Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2010638

Sales Strategy Recommendation Based on Commodity Data Mining

Summary

We conduct data analysis and information mining from the following four aspects: correlation between review, star rating and helpfulness rating, product brand rating, prediction of product's reputation and impact of star rating on reviews, so as to propose reliable sales strategies and suggestions for product improvement.

Firstly, after data pre-processing, We perform word segmentation and preliminary sentiment analysis on the text data through the NLTK tool, and quantify it as emotion score, ranging from [-1,1].We adopt the methods of data visualization, descriptive statistics and correlation analysis, and further construct a multivariate Logistic regression model to analyze the relationship between helpfulness rating and review length, star rating and compound. The results show that helpfulness rating has an inverted "U-type" relationship with review length and a positive "U-type" relationship with star rating.

The next step, we conduct analysis based on the rating and evaluation model. The LDA analysis model is constructed to find the topic feature of each product, based on which we can propose suggestions for improvement of products. At the same time, we summarize five indexes that affect the product sales from the topic features, namely quality, price, appearance, service and size. Then a computer search algorithm based on the text similarity is used to calculate the index score of each review. In addition, we combine the score with the analytic hierarchy process to determine the weight of each index and build a weighted brand scoring system. Finally, we cluster all product brands through systematic clustering to select potential high-quality brands and recommend them to sunshine company.

Further, we calculate the comprehensive score of review and star rating, and take it as the product's reputation, which is conducive to forecast the future reputation of three products through time series analysis. And it depicts that the three products have seasonal characteristics and will probably maintain a stable seasonal cycle in the near future. However, the peak time of reputation comprehensive score of the three products is discrepant, and further analysis illustrates that the product sales figure is larger during the peak period of reputation score, according to which sunshine company can make a sales plan.

Finally, we analyze the relationship between star rating and review. Through the establishment of distributed lag model, it is found that customers' reviews in the current period will be affected by other customers' ratings and reviews. Meanwhile, we observe the time-varying synchrony between emotional score and star rating, which is clear that reviews containing positive words result in higher star ratings, while reviews containing negative words result in lower star ratings. In other words, there is a strong correlation between star ratings and specific quality descriptors.

Key words: correlation analysis, multinomial logistic regression, natural language processing, time series analysis

Content

1. Introduction

1.1 Background

With the popularization and development of the Internet, online sales are gradually replacing offline sales and occupy a major position in the sales industry. The transformation of sales mode will also be a huge challenge for commodity companies. Analyzing and mining market information and consumer feedback has become the key of network marketing.

The online marketplace created by Amazon provides an opportunity for customers to evaluate their purchases. Customers can use quantities from 1 (low rating, low satisfaction) to 5 (high rating, high satisfaction) to express their satisfaction with the product. In addition, customers can submit their own reviews on the products to express their opinions. Other customers can use these reviews to gain an initial understanding of the product before purchasing it, and to distinguish between the reviews (called "helpfulness ratings") that are helpful or not. The data will be fed back to the company which can use the data to conduct in-depth analysis of the markets, determine the timing of their participation, and modify products based on customer suggestions.

1.2 Problems Restatement

Sunshine Company plans to launch and sell microwave ovens, hair dryers and baby pacifiers in the online marketplace. In order to understand these three commodity markets and develop sales strategies, it is necessary to analyze the customer feedback data. We will accomplish the following tasks according to the given data:

(1) Develop sales strategy for sunshine company.

(2) Identify potential important design features that would enhance product desirability.

In order to accomplish the above two tasks, our specific work is as follows:

- Analyze the relationship between star rating, helpful vote and review.
- Make in-depth analysis of reviews and ratings, and distinguish the advantages and disadvantages of products, so as to make suggestions for product improvement.
- Establish a product rating system based on the analysis of reviews and ratings, and select high-quality brand products to recommend to Sunshine Company for sales.
- Establish the reputation scoring system and predict the development trend of reputation.
- Analyze whether current reviews are influenced by prior star ratings and reviews.
- Analyze whether specific quality descriptors of text-based reviews strongly associated with rating levels.

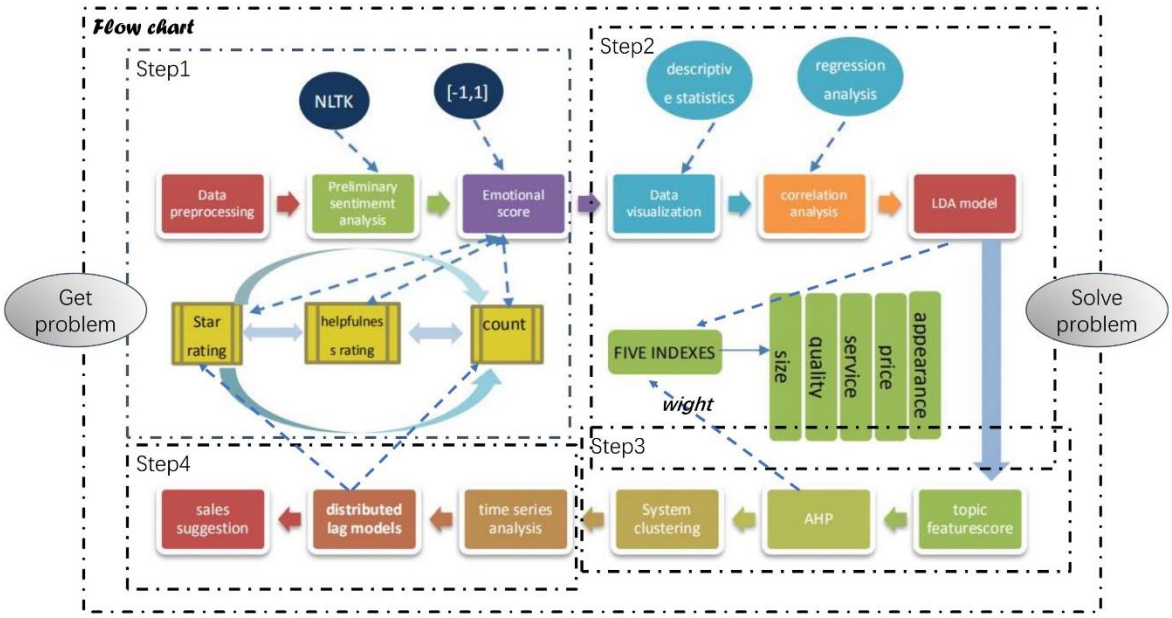


Figure 1: The Flow Chart in This Paper

2. Assumptions and Nomenclature

2.1 Assumptions

We make several assumptions in our model. Later we may relax some of these assumptions to optimize our model making it more applicable in the complex reality environment.

- All reviews are from customers and there is no automated review.
- There is no deliberate attempt to smear the product in customer rating and review.
- Unverified orders are not sold on Amazon. The sum of confirmed sales in the data set is total sales.
- The review of members certified by Amazon Vine has high credibility.

2.2 Nomenclature

Symbols	Definition
SS	Star ratings squared
count	The number of words in each review.
CS	The number of words squared
compound	Emotional score
H_r	The percentage of helpful votes in the total votes.
w_i	The weight of the i^{th} index
SQ	Total score of six indexes(quality、size、service、price、appearance、star ratings)
SR	Total score of product reputation

3. Model 1-Analysis of Star rating, Helpfulness rating and Review

3.1 Data Pre-processing

Before data analysis, the availability of data must be guaranteed. No measures, regardless of its value, can provide accurate assessments if based on unreliable data. We first remove useless information including product title, marketplace and product category, on the basis of which we can carry out data pre-processing.

To ameliorate the condition of data set, there are four steps: data classification, data cleaning, information filtration, set-up of new attribute and data-measuring, as shown in the figure.

