World Scientific Research Journal Volume 6 Issue 12, 2020
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

Data Utilization in Online Marketplace

Xuantao Zeng, Nancy Shi, Steven Tong
Shenzhen Middle School, China

Abstract

As a consultant hired by Sunshine Company, our team do our best to help Sunshine Company to inform their online sales strategy and identify potentially important design features that would enhance product desirability, including specific justification that our team most confidently recommends to the Marketing Director. Overall, our team build a online review data processing model, which adopt with single factor variance, principal component analysis, correlation analysis, and decentralization methods. For requirement 1, We pre-processed, quantify and evaluate the time base data, and the data excluding invalid data, blank data, and comments from people who did not purchase items in fact. According to the star ratings, helpful votes, total votes, reviews and other factors of data, using principal component analysis to figure out the main influencing factors are star ratings and comment length. For requirement 2a, we normalize the data and use entropy deweighting to weigh each factor, and get the star rating, total votes, and the comment length is 0.5064, 0.3025, 0.1911, respectively. According to the weights, the comprehensive coefficients of hair dryer, microwave, and pacifier are 2.8394, 3, and 2.815, respectively. The main product to be tracked is microwave. For 2b, we preprocess the data and perform a linear fit to get the relationship between the year and the star ratings of hairdryer, microwave and pacifier. For 2c, the data is integrated, the length of the reviews are used. We use the star ratings of the reviews as independent variables, and helpful reviews as dependent variables to do the correlation analysis. It is found that the comments in the correlation analysis table below are more relevant to star rating comments. For 2d, statistically process data is used to make a line chart between star ratings and review length. For 2e, the "enthusiasm" and "disappointment" words are classified and counted in hair dryer, microwave and pacifier to obtain their respective proportions. For requirement 3, the letter to the Marketing Director of Sunshine Company summarizing our team's analysis and results, which is at the end of the paper.

Keywords

ANOVA; PCA; correlation analysis; entropy method; normalization.

1. INTRODUCTION

Amazon transaction data has great value. As an "information company", Amazon not only obtains information from each user's purchase behavior, but also records all the behaviors on their website, such as time on page, whether users view reviews or not, keywords searching, products viewed, etc. With high sensitivity and emphasis on data value, as well as strong mining capabilities, these have made Amazon far beyond its traditional way of operating. Many e-commerce companies can use the value of big data to make data-driven decisions, mainly by increasing the probability of prediction to improve the success rate of decisions. For data-driven products, the focus should be on personalization during the product design phase. During the product operation stage, iterative innovation is emphasized. [1] With the increase in the number of online reviews, in order to make it easier for consumers to obtain useful information,

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

major e-commerce sites have changed from merely presenting reviews to identifying the usefulness of reviews. Placing highly useful reviews in front can greatly reduce the time cost of consumers, and improve the efficiency of decision-making.

1.1. Problem Review

Sunshine Company is planning to introduce and sell three new products in the online marketplace: a microwave oven, a baby pacifier, and a hair dryer. They have hired our team as consultants to identify key patterns, relationships, measures, and parameters in the past customer-supplied ratings and reviews associated with other competing products to 1) inform their online sales strategy and 2) identify potentially important design features that would enhance product desirability. Sunshine's data center has provided three data files for this project. These data represents customer-supplied ratings and reviews for microwave ovens, baby pacifiers, and hair dryers sold in the Amazon marketplace over the time period(s).

1.2. Decompose the Problem and Our Tasks

To totally solve the problem better and provide support for Sunshine Company, we need to decompose the problem. According to the requirements and specific questions from the Sunshine Company Marketing Director, we can divide the requirement 2 into 5 tasks:

Task1: Identify data measures based on ratings and reviews that are most informative to track.

Task2: Identify and discuss time-based measures and patterns within each data set.

Task3: Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product.

Task4: Do specific star ratings incite more reviews?

Task5: Are specific quality descriptors of text-based reviews strongly associated with rating levels?

