no. 4, pp. 397–401, 1987.

[20] https://github.com/uiuc-cs498/amazon-product-sentiment-analysis.

[21] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1–4, pp. 43–52, 2010.

[22] J. Cheng, “Analyzing the factors influencing the choice of the government on leasing different types of land uses: evidence from Shanghai of China,” *Land Use Policy*, vol. 90, Article ID 104303, 2020.

[23] J. Cheng, “Data analysis of the factors influencing the industrial land leasing in Shanghai based on mathematical models,” *Mathematical Problems in Engineering*, vol. 2020, Article ID 9346863, 11 pages, 2020.

[24] W. Pan, H. Ming, C. K. Chang, Z. Yang, and D.-K. Kim, “ElementRank: ranking java software classes and packages using multilayer complex network-based approach,” *IEEE Transactions on Software Engineering*, 2019.