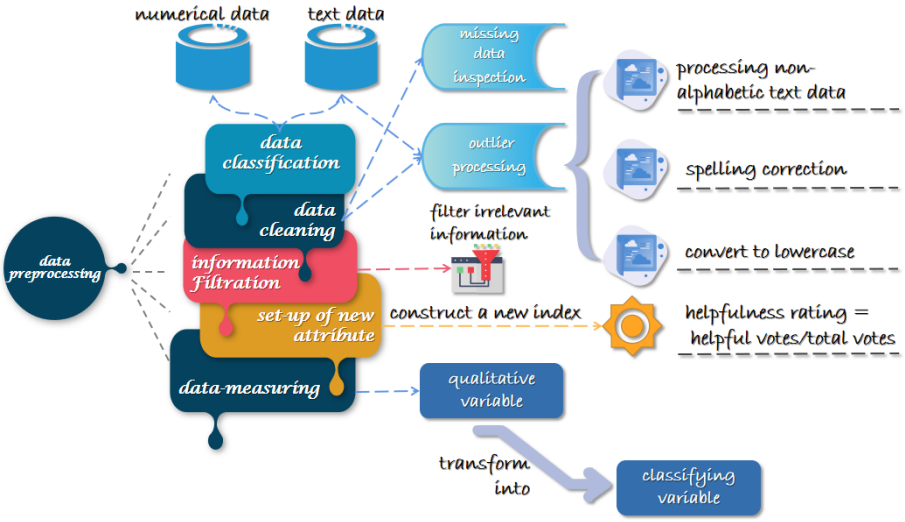


Figure 2: Data Pre-processing Flow Chart

Step 1: Firstly, divide the data into numeric and textual types and process them separately.

Step 2: In the stage of data cleaning, initially we check the missing value by using the "duplicated () method" in python, while no item with a missing value was found.

Step 3: Outlier processing is mainly used in the text data. Take the pacifier as an example, it includes the following contents:

- Processing non-alphabetic text data. There are some non-alphabetic data like: 😁 , 👍 , ❤, 💔 ,They can't be recognized by a computer program, while their emotional inclination can be intuitively identified by us. As a result, we transform all of them into "five stars", so as to facilitate the later computer programming.
- Spelling correction. Generally, English text may contain spelling errors, so a spelling check is required. Based on "PyEnchant library" in python program, all the review texts are checked and corrected.
- Lowercase: the computer will react differently to the two texts: "Five stars" and "five stars", whereas we expect them to be the same word when we count them. Therefore, we unify all text data to lowercase characters.

Step 4: After the above steps, our data set has become more acceptable, but when we look more closely at the review content, we find that a lot of the review content is irrelevant to the product and can be regarded as redundant information, such as:

- “good cleaning cloths”
- “I bought hits basket to store magazines and newspapers. It looks very nice in our room”
- “I received this blanket as a gift from my mom and have never felt anything so soft.”

Take the pacifier as an example, reviews that have nothing to do with the pacifier is needed to be weeded out. We identify the redundant information according to the text similarity, and the process is as follows:

The product review is regarded as a set of words, and the feature vector of the text is established by calculating the number of each word in the text. Afterwards, the cosine similarity between vectors is used to calculate the similarity between texts. If the average similarity between a review and others is less than the threshold value set by us, the review is removed.

Step 5: Set-up of new attribute. To better analyze the usefulness of reviews, we establish a new attribute named helpfulness rating (Hr), which is obtained by dividing “helpful votes” by “total votes”

3.2 Word Segmentation and Sentiment Analysis

Although the text data has been initially processed aforementioned, the information from reviews is s till needed to assess.However, the review data is unstructured and requires a different mechanism to ex tract the information. Emotion analysis or opinion mining is the process of computing to identify wh ether the author's attitude towards a text is positive, negative or neutral. For the review data, we con ducted a preliminary Sentiment analysis through data mining to determine and quantify the emotional inclination of each review, and further analysis will be discussed in the following questions. Based on this, NLTK (natural language toolkit),which is a leading platform that is able to work with human langu age data by python program, is used for sentiment analysis.

A basic sentiment analysis model is built in Python 3 using the NLTK library. The preprocessin g is operated by marking the text, normalizing the words, and removing the noise. Next, the NTLK emotional analyzer is used to build a model and associate the text with a specific emotion and quant ify it.

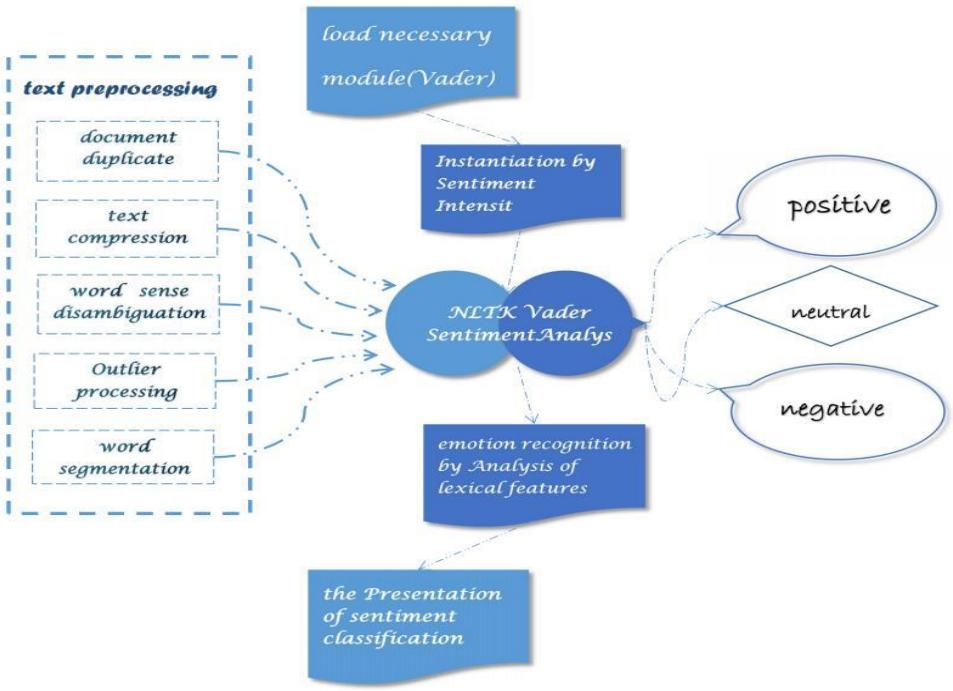


Figure 3: Implementation process

3.3 Data Description and Visual Analysis

The table below describes the total, average and variance of the star rating of these three products. The average of the helpful votes and the number of the words in the review are also calculated. The figures below describe their distribution.

Table 1: Data Statistic

Descriptive statistics	Total	Star rating average	Star rating variance	Helpful votes average	Count	Count average
Pacifier	18939	4.30456	1.19043	0.827182	4841870	256
Microwave oven	1615	3.44458	1.64524	5.62167	748130	463
Hair dryer	11470	4.11604	1.30033	2.17908	3268716	285

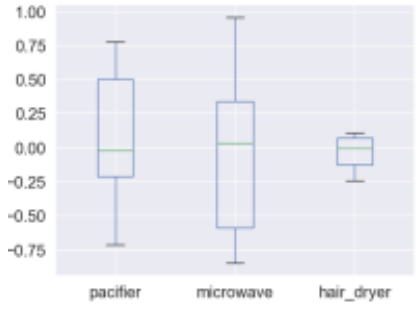


Figure 4: Box Figure

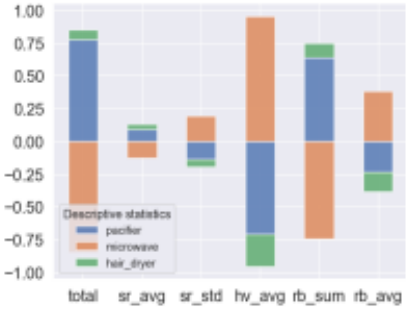


Figure 5: Stacked Barplot

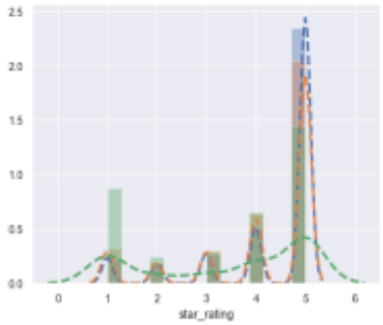


Figure 6:Star Rating Distributon

The data is visualized to dig into the inherent rules, which is helpful for modeling.