1.3. Process Analysis Diagram

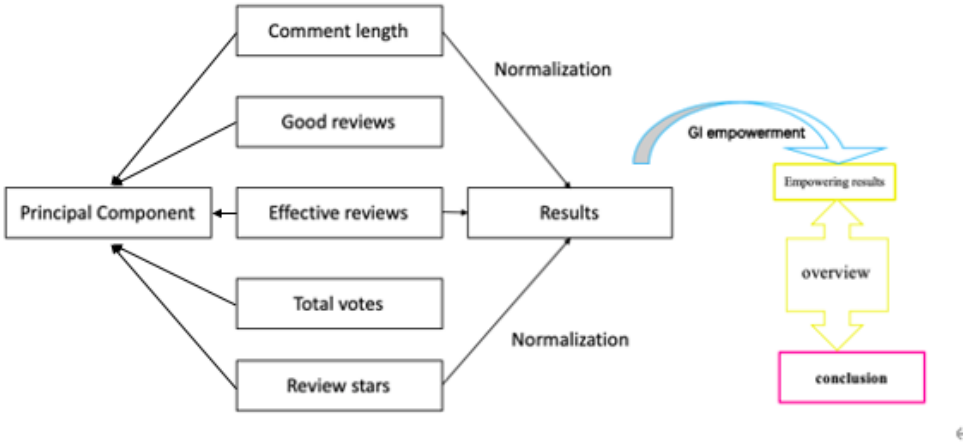

Figure 1. Process analysis diagram

2. ASSUMPTIONS AND JUSTIFICATIONS

Due to the data limitation:

a. We suppose that comment length has impact on comment usefulness.

b. We assume that buyer reviews are authentic.

c. We assume that no errors in data processing.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

We will list more detailed assumptions in the following sections when using them.

3. GENERAL SYMBOLIC DESCRIPTION

Table 1. Symbolic description table

order number	symbol	symbol description
1	Σ=(sij)p×p	Covariance matrix
2	λ_i	Eigenvalues
3	a_i	Unit norm eigenvector
4	f_i	i^{th} principal component
5	x_j	Independent variable
6	l_{ij}	Factor loadings
7	σ	Standard Deviation
8	μ	Measured average
9	σ^2	variance
10	σ_0^2	Variance from multiple determinations
11	R	correlation matrix
12	v_j	Weight
13	e_j	Entropy
14 15	p_{ij}	Proportion
16	x_{k-1}, x_k	adjacent rule layer
17	$x_i^{'}$	input variable for multivariable linear regression
18	$y_i^{'}$	output variable for multivariable linear regression
19	x_i	input variable for multivariable non-linear regression
20	y_i	output variable for multivariable non-linear regression
21	β	regression coefficient
22	$\hat{\beta}$	Estimated regression coefficient
23	$\hat{y}$	Estimated output variable

4. ANALYSIS AND MODELING OF REQUIREMENT 1

4.1. Introduction

Firstly, we use qualitative analysis to summarize the factors that affect online reviews. Secondly, we deal with the factors with quantitative methods. Finally, quantitative data is used for analysis. The main research factors include star ratings, help votes, total votes, comment head length as independent variables, help votes as dependent variables, and finally perform principal component analysis.

We pre-processed the data excluding invalid data, blank data, and comments from people who did not purchase items in fact. We also quantify the evaluation and time base data. According to the star ratings, helpful votes, total votes, reviews and other factors of data,

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

principal component analysis is used to figure out that the main influencing factors are star ratings and comment length. [2]

4.2. Construction of the Model

(1) Calculate covariance matrix

Calculate the covariance matrix of the sample data: Σ=(sij)p×p, where

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \qquad \text{i,j=1,2,...,p} \tag{1}$$

(2) Find the eigenvalues of λ_i and the corresponding orthogonal unit eigenvectors a_i, The first m large eigenvalues of λ1≥λ2≥...λm>0, which is the variance corresponding to the first m principal components, the corresponding unit feature vector λ_i is the coefficient on the original variable of the principal component f_i, then the a_i principal component f_i of the original variable is:

$$f_i = a_i'X \tag{2}$$

The variance (information) contribution rate of the principal component is used to reflect the amount of information, which is:

$$\alpha_i = \lambda_i / \sum_{i=1}^{m}\lambda_i \tag{3}$$

(3) Choosing principal components

Finally, several principal components are selected, that is, the determination of m in f_1, f_2, ..., f_m is determined by the cumulative contribution rate of variance G (m)

$$G(m) = \sum_{i=1}^{m}\lambda_i / \sum_{k=1}^{p}\lambda_k \tag{4}$$

When the cumulative contribution rate is greater than 85%, it is considered to be sufficient to reflect the information of the original variable, and the corresponding m is the first m principal components extracted.

(4) Calculating principal component loads

The principal component load reflects the degree of correlation between the principal component $\boldsymbol{f_i}$. and the original variable x_j. The original variable x_j (j=1,2 ,..., p) is on the principal components $\boldsymbol{f_i}$(i=1,2,...,m)Load l_{ij} (i=1,2,...,m; j=1,2 ,...,p):

$$l(Z_i, X_j) = \sqrt{\lambda_i}a_{ij}(i = 1, 2, \cdots, m;\ j = 1, 2, \cdots, p) \tag{5}$$

(5) Calculating the principal component score

Calculate the sample's score on m principal components:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

$$F_i = a_{1i}X_1 + a_{2i}X_2 + ... + a_{pi}X_p \qquad \text{i = 1,2,...,m} \tag{6}$$

In practical applications, the dimensions of the indicators are often different, so the influence of the dimensions should be eliminated before the principal component calculation. There are many ways to eliminate the dimension of the data. The common method is to standardize the original data, that is, to perform the following data transformation:

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad i = 1,2,...,n;\ j = 1,2,...,p \tag{7}$$

And: $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$

4.3. Results Analysis

Descriptive statistics of the processed data and principal component analysis were performed to obtain descriptive statistics as shown in Tables 2, 4, and 6. Correlation matrices are shown in Table 3, Table 5, and Table 7 respectively.

Table 2. Hair_dryer descriptive statistics

	Average	Standard deviation	Analysis N
star_rating	4.12	1.301	11458
helpful_votes	2.18	14.249	11458
total_votes	2.57	15.390	11458
vine	.02	.124	11458
verified_purchase	.86	.352	11458
review_headline_len	22.28	16.542	11458

Table 3. Hair_dryer correlation matrix

		star_rating	total_votes	vine	verified_purchase	review_headline_len
correlation	helpful_votes	0.995	.54	.006	.091	.893
Saliency	helpful_votes	.000	.000	.000	.000	.000

Table 4. Microwave descriptive statistics

	Average	standard deviation	Analysis N
star_rating	3.44	1.645	1615
helpful_votes	5.62	27.772	1615
total_votes	6.67	29.263	1615
Vine	.01	.108	1615
verified_purchase	.68	.467	1615
review_headline_len	27.17	19.280	1615

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

Table 5. Microwave correlation matrix

		star_rating	total_votes	vine	verified_purchase	review_headline_len
correlation	helpful_votes	.997	.51	.239	.108	.700
Saliency	helpful_votes	.000	.000	.000	.000	.002

Table 6. Pacifier descriptive statistics

	Average	standard deviation	Analysis N
star_rating	4.12	1.301	11458
helpful_votes	2.18	14.249	11458
total_votes	2.57	15.390	11458
vine	.02	.124	11458
verified_purchase	.86	.352	11458
review_headline_len	22.28	16.542	11458

Table 7. Pacifier correlation matrix

		star_rating	total_votes	vine	verified_purchase	review_headline_len
correlation	helpful_votes	.995	.406	.006	.091	.853
Saliency	helpful_votes	.000	.000	.245	.000	.000

In summary: According to the table, helpful reviews of products are most relevant to star ratings, followed by review headline length, and moderately relevant are total votes, vine and verified purchase are almost irrelevant.

5. ANALYSIS AND MODELING OF REQUIREMENT 2

5.1. Analysis and Modeling of Task 1

5.1.1 Analysis and modeling of task 1 - Entropy decentralization

For task 1, we normalize the data and use entropy deweighting to weigh each factor, and then get the weights of the star ratings, total votes, and the comment length is 0.5064, 0.3025, 0.1911, respectively. According to the weights, the comprehensive coefficients of hair dryer, microwave, and pacifier are 2.8394, 3, and 2.815, respectively. The main product to be tracked is microwave.

The basic steps are:

Information entropy can be used to reflect the degree of variation of the index and can be comprehensively evaluated. The main steps are as follows:

(1) With n indicators, the data for m years will be converted into p_{ij} in the form of specific gravity, then:

$$p_{ij} = \frac{X_{ij}}{\sum_{i=1}^{m} X_{ij}}, (i = 1,2,\cdots, m; j = 1,2,\cdots, n) \tag{8}$$

(2) Define the entropy value of the e_j evaluation index, then:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

$$e_j = -k\sum_{i=1}^{m} p_{ij}\ln(p_{ij}), i = 1,2,\cdots,m; j = 1,2,\cdots,n$$,

and: $k = \frac{1}{\ln m}$,

The multiplication by a constant in the above formula is to ensure that when the specific gravity of the p_{ij} evaluation index is equal (I / m), it can be $e_j = 1$. At this time, this index cannot provide any information and cannot be used for the comprehensive evaluation.