The figures below describe the relationship between the helpfulness rating, and star ratings, review length and emotional score.

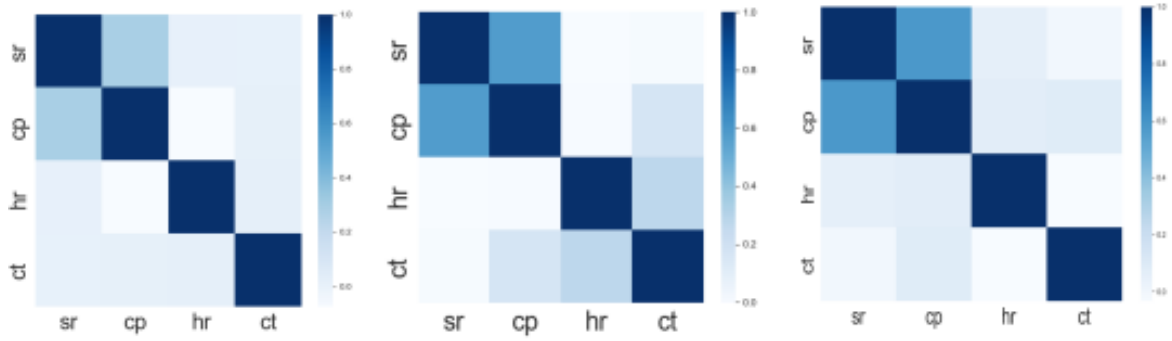


Figure 7:Heat Map of Correlation

Note: sr represents "star rating"; cp represents "compound"; ct represents "count"; hr represents "helpfulness rating"

The darker the color in the figure is, the greater the correlation between the two indicators will be. We can find that the emotional score and star ratings has a strong correlation in the three figures. The helpfulness rating and the review length has a light correlation.

3.4 Correlation Analysis

We apply the methods of correlation analysis on the percentage of helpful votes (H_r), star ratings, the number of words in each review (count) and emotional score (compound)., which is measured by Pearson correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n}(\mathrm{X}_i - \bar{\mathrm{X}})(\mathrm{Y}_i - \bar{\mathrm{Y}})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (1)$$

X_i represents the star rating, review length, emotional score, etc. Y_i represents the helpfulness rating,. $\overline{X}$, $\overline{Y}$ represents the average number of these indexes. The r is the correlation coefficient between H_r and star rating, count, compound, etc.

The results of the correlation analysis are

Table 2: Correlation Coefficient

Index	Correlation coefficient(Microwave oven)	Correlation coefficient(Hair dryer)	Correlation coefficient(Pacifier)
star rating	-0.149**	-0.138**	-0.151**
count	-0.071**	0.356**	0.332**
compound	-0.024**	0.057**	-0.101**

*Note: ** means the result is significant at 0.01 level (two-side test)*

From the table above we can see that, for the three data sets, there is a significant correlation between H_r and star rating, count and compound (all of which are significant in the two-tailed test), which is consistent with the results of visual analysis. More concretely, the star rating and emotional score of microwave oven, hair dryer and pacifier are negatively correlated with H_r. The counts of hair dryer and pacifier are positively correlated with H_r, while the counts of microwave oven are negatively correlated with H_r. Besides, there is a negative correlation between the compound of microwave oven, pacifier and H_r, while the correlation between the compound of hair dryer and H_r is positive.

3.5 Multinomial Logistic Regression Model

Multinomial logistic regression model is adopted to further analyze the impact of star rating, review length and emotional score on the helpfulness rating. In fact, the helpfulness rating is the cumulative result of each consumer's voting (yes or no) . Therefore, the number of helpful votes is subject to binomial distribution, and logistic model is suitable for empirical analysis of such kind of data.

Multivariate logistic regression can determine the role and intensity of the explanatory variable X_n in predicting the probability of occurrence of strain Y. Suppose X is the response variable and P is the response probability of the model, and the corresponding regression model is as follows:

$$\ln\left(\frac{p_1}{1-p_1}\right) = \alpha + \sum_{k=1}^{k}\beta_k x_{ki} \quad (2)$$

$p_1 = P(\mathrm{y}_i = 1 | x_{1i}, \mathrm{x}_{2i}, \cdots, \mathrm{x}_{ki})$ is the possibility that case i happens in case the values of $x_{1i}, \mathrm{x}_{2i}, \cdots, \mathrm{x}_{ki}$ are given. Also p_1 is a binomial distribution parameter, which is the probability that a review receives a helpful vote. x_{ki} represents a series of independent variables. β is the corresponding estimation coefficient, and α is the intercept, which reflects the random effect at the product level.

The probability of an event happening of an event is a non-linear function composed of the explanatory variable X_i. Here is the expression:

$$p = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)} \quad (3)$$

On top of these 3 indexes, we introduce two additional variables: star rating square(SQ) and review length square(CS) in order to further explore the influence of star rating and review length on the helpfulness ratings of reviews.

Meanwhile, considering the impact of whether the customers are vine members and whether the purchase is verified on the authenticity and validity of reviews, we take these two indicators as independent variables.

Independent variables are divided into three levels, which represent the intuitive evaluation (star rating, star rating square) respectively, characteristics of review content (review length, review length square, emotional score), and *reviewers'* characteristics (vine, VP), to explore the various reasons that affect the helpfulness ratings in detail.

As for dependent variables, we divide them into four categories according to their numerical values, namely:

$$helpfulness\ ratings = \begin{cases} 1, 0.75 < compound \le 1 \\ 2, 0.50 < compound \le 0.75 \\ 3, 0.25 < compound \le 0.50 \\ 4, 0 \le compound \le 0.25 \end{cases} \quad (4)$$

The independent variables are divided into blocks and do regression analysis respectively. The independent variables in models 1, 2 and 3 and the regression coefficients obtained by SPSS are summarized as Table.For reasons of space, the results of hair dryer and pacifier are included in the appendix.

	independent variables	**Abbreviations**	**Model 1**	**Model 2**	**Model 3**
Intuitive evaluation	star rating	star rating	0.213**	0.469**	0.423**
	star rating square	SS	0.314**	0.207**	0.110**
Characteristics of review content	review length	count		-0.020**	-0.019**
	review length square	CS		0.002**	0.002**
	emotional score	compound		-0.146**	-0.160**
Reviewers' characteristics	vine	vine			0.033**
	verified purchase	verified purchase			0.398**

Table 3: Index and regression coefficient

Note:: p<0.05; **: p<0.01; +: p<0.1; ns: not significant*

It is clear from the above table, each variable passed the significance test. The regression coefficients of star ratings of microwave oven, hair dryer and pacifier are all positive, and the regression coefficients of star rating square terms are also positive. It indicates that there is a "U-type" relationship between star ratings and

the helpfulness ratings of reviews, which is contrary to the research conclusions of Mudambi and Schuff [8]. Our analysis suggests that the reviewers who score higher or lower star ratings may be more willing to express a clear attitude towards the products, so as to provide more valuable reviews; while the reviewers who give intermediate star ratings may lack reference value due to their less distinctive attitude.