(3) Under the condition $\sum_{j=1}^{n} w_j = 1$, the right w_j to define the evaluation index is:

$$w_j = (1-e_j)(n-\sum_{j=1}^{n} e_j) \quad (9)$$

Table 8. Normalization processing

star_rating	total_votes	review_headline_len
1	0	0.166667
0.8	0	0.893939
1	1	0.151515
1	0	0.151515
0.8	0	0.651515
1	0	0.151515
0.2	0	0.742424
0.6	1	0.393939
1	0	0.151515
0.2	0	0.924242
1	0	0.424242
0.8	0	0.151515
0.6	1	0.075758
0.8	0	0.151515
1	0	0.227273
0.2	0	0.439394
0.8	0	1
1	0	0.393939
1	0	0.151515
1	0	0.136364
1	0	0.151515
1	0	0.151515
1	0	0.151515
1	0	0.151515
1	0	0.151515
0.6	0	0.181818
1	0	0.378788
1	0	0.909091

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

The characteristic of the weighting method is that in the selected sample, the greater the value difference between the same indicators, the greater the weight.

Combination weighting method of modified G1 method based on entropy value:

Aiming at the above problems, this paper uses the method of combining weights by modifying the G1 method through entropy.

(1) The indicators are ordered by experts to determine the essentials of the indicators.

(2) By calculating the entropy value of each index data, the information changed is reflected by the entropy value of the index.

(3) According to the entropy value of each indicator, to determine the index importance scale is required by the G1 weighting method, and to determine the combination weight of the indicator through the idea of modifying the entropy value to 0. [3]

5.1.2 Solution of task 1

From model one, it can be seen that the helpful reviews are the most related to star_rating, followed by review_headline_len, and the medium ones are total_votes and vine, and verified_purchase is almost unrelated. This model mainly selects the star_ratings, review_headline_len and total_votes for weight analysis, and each of them uses 10 groups of data for normalization processing and then decentralization processing.

Using Matlab get the weight coefficients are: 0.5064, 0.3025, 0.1911.

Get the comprehensive evaluation formula:

Y=0.5064X1+0.3025X2+0.1911X3

Table 9. Evaluation Summary

	hair_dryer	microwave	pacifier
star_rating	4.115455	3.444582	4.304201
Normalized data	0.823091	0.688916	0.86084
total_votes	2.565407	6.66935	1.131466
Normalized data	0.256541	0.666935	0.113147
review_headline_len	22.27605	27.16842	22.44425
Normalized data	0.337516	0.411643	0.340064
General (1-5)	2.8394	3	2.815

In summary, the main tracking product is microwave.

5.2. Analysis and Modeling of Task 2

5.2.1 Analysis of task 2

It is needed to pre-process the data and perform a linear fit to get the relationship between the year and star rating of hairdryer, microwave and pacifier.

5.2.2 Building a model of task 2

(1) We firstly assume that the relationship between input variables and output variables is a linear function relationship, and then establish a multiple linear regression model:

$$Y' = \beta_0 + \beta_1 x_1' + \ldots + \beta_m x_m' + \varepsilon$$

$$\begin{cases} \varepsilon \sim N(0, \sigma^2) \end{cases}$$

(2) In order to study the correlation between two specified variables while controlling other variables that may affect it. We use partial correlation analysis to determine the interaction

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

between any two input variables. Determine whether there are interactions between the variables. Assume that there is a correlation between the random variables X, Y, and Z. In order to figure out the relationship between X and Y, we must calculate the partial correlation coefficient with Y on the assumption that Z is constant. Record as: $r_{xy.z}$. When for multiple variables, the partial correlation of order p-1 between X_i (i=1,2...,p)can be defined recursively as follows:

$$r_{0i.12\ldots(i-1)(i+1)\ldots p} = \frac{r_{0i.12\ldots(i-1)(i+1)\ldots(p-1)} r_{0ip.12\ldots(p-1)} r_{0ip.12\ldots(i-1)(i+1)\ldots(p-1)}}{\sqrt{1-r^2_{0p.12\ldots(p-1)}}\sqrt{1-r^2_{ip.12\ldots(i-1)(i+1)\ldots(p-1)}}}$$

A correlation test is calculated for the output variables.

We build a partial multivariate nonlinear regression model to judge the forms of interaction in the model of Y and X_i:

$$y = b_0 + b_1 x_1 + b_2 x_3 + b_{11} x_1^2 + b_{22} x_2^2 + b_{33} x_3^2 + b_{12} x_1 x_2 + b_{13} x_2 x_3 + b_{23} x_2 x_3 + \varepsilon$$

And: $\varepsilon \sim N(0, \sigma^2)$

After that we build all multivariate nonlinear regression models with Y.

(3) After inputting the data, we use SPSS 22 software to get the unknown coefficients, so as to get the functional relationship between them. Then, perform parameter estimation, statistical analysis, hypothesis test, regression coefficient test, and correlation coefficient test. If the test passes, a better model will be obtained. If the test fails, further adjustment and optimization will be performed.