Based on model 1, review length, review length square and emotion score are added into the model. From the data in the table, it is clear that the regression coefficient of the review length term is negative, while one of the square term of the review length is positive, which indicates that the review length and helpfulness rating are of an inverted-U-type. According to our analysis, reviews that are too short are often limited in content and cannot provide sufficient useful information. However, overlong reviews may provide a lot of valuable information though, other customers may not be patient to spend time and energy reading them. Therefore, reviews with high helpfulness ratings should be of moderate length.

On the basis of model 2, in model 3 we introduce other reviewer-related features such as Vine and Verified Purchase. It can be seen that most customers are not vine members and many purchase are not verified in the sample data, thus it is difficult to analyze and draw valuable conclusions about these two indicators. In spite of this, they make the model more complete and more applicable.

4. Model 2-Establish a Scoring System to Determine Product Positioning

According to the three data sets, each product has hundreds of different brands. If sunshine company hopes to enter the market of microwave oven, hair dryer and pacifier, it is of great significance to accurately grasp the market positioning of each brand. The market positioning of different brands is mainly determined by the star rating and review from customers. The processing of star rating is relatively simple, so we mainly process text of customers' reviews in detail.

4.1 LDA Model

Latent Dirichlet Allocation is a generative thematic model proposed by Blei et al [2] in 2003. It is also known as three-tier Bayesian probability model with three-tier structure of document (d), topic (z) and word (W), which can effectively model the text. Based on LDA topic model, we are able to mine the potential topics in the data set, and then analyze the main information of the data sets and related feature words.

Table 4:Symbol Explanation

α , β	prior parameters of the Dirichlet function
θ , ϕ	the parameter of the topic's multiple distribution in the document.

LDA model assumes that each review is randomly mixed by each subject according to a certain proportion. And the mixing proportion is subject to multiple distribution, which is recorded as follows:

$$Z \mid \theta = Multionmial(\theta) \quad (5)$$

Each topic is made up of the words in the glossary in a certain proportion, and the proportion of the mixture is also subject to multiple distribution.

$$W \mid Z, \phi = Multionmial(\phi) \tag{6}$$

Under the condition of review d_j, the probability of generating word w_i is expressed as:

$$P(w_j \mid d_j) = \sum_{i=1}^{K} P(w_i \mid z = s) \times P(z = s \mid d_j) \tag{7}$$

$P(w_i \mid z = s)$ is the probability that the word w_i belongs to the s^{th} topic. $P(z = s \mid d_j)$ indicates the probability that the s^{th} topic is in the review.

After sentiment analysis based on LDA , the review text is gathered into three topics, as a result 10 most frequently used words and their corresponding probability are generated under each topic. Table 5 shows the potential topics in the positive evaluation text of microwave ovens, and the other one depicts the negative evaluation. Considering the limited space of the article, we put the results of hair dryer and pacifier in the appendix.

Table 5:Microwave Oven Positive Topics

Theme 1	Theme 2	Theme 3	Theme 1	Theme 2	Theme 3
great	good	well	product	product	small
well	great	great	space	no	nice
like	well	good	price	price	easy
good	use	use	use	nice	product
no	small	size	like	like	like

Table 6:Microwave Oven Negative Topics

Theme 1	Theme 2	Theme 3	Theme 1	Theme 2	Theme 3
new	no	like	time	time	no
use	use	new	more	service	old
easy	like	use	product	great	more
no	more	service	model	small	small
like	new	time	service	good	well

According to the extraction of three potential subject words for positive and negative evaluation, we can determine the customer's attitudes towards 5 aspects of microwave oven characteristics.

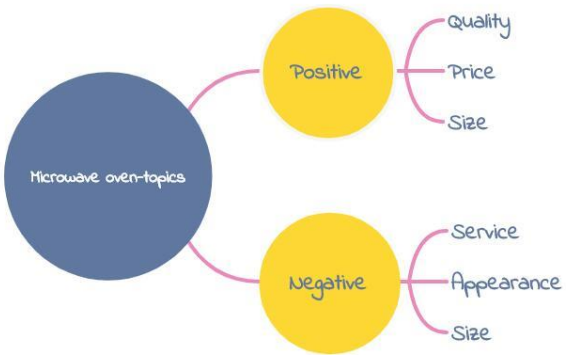


Figure 8:Index of Microwave Oven

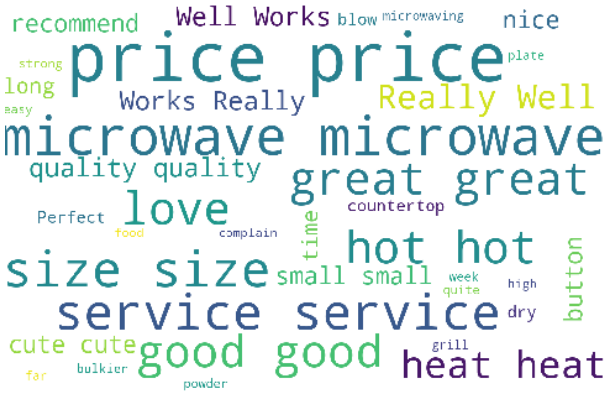
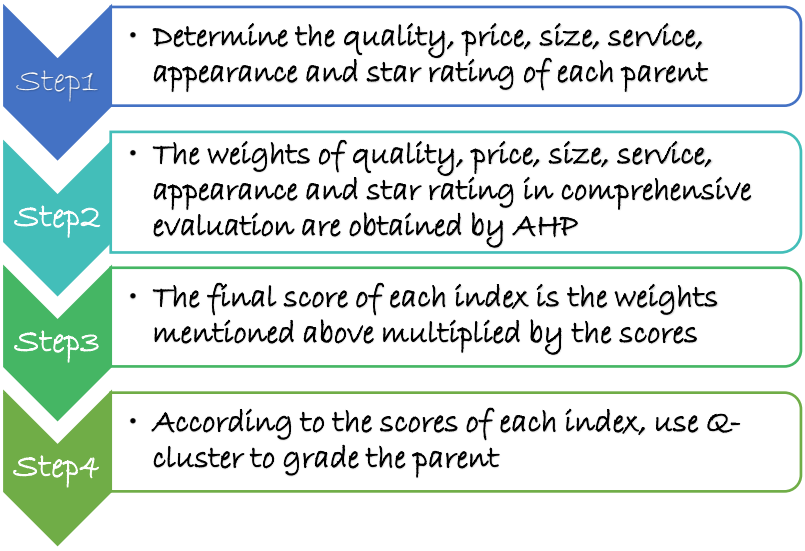


Figure 9:Word Cloud

These high-frequency words in reviews can also reflect the focus of customers when purchasing microwave ovens, that is the quality, price, size, service and appearance. These five indicators can reflect a product comprehensively. We conduct the same analysis on pacifiers and hair dryers, then it is found that these five indicators can also be taken as the criteria to consider whether they are worth purchasing.