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \sum_{j=1}^{m} \beta_j \times x_j^2 + \sum_{1 \le j \ne k \le m} \beta_{jk} x_j x_k$$

(4) After we get the functional relationship, we need to estimate its parameters. Suppose there are n independent observations $(x_{i1}, x_{i2}, \ldots, x_{im}, y_i), i = 1,2,\ldots,n,$ and then determine the regression coefficient $\beta_0, \beta_0, \ldots, \beta_m$

By least squares:

$$\min Q(\beta_0, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im})]^2$$

Find an estimate:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1m} \\ \ldots & & & \ldots \\ 1 & x_{n1} & \ldots & x_{nm} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \ldots \\ y_n \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \ldots \\ \hat{\beta}_m \end{bmatrix}$$

The estimated value of Y is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_m x_m$$

Fitting error, called the sum of squared residuals:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(5) Statistical analysis:

Firstly, find the sum of squared residuals Q and the unbiased estimate obtained from σ^2.

$$\frac{Q}{\sigma^2} \sim \chi^2(n-m-1); \sigma^2 = \frac{Q}{n-m-1}$$

Then, the sample variance of Y is decomposed use S^2.

$$S^2 = Q + U, U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

(6) Hypothesis test:

Construct F-statistics and rejection fields H_0 for the test:

$$H_0 : \beta_0 = \beta_1 = ... = \beta_m = 0$$

$$F = \frac{U/m}{Q/(n-m-1)} \sim F(m, n-m-1)$$

Rejection fields: $\chi_0 = \{F > F_{1-\alpha}(m, n-m-1)\}$

(7) Test of regression coefficient:

Determine whether each independent variable x_i has a significant effect on y.

$$T_i = \frac{\hat{\beta}_i / \sqrt{c_{ii}}}{\sqrt{\frac{Q}{n-m-1}}} \sim t(n-m-1)$$

And: $H_0 : \beta_i = 0, H_1 : \beta_i \neq 0, \hat{\beta}_i \sim N(\beta_i, c_{ii}\sigma^2), i = 1,..., m$

(8) Correlation coefficient test:

$$R^2 = \frac{U}{S^2}$$

The complex correlation coefficient R is an index to measure the correlation between y and $x_1, x_2, ..., x_m$. The closer the value of R is to 1, the closer their correlation. [4]

5.2.3 Solution of task 2

It is known from the model that the most important weight is star reviews. Therefore, the star reviews of each year in the data are counted, few data are deleted, and the three years with the average of the years as a representative value, after finishing, the comprehensive data in the following table is obtained.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

Table 10. Time and star review statistics table

Year	Hair_dryer	Microwave	Pacifier
2006	3.19	4.21	3092
2007	3.97	4.42	3.84
2008	3.76	3.48	4.17
2009	3.97	2.75	4.14
2010	3.84	3.03	4.22
2011	3.86	2.58	4.11
2012	3.98	2.93	4.24
2013	4.15	3.21	4.31
2014	4.18	3.51	4.34
2015	4.23	3.77	4.35

Take the average number of stars for each year according to the obtained data, draw a scatter plot by Matlab, and then perform a linear fit to obtain a linear fit graph and relationship.

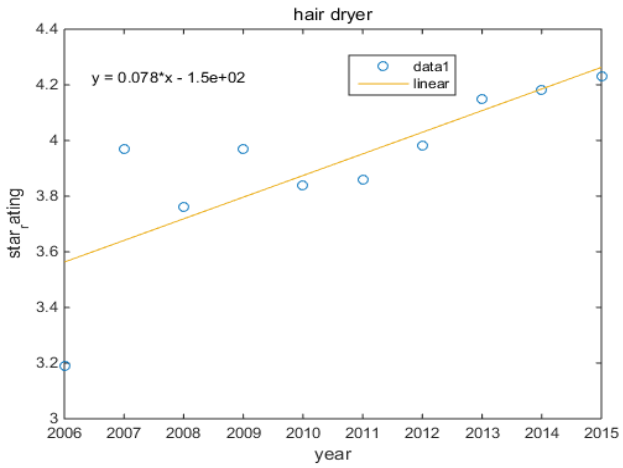

Figure 2.

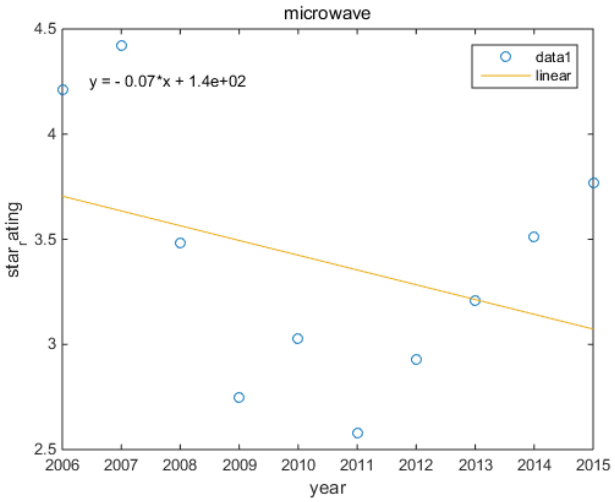

Figure 3.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

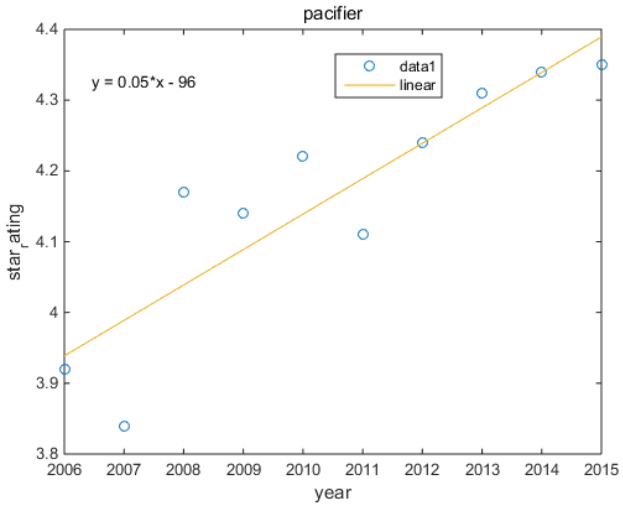

Figure 4.

Polynomial fitting, fit m polynomials with data x_0, y_0,

If taken: $\{r_1(x), \bullet\bullet\bullet, r_{m+1}(x)\} = \{1, x, \bullet\bullet\bullet, r_{m+1}(x)\}$,

$a = polyfit(x_0, y_0, m)$

Where the input parameter x_0, y_0 is the data to be fitted, m is the degree of the fitted polynomial, and the output a is the fitted polynomial:

$y = a_m x^m + \ldots + a_1 x + a_0$ where $a = [a_m, \ldots, a_1, a_0]$

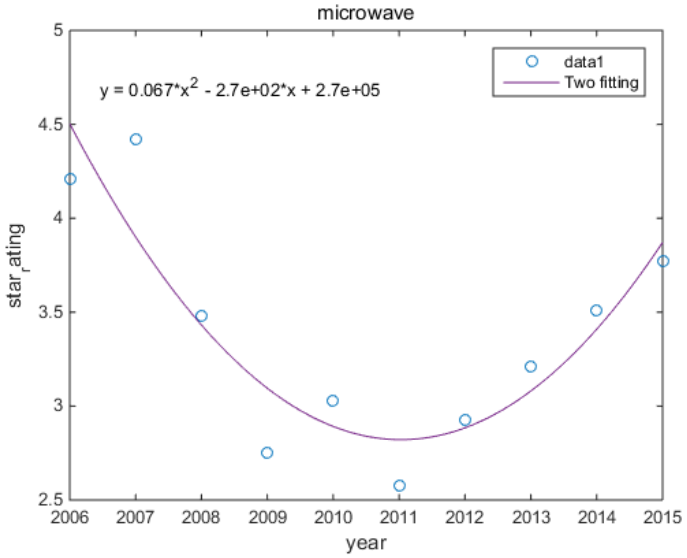

Figure 5.

In conclusion, over time, the sales star of hairdryer and pacifier tend to increase. Microwave tends to decrease. But in the next few years, there has been a significant increase, and the decreasing trend is weakening.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

5.3. Analysis and Modeling of Task 3

5.3.1 Analysis of task 3

Integrate the data, and use the length of the reviews, the star ratings of the reviews as independent variables, use helpful reviews as dependent variables to do the correlation analysis.