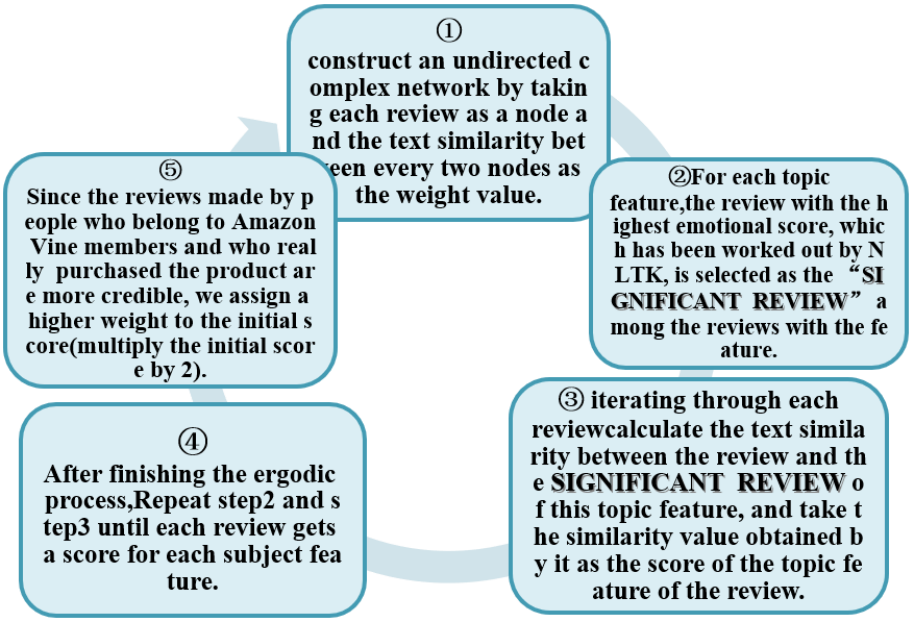
Taking the quality, price, size, service, appearance and star rating as the evaluation indexes, we establish a scoring system to comprehensively evaluate these 3 product sold in the current market, so as to obtain the accurate market positioning of each brand of products, and then formulate a good sales strategy for sunshine company.

The steps of our brand scoring system are as below:



4.2 Determine index score

After implementing LDA algorithm, complex network is introduced to calculate the importance degree of each review. We use traversing search algorithm to implement the process, through which we can get weighted average score of each topic feature for all products. Here is the algorithm:



4.3 Analytic hierarchy process (AHP)

4.3.1 The introduction of analytic hierarchy process

Analytic hierarchy process (AHP) can be used deeply analyze the essence, influencing factors and their internal relations of complex decision problems. It makes use of less quantitative information to make the thinking process of decision mathematical, so as to provide a simple method for complex decision problems. AHP can decompose the evaluation criteria into different hierarchical structures, and then calculate the weight of each factor by solving the eigenvectors of judgment matrix.

According to the LDA model, the indicators we extracted are quality, appearance, price, size and service satisfaction level. Now we construct the judgment matrix of the first-level evaluation index:

$$A=\begin{bmatrix} \frac{W_1}{W_1} & \frac{W_1}{W_2} & \frac{W_1}{W_3} & \frac{W_1}{W_4} & \frac{W_1}{W_5} & \frac{W_1}{W_6} \\ \frac{W_2}{W_1} & \frac{W_2}{W_2} & \frac{W_2}{W_3} & \frac{W_2}{W_4} & \frac{W_2}{W_5} & \frac{W_2}{W_6} \\ \frac{W_3}{W_1} & \frac{W_3}{W_2} & \frac{W_3}{W_3} & \frac{W_3}{W_4} & \frac{W_3}{W_5} & \frac{W_3}{W_6} \\ \frac{W_4}{W_1} & \frac{W_4}{W_2} & \frac{W_4}{W_3} & \frac{W_4}{W_4} & \frac{W_4}{W_5} & \frac{W_4}{W_6} \\ \frac{W_5}{W_1} & \frac{W_5}{W_2} & \frac{W_5}{W_3} & \frac{W_5}{W_4} & \frac{W_5}{W_5} & \frac{W_5}{W_6} \\ \frac{W_6}{W_1} & \frac{W_6}{W_2} & \frac{W_6}{W_3} & \frac{W_6}{W_4} & \frac{W_6}{W_5} & \frac{W_6}{W_7} \end{bmatrix} \tag{8}$$

Calculate the product of the elements:

$$M_i=\Pi_{i=1}^{n}a_{ij}\,(i=1,2,3,4,5,6) \tag{9}$$

Take the NTH root of M_i:

$$\overline{W_i}=\sqrt[n]{M_i} \tag{10}$$

Calculate the eigenvector:

$$W_i=\frac{\overline{W_i}}{\sum_{j=1}^{n}\overline{W_j}} \tag{11}$$

Calculate the maximum eigenvalue:

$$\lambda_{\max}=\frac{1}{n}\sum_{i=1}^{n}\frac{(AW)_i}{W_i} \tag{12}$$

Where $W=\begin{bmatrix}W_1 & W_2 & W_3 & W_4 & W_5 & W_6\end{bmatrix}^T$

4.3.2 AHP Consistency Test

After the judgment matrix is constructed, the relative weight of each element in two levels is calculated by the judgment matrix, and the consistency test is carried out. It is not allowed for the judgment to deviate too much from the consistency, so the consistency test of the judgment matrix is needed. The specific test steps are as follows:

Step1: Calculated consistency index CI :

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{13}$$

Step2:Check the criteria for testing the consistency of the judgement matrix from the relevant data $RI(n)$

Table 7: Relation Between RI and n

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46

Note: if the consistency ratio is less than 0.1, the judgment matrix of AHP has satisfactory consistency

Step3:Calculate the random consistency ratio of the judgment matrix CR **:**

$$CR = \frac{CI}{RI}\ ^{1} \tag{14}$$

If the consistency ratio is less than 0.1, the judgement matrix of AHP has satisfactory consistency, namely its consistency degree is acceptable.

4.3.3 Results of AHP

After the sentiment analysis based on LDA , we screened out 5 indexes with significant influence, and combined with the star ratings, there were 6 evaluation indexes. The judgment matrix of each index is given by the expert scoring method to reflect its relative critical degree. According to the result of data processing, the judgment matrix of these indexes has passed the consistency test, which is shown in the appendix.

Table 8:Index judgment matrix

	quality	appearance	price	size	service	star ratings
quality	1	2	1	2	2	1
appearance	0.5	1	1	1	2	1
price	1	1	1	2	1	1
size	0.5	1	0.5	1	0.5	1
service	0.5	0.5	1	2	1	0.5

Table 9:Evaluation Index and Weight

	quality	appearance	price	size	service	star ratings
Weight(microwave oven)	0.21	0.13	0.20	0.13	0.13	0.20
Weight(hair dryer)	0.22	0.14	0.21	0.10	0.13	0.20
Weight(pacifier)	0.28	0.27	0.10	0.15	0.10	0.10

4.4 System Clustering Analysis

4.4.1 Model establishment

Cluster analysis is a method of classification step by step. Its main idea is to reasonably merge and classify the research objects according to certain similarity indexes. It is called system clustering when it is used to solve the classification problem of samples and R-clustering when it is used to solve the classification problem of variables. We mainly uses cluster analysis to solve the problem of brand classification and evaluation. According to the observation indexes (star rating, quality, service, appearance, price, size) and system clustering algorithm of different brands, it calculates the similarity degree between brands, and classifies the similar brands into one category and the different brands into another. The closely related to a small classification unit, the not closely related to a large classification unit.

In a word, the result of system clustering analysis is to form a large to small classification pedigree or cluster diagram. Clustering graph can not only intuitively represent the similarity relationship and classification among the research objects, but also reflect all kinds of brands and quantitatively indicate the degree of similarity, so as to provide a good basis for the comprehensive evaluation.

Distance coefficient is a common statistic in system cluster analysis. If n brands observed on m variables are regarded as n points in m dimension space, the similarity between any two brands points x_j and x_k can be expressed by the distance between two points in m dimension space, then the distance coefficient is defined as:

$$d_{jk} = [\frac{1}{m}\sum_{i=1}^{m}(x_{ij} - x_{ik})^2]^{\frac{1}{2}} \tag{15}$$

Similarity coefficient is a measure of similarity between brands. Each brand is regarded as a vector of m-dimensional space, and the similarity between two brands x_j and x_k is defined as the cosine of the angle between two vectors, that is

$$\cos\theta_{jk} = \frac{\sum_{i=1}^{m} x_{ij}x_{ik}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2 \bullet \sum_{i=1}^{m} x_{ij}^2}} \tag{16}$$

4.4.2 Classification results

We use SPSS to cluster different parents and divide them into 10 grades. The first-class products are the best, and the tenth class products are the worst. In our sales strategy, we will recommend the first-class brands with better response in the sales market of sunshine company. Limited to space, only 5 first-class brands of each product are listed in the text.