5.3.2 Building a model of task 3

To study the correlation between two sets of random variables, a complex correlation coefficient (also called a full correlation coefficient) is available. In 1936, Hotelling extended the simple correlation coefficient to the discussion of the correlation between multiple random variables and multiple random variables and proposed a typical correlation analysis. [5]

For the correlation, we can use the most primitive method to calculate all the correlation coefficients between the two groups of variables. There are total of pq simple correlation coefficients. It would be simpler if we could use a similar idea to the principal component to find a linear combination of each of the two groups of variables and discuss the correlation between the linear combinations. Firstly, find the first pair of linear combinations in each group of variables to make them have the greatest correlation.

$$\begin{cases} u_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ v_1 = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}y_q \end{cases}$$

Then, find a second pair of linear combinations in each group of variables, so that they are irrelevant to the first linear combination in this group, and the second pair has the second largest correlation.

$$\begin{cases} u_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ v_2 = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}y_q \end{cases}$$

5.3.3 Solution of the task 3

Correlation analysis was performed using review length, review star ratings as independent variables, and helpful reviews as dependent variables. Get the following correlation analysis table:

Table 11. Hair_dryer

		star_rating	helpful_votes	review_headline_len
helpful_votes	Pearson	.970	1	.857
	Distinctiveness	.000		.000
	N	11459	11459	11458

Table 12. Microwave

		star_rating	helpful_votes	review_headline_len
helpful_votes	Pearson	.931	1	.770
	Distinctiveness	.000		.005
	N	1615	1615	1615

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

Table 13. Pacifier

		star_rating	helpful_votes	review_headline_len
helpful_votes	Pearson	.870	1	.613
	Distinctiveness	.000		.000
	N	18925	18925	18924

In summary, it is found that the comments in the correlation analysis table below are more relevant to star rating comments.

5.4. Analysis of Task 4

To find out whether a particular star rating will lead to more reviews, for example, whether a customer is more likely to write some type of reviews after seeing a series of low star ratings. We have analyzed the length of the reviews of three types of products with different star ratings. [6]

Table 14. Three types of products with different star ratings.

Star rating	Hair_dryer Comment length	Microwave Comment length	Pacifier Comment length
1	25.72	31.19	26.02
2	25.1	31.65	26.85
3	23.43	30.54	27.9
4	22.96	28.25	24.97
5	20.77	22.82	19.29

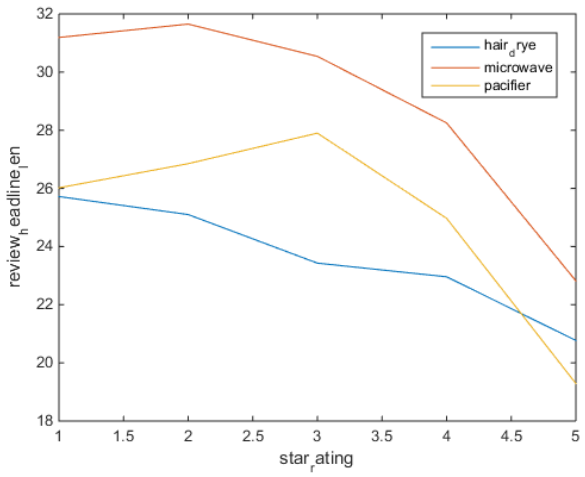

Figure 6.

In summary, according to the figure, as the star rating increases, the length of the reviews decreases. The number and length of extreme reviews are more, meanwhile 2, 3, and 4 stars are less.

5.5. Analysis of Task 5

Through the processing of the data, we can see the star heat map of the positive and negative words in the comments of hair dryer, microwave, and pacifier.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

Table 15. Pacifier 'comments' star rating distribution heat map with negative words

Words \ Star Rating	5 stars	4 stars	3 stars	2 stars	1 star
worst	11.4%	17.1%	0.0%	11.4%	60.0%
poor	21.8%	5.0%	11.8%	21.8%	39.5%
awful	20.0%	5.0%	12.5%	25.0%	37.5%
disappointed	17.0%	9.2%	14.5%	24.0%	35.3%
hate	42.8%	13.2%	11.9%	12.6%	19.5%

Table 16. Pacifier 'comments' star rating distribution heat map with positive words

Words \ Star Rating	5 stars	4 stars	3 stars	2 stars	1 star
excellent	87.80%	8.30%	1.70%	0.70%	1.70%
wonderful	84.40%	9.80%	2.70%	1.60%	1.60%
love	82.50%	11.10%	3.40%	1.50%	1.60%
best	82.20%	10.00%	3.60%	2.40%	1.90%
perfect	81.70%	12.30%	3.40%	1.40%	1.20%
great	74.50%	15.50%	5.00%	2.90%	2.10%
nice	61.80%	23.30%	8.10%	4.00%	2.70%

Table 17. Hair Dryer 'comments' star rating distribution heat map with negative words

Words \ Star Rating	5 stars	4 stars	3 stars	2 stars	1 star
worst	10.4%	4.2%	4.2%	18.8%	62.5%
awful	9.5%	9.5%	14.3%	14.3%	52.4%
poor	13.6%	10.6%	13.6%	25.8%	36.4%
disappointed	21.4%	7.0%	19.3%	19.8%	32.5%
hate	51.8%	12.2%	12.2%	7.2%	16.5%