Table 10:Part of Classification results

Microwave oven	862802057	423421857	692404913	464779766	423421857
Hair dryer	244516305	266176173	468944538	741916038	112413045
Pacifier	22060147	22189989	51496920	62352351	79207704

Note: The number in the table corresponds to the product parent.

5. Model 3-Time Series Analysis

In this section, we construct a time series analysis model to analyze microwave reputation, hair dryer reputation and pacifier reputation. Afterwards the product reputation could be predicted after analyzing the reviews and star ratings about the products.

The product reputation is consisted with star rating and emotional inclination. According to statist ics, the customers are apt to pay more attention to the reviews when they buy goods, so the review shows greater impact on the product reputation than the star rating. Because of this, we assign 70% weight to the average emotional score and 30% weight to the average star rating. The sum of them is the comprehensive score of product reputation.

Due to hypotheses, the product reputation score will be influenced by autocorrelation in time seri es. As a consequence, we choose p^{th} -order Auto-Regression Model to fit curve, namely AR(p).

$$Y_n = a_1 Y_{n-1} + a_2 Y_{n-2} + ... + a_p Y_{n-p} + \mu_n \tag{17}$$

Where Y represents the product reputation score in year n.

$a_1 \ldots\ldots a_p$ represent the influential coefficients of different lag orders.

μ_n represents the error term whose mean value is 0 and variance is σ^2 .The distribution matches White Noise Process WN(0, σ^2)

Firstly, we collected the time series data of three products' reputation scores. Then we solve the auto-regession model of the reputation score and calculate the time-series autocorrelation function and partial autocorrelation function of the reputation scores to identify the order, and specifically calculated the parameter p .

The sequence correlation of the comprehensive scores of each product's reputation depicts an obvious autoregressive structure AR1, which means the involved functions don't possess the property of truncation, and the partial functions possess the property of truncation. Therefore, we conclude that p=1 and the reputation score of all the three products share the same model setting.

$$N_t = \phi_1 N_{t-1} + \varepsilon_t \tag{18}$$

The mean variance of the coefficient ϕ_1 and the error term ε is different from the reputation changing. By calculation, we apply the MLE to estimate the coefficient of these three products. Here we list the results:

Table 11: ARE Model for Different Products

product	AR（1）	σ	P-Value	R^2
Hair dryer	0.125	0.046	0.007	0.901
pacifier	0.054	0.043	0.021	0.930
Microwave oven	-0.075	0.050	0.000	0.952

It is clear that the results of coefficients testing are significant, meanwhile the overall fitting gets a good result with a high R-squared.

Afterwards we conduct over-fitting and under fitting test, which contributes to the accuracy of p=1.The test uses the Algorithm of information guidelines to consider both the error term and the parameters' complexity.

$$AIC = \ln \sigma^2 + \frac{2p}{n} \tag{19}$$

$$BIC = \ln \sigma^2 + \frac{p}{n} \ln n \qquad (20)$$

Where n represents the size of sample.

The test illustrates that when p=1, AIC and BIC get the minimum value. Therefore, we consider the fitting of AR(1) is reasonable.

The AR (1) model is constructed to predict the comprehensive scores of product reputations. The results are as follows:

Taking the hair dryer as an example, in order to observe the relationship between the comprehensive score of product reputation and the time changing trends, we analyzed and predicted the time series in SPSS.

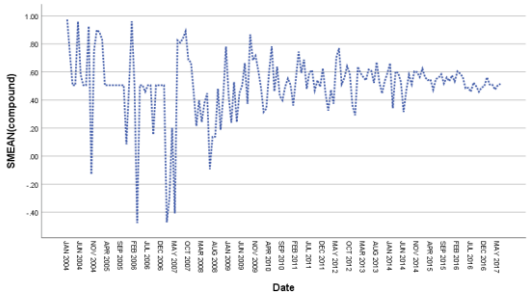
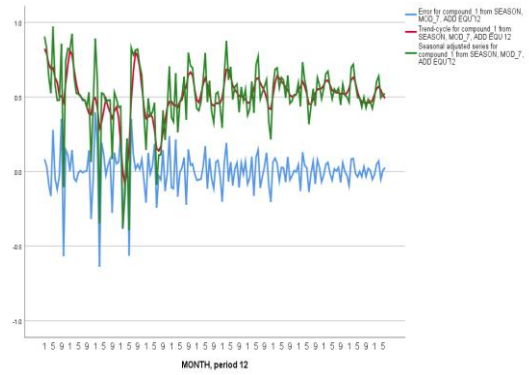


Figure 10: Sequence Figure

Through the above sequence diagram, we find that the seasonal fluctuation of the time series is basically constant, so we choose the addition model to decompose the seasonal factors. After removing the seasonal factors, the value of error sequence is very small, so the long-term trend and cyclic change sequence (long-term trend + cyclic change) and the sequence after seasonal factor correction (long-term trend + cyclic change + irregular change, that is, error Poor) can basically coincide.



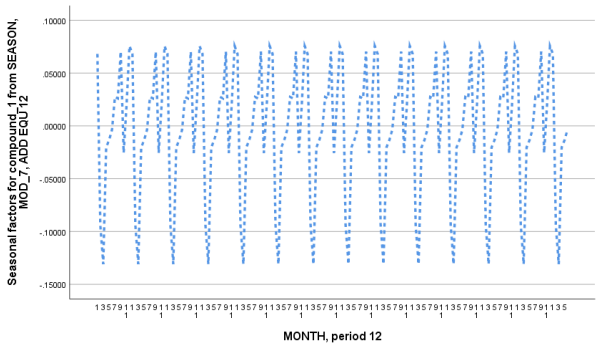


Figure 11:A sequence* chart that eliminates seasonal trends** ***Figure 12:The original sequence

By analyzing the seasonal trend, it can be found that the trend first drops in the first quarter of each year, then rises to the first peak. Then after a sharp decline, there is an obvious upward trend, reaching the second peak in about the eighth month, then comes a small decline. After that it reaches the peak in about October, and then falling till the beginning of the next year. In order to analyze the reasons for the seasonal changes, we introduced the annual sales volume with the time chart. It can be seen that the timing of reputation and sales have roughly the same seasonal trend, indicating that with the increase of sales, customers are more likely to give favorable reviews. So we conclude that it is consistent with the actual situation. In summer, people wash their hair more frequently. Therefore, they are apt to buy hair dryers from June to August and have a higher probability of making favorable reviews. However, when winter comes around October, the temperature is

colder and the blowing time becomes longer, which might also improve the demand and reputation of hair dryer.

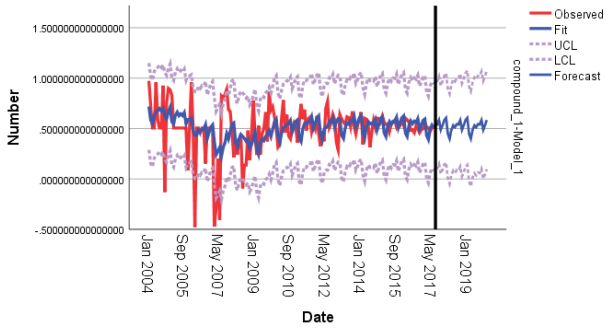


Figure 13: Time Series Forecasting

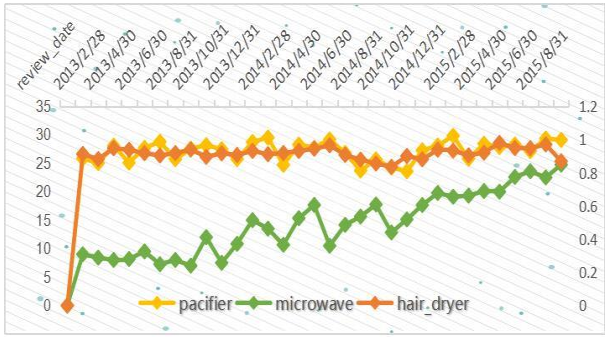


Figure 14:Annual sales for Three products

Through the prediction of time series, the data of 2016-2019 are obtained. It is found that the reputation and demand of hair dryer are highly related to the seasonality, and it is very likely to maintain a stable seasonal cycle mode in the future. Microwave ovens and pacifiers are about the same. The comprehensive score of the microwave oven reaches the peak in April and October, and that of the pacifier reaches the peak in November.

From the analysis above, it is believed that the market of hair dryer, microwave oven and pacifier is roughly stable. By referring to seasonal trend, sunshine company is capable to increase the sales share when the score of the product reputation rises up, or reduce the sales investment before the score is about to slump .

6. Model 4-Distributed lag model

6.1 Distributed lag model

In this part, we apply the distributed lag model to determine if the customer's reviews will be affected by others' star ratings.

The distributed lag model is based on the fact that the explained variables are affected by the explanatory variables and distributed on the lagged values of the explanatory variables in different periods.

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_s X_{t-s} \tag{21}$$

s is the lag length.

Each coefficient in this model reflects the different influence of each lag value of the explanatory variable on the explained variable, which is commonly referred to as multiplier effect:

β_0 : Short-term multiplier, which represents the average influence of one unit of star rating change on the emotional score of the review;

β_i : Delay multiplier, which represents the average influence of the change of one unit in the previous period star rating on the emotional score of the review;

$\sum_{i=0}^{s} \beta_i$: The long-term multiplier, which indicates the total impact of the star rating changing due to the lag effect.

6.2 Almon Method

The second-order Almon polynomial distribution lag model with a lag of 3 periods is used to establish the regression equation between current reviews and previous star ratings and reviews, so as to determine whether there is a phenomenon that previous star ratings affect current customer reviews. We select the data of the hair dryer whose product parent is 732252283, the pacifier of 392768822 and the microwave oven of 423421857 for analysis.

The following finite distribution lag model is estimated by Almon method. And the coefficients are approximated by quadratic polynomials.

$$\begin{aligned} Y_t &= \alpha + \beta_0 X_t + \beta_1 X_{t\text{-}1} + \beta_2 X_{t\text{-}2} + \beta_3 X_{t\text{-}3} + u_t \\ \beta_0 &= \alpha_0 \\ \beta_1 &= \alpha_0 + \alpha_1 + \alpha_2 \\ \beta_2 &= \alpha_0 + 2\alpha_1 + 4\alpha_2 \\ \beta_3 &= \alpha_0 + 3\alpha_1 + 9\alpha_2 \end{aligned} \tag{22}$$

Then the original model can be transformed into

$$\begin{aligned} Y_t &= \alpha + \alpha_0 Z_{0t} + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \mu_t \\ Z_{0t} &= X_t + X_{t-1} + X_{t-2} + X_{t-3} \\ Z_{1t} &= X_{t-1} + 2X_{t-2} + 3X_{t-3} \\ Z_{2t} &= X_{t-1} + 4X_{t-2} + 9X_{t-3} \end{aligned} \tag{23}$$

The final estimation formula of the distribution lag model can be obtained by regression. The regression results are as follows:

Table 12:Regression Results

	β_0	β_1	β_2	β_3	R^2
Hair dryer	1.13221**	0.32379**	-0.05354	0.00020	0.902197
Pacifier	0.79191**	0.21743**	-0.03039（0）	0.04845	0.929296
Microwave	1.05871**	0.17710**	-0.13628	0.11858	0.835728

Note:: p<0.05; **: p<0.01; +: p<0.1; ns: not significant*

The regression equation between the comprehensive scores of past star ratings and the emotional score of the current reviews of the three products is as follows:

$$\begin{aligned} Y_t &= -0.342472 + 1.13221\mathrm{X}_t + 0.32379 X_{t-1} \\ Y_t &= -0.054537 + 0.79191\mathrm{X}_t + 0.21743 X_{t-1} \\ Y_t &= -0.226388 + 1.05871\mathrm{X}_t + 0.17710 X_{t-1} \end{aligned} \tag{24}$$

Therefore, it can be explained that there is a correlation between the current reviews of customers and the past star ratings and reviews, that is, the current evaluation of customers will be affected by the past ratings and evaluations.

6.3 Correlation Analysis between Star Rating and Comments

According to the time series model, the inflection points of customer reviews can be gotten quickly. By analyzing the specific quality descriptors in the reviews corresponding to these inflection points, and

comparing them with the star ratings corresponding to the reviews, it is easy to know whether these words are closely related to the star ratings. In the above model, the average score of reviews and star ratings in each time period has been obtained. By comparing the synchronicity of the two changes over time, we can draw the conclusion that the star ratings corresponding to the reviews with positive words are higher, while the star ratings corresponding to the reviews with negative words are lower. (Due to the space, we only put the picture of microwave oven in the text, the picture of hair dryer and pacifier in the appendix.)

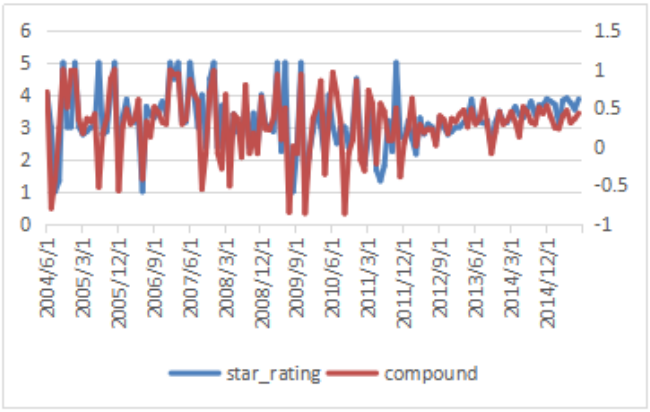


Figure 15: Sequence Diagram of Star rating and Review

However, it still can be seen from the diagrams that some reviews with higher star ratings have lower emotional scores, so it is not excluded that there are high scores, poor reviews (or vice versa) and false reviews.

7. Sensitivity Analysis

In the time series analysis model, we give star ratings and reviews 0.3 and 0.7 weight respectively, representing the comprehensive evaluation of each customer on the products they buy. Based on the principle that the importance of reviews is higher than that of star ratings, in order to test the rationality of the weight, we change the weight of star ratings to test whether the results of the model change significantly under the weight ratio of 0.4 and 0.2 respectively. The results are as follows:

Table 13: Sensitivity Analysis on Weight of Star Rating

Product	$AR(1)^a$	Change (in %)	$AR(1)^b$	Change (in %)
Hair Dryer	0.126	0.8%	0.127	20%
Pacifier	0.054	-0.9%	0.054	18%
Microwave Oven	-0.075	0.3%	-0.074	-0.6%

Note: $AR(1)^a$*: weight of star ratings: 0.4 ;* $AR(1)^b$*: weight of star ratings: 0.2*

From the table, we can see that the deviation of regression coefficient is not more than 1.8%, which shows that the change of weight has no significant impact on the results of the model, so the weight of 0.3 and 0.7 we set has certain rationality.

8. Model Assessment

8.1 Strengths

- Data visualization technology is applied to interpret the original data, and the results are presented intuitively and concisely.
- Through the analysis of LDA topic model, the strengths and weaknesses of the products can be given, which is helpful for the design of products.
- Our products brand rating system has a wide range of applicability, which can help sales companies grasp customer reviews on the market.