Table 18. Hair Dryer 'comments' star rating distribution heat map with positive words

Words \ Star Rating	5 stars	4 stars	3 stars	2 stars	1 star
excellent	83.0%	13.1%	1.5%	0.6%	1.8%
love	80.3%	12.1%	3.2%	1.5%	2.8%
wonderful	80.3%	11.2%	2.6%	5.3%	0.7%
best	79.8%	10.9%	4.4%	3.2%	1.7%
perfect	77.0%	15.6%	2.9%	2.3%	2.2%
great	70.3%	19.1%	4.8%	2.9%	3.0%
nice	62.4%	22.3%	9.1%	3.4%	2.8%

Table 19. Microwave 'comments' star rating distribution heat map with negative words

Words \ Star Rating	5 stars	4 stars	3 stars	2 stars	1 star
worst	0.0%	0.0%	0.0%	9.1%	90.9%
awful	0.0%	0.0%	22.2%	0.0%	77.8%
poor	3.0%	9.1%	3.0%	9.1%	75.8%
disappointed	16.7%	12.5%	4.2%	12.5%	54.2%
hate	20.8%	20.8%	29.2%	12.5%	16.7%

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

Table 20. Microwave 'comments' star rating distribution heat map with positive words

Words \ Star Rating	5 stars	4 stars	3 stars	2 stars	1 star
excellent	73.80%	9.50%	4.80%	0.00%	11.90%
love	66.90%	14.50%	4.10%	4.70%	9.90%
perfect	66.80%	19.60%	5.10%	4.20%	4.20%
great	62.50%	18.80%	5.70%	4.30%	8.70%
best	61.40%	17.50%	5.30%	1.80%	14.00%
wonderful	61.10%	11.10%	0.00%	11.10%	16.70%
nice	42.90%	28.60%	12.60%	5.10%	10.90%

In summary, review with specific quality description is indeed strongly associated with rating levels. We can see that reviews with negative words are indeed more likely to be star rating 1-2 negative reviews, while reviews with positive words are more 5 and 4 positive reviews. According to the distribution map of these three products containing negative words, it can be seen that star rating 1 is the comment containing the word 'worst'. Based on it the three products containing positive words, star rating 5 are containing the word 'excellent' most. So we recommend regularly reviewing these reviews with specific quality descriptors. The company need to continue to maintain the advantages of the product given in the comments that contain positive words, and to improve the disadvantages of the product immediately given in reviews containing negative words. [7]

6. STRENGTHS AND WEAKNESSES OF THE MODEL

Strengths of the model:

(1) The model in this paper has been introduced and studied in details. The algorithm is flexible and easy to understand.

(2) Correlation analysis of each factor was determined using SPSS software.

(3) Determining and analyzing most factors, the model is reliable.

(4) The results obtained are more in line with reality and have higher accuracy.

Weaknesses of the model:

(1) Ignoring the influence of factors such as gender, removing some factors, will have a small impact on the results.

(2) The model is not detailed enough. And the model research is not deep enough.

ACKNOWLEDGEMENTS

These authors are contributed equally to this work.

REFERENCES

[1] How Amazon uses Big Data in practice, https://www.bernardmarr.com/default.asp?contentID=712

[2] Kim S M, Pantel P. Automatically Assessing Review Helpfulness[C]. Proceedings of The 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006),2006:423-430.

[3] Wang Hezhu, Zhu Yong, Zhu Yi, He Youxun, Qin Likang, Liang Yali.Quality Evaluation of Kidney Beans in Guizhou Based on Principal Component Analysis [J / OL] .Food and Machinery: 1-7 [2020-03-10]

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202012_6(12).0008

[4] Mei Xueyi, Zhou Meihua.Research on Drug Sales Forecast in Retail Pharmacy under Limited Time Data [J / OL]. Journal of China University of Mining & Technology (Social Science Edition): 1-12 [2020-03-10].

[5] Li Xiaozhang, Zhang Yanqiong, Qiao Jun, Qin Chenghua, Analysis of Sources of Shallow Groundwater Pollution Based on Multivariate Statistical Methods [J] .China Environmental Monitoring, 2020,36 (01): 88-95.

[6] Wang Junkui. Research on the Usefulness of Online Reviews of E-commerce Websites [D]. Xidian University, 2014.

[7] Lin Jie, Wang Pingchun.Research on the Influential Factors of the Usefulness of Online Reviews in E-commerce Trade [J]. Commercial Economic Research, 2017 (10): 73-75.