8.2 Weaknesses

- Our product scoring system uses AHP to determine the weight of each index, which is to some extent subjective.
- Our time series model uses monthly data, which is not very accurate and does not take into account special circumstances such as holidays, and may have a large gap with the actual results, and cannot effectively predict the product's reputation in the long term.

Reference

[1]Arreola Elsa Vazquez,Wilson Jeffrey R. Bayesian multiple membership multiple classification logistic regression model on student performance with random effects in university instructors and majors.[J]. PloS one,2020,15(1).

[2]Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:2003.

[3]Cao Juan, Xia Tian, Li Jin Tao, A density method for adaptive LDA model selection[J]. Neurocomputing 2009(72):1775-1781.

[4]Connors, L., Mudambi, S.M., Schuff, D.. Is It the Review or the Reviewer? a Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness[P]. System Sciences (HICSS), 2011 44th Hawaii International Conference on,2011.

[5]Changxuan Wan,Yun Peng,Keli Xiao,Xiping Liu,Tengjiao Jiang,Dexi Liu. An association-constrained LDA model for joint extraction of product aspects and opinions[J]. Information Sciences,2020,519.

[6]Cheung, Christy M. K., and Dimple R. Thadani. "The Effectiveness of Electronic Word-of-Mouth Communication: A Literature Analysis." Bled EConference, 2010, p. 18.

[7]Kim, Soo-Min, et al. "Automatically Assessing Review Helpfulness." Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 423–430.

[8]Mudambi, Susan M.,Schuff, David,Zhang, Zhewei. Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings[P]. ,2014.

[9]Yan Leng,Weiwei Zhao,Chan Lin,Chengli Sun,Rongyan Wang,Qi Yuan,Dengwang Li. LDA-based data augmentation algorithm for acoustic scene classification[J]. Knowledge-Based Systems,2020.

Letter

Dear Sunshine Company Marketing Director:

We are honored to inform you of our recommendations for product improvement and sales strategies for your company after data analysis and modeling. The following are some suggestions based on our analysis.

1. Suggestions for product improvement

We apply the LDA analysis model to find the topic feature of each product, in which the negative topic words can reflect the customers' dissatisfaction with the existing products in the market.

Analysis of the review text revealed that the words like "cost", "heavy", "smoking" and "noise" appeared on the top of the list of negative reviews on the hair dryer. After analyzing these high-frequency words one by one, we discovered that there were some drawbacks with hair dryers from the customers' point of view:

- Inefficiency: it takes a long time to blow-dry.
- Weight: it is too heavy to use conveniently.
- Bad quality: when using the product, there will be a lot of noise sometimes.

Therefore, according to the feedback above, we believe that the improvement of hair dryer products is supposed to start from improving the working efficiency and quality of hair dryer. Lightweight materials can be used in the production, reducing the weight of the hair dryer, should also reduce the noise when it works. For instance, lightweight materials can be used to reduce the weight of the hair dryer and reduce the noise while it is working.

As for microwave oven, words like "service" and "small" appeared on the top of the list of negative reviews. According to the feedback, we believe that the microwave oven may have the following defects:

- Size: someone considered that microwave ovens are too small, while others reckon that small ones are more convenient to use.
- Service: the after-sale service of some microwave ovens is not consummate.

In view of the above two points, we hold that microwave oven products can be designed in different sizes to meet the diverse needs of different customers. Moreover, improving service level and customer satisfaction is also a top priority.

"Hot" is the most frequently used negative word for a pacifier. Some customers think that the insulation function of the pacifier is poor, which reminds us that we should enhance the thermal insulation function in the design of the pacifier. In addition, we have noticed that customers attach great importance to the "appearance" of the pacifier and hence the cute product design will help to sell the product.

2. Sales strategy

Through the in-depth analysis of star ratings and reviews, we will propose sales suggestions and strategies from three aspects: the product brand, the time of selling products and the evaluation psychology of customers.

- Recommend the product brand for sale.

Based on the analysis of star rating and review, we summarized five indexes that affect product sales, namely quality, price, appearance, service and size, as well as index weight, then built a brand scoring system based on this. Through the systematic clustering of each score, all product brands were clustered, and the high-quality product brands were selected eventually. The product parent of some premium brands is as follows:

microwave	862802057	423421857	692404913	464779766	423421857
Hair-dryer	244516305	266176173	468944538	741916038	112413045
pacifier	22060147	22189989	51496920	62352351	79207704

Our brand rating system takes full account of the characteristics of each product, so the conclusions are very reliable. It is believed that the high quality brands recommended for sale will be favored by more customers.

- Sales opportunities analysis.

we predict the future reputation of the three products through time series analysis. The results demonstrate that the three products are highly correlated with seasonality, and the seasonal cycle pattern is likely to be stable in the future. However, the peaks of reputation composite scores of the three products occurred at different times.

Pacifiers peaked around November, hairdryers around August and October, and microwave ovens around April and October. The peak of reputation will lead to an increase in sales, so we suggest that your company adjust the sales structure according to the peak time of different products. When the product reputation is about to peak, your company can increase the sales investment of the product to obtain higher revenue.

- Psychological analysis of customer reviews.

We analyze the relationship between star rating and review. Through the research of distributed lag model, it is found that customers' reviews in the current period will be influenced by other customers' ratings and reviews. Therefore, we suggest that your company should still pay attention to the customer's rating of the goods after selling the products, and try to make the rating of its own products at a high level, so as not to gain a bad impact on future sales. To sum up, by keeping an eye on customers' preferences and regarding them as standards for improving products and services, your company can gradually increase product sales and market share.

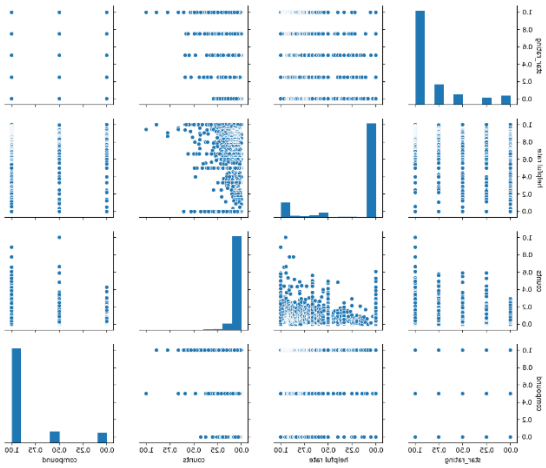
These are all suggestions and strategies our team has provided to your company. Thank you again for taking the time to read our suggestions.

Hope that our models and these suggestions can be helpful to you!

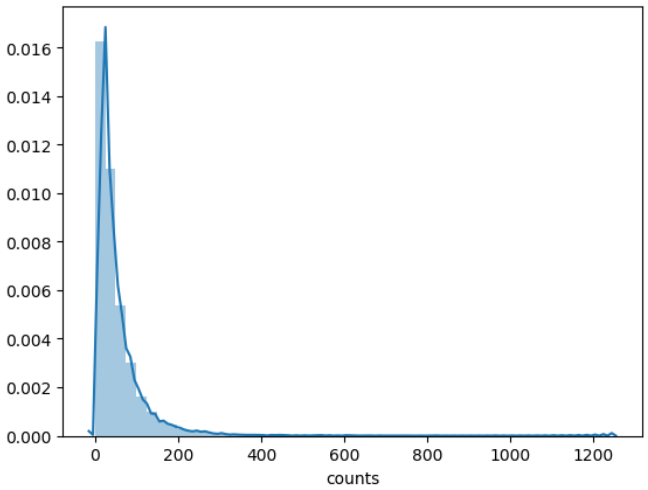
Sincerely,
MCM Team Members

Appendix

Appendix 1



Correlation analysis



The distribution of the number of words

- **Consistency test of AHP**

Consistency check ratio table

Consistency ratio	quality	appearance	price	size	service	Star ratings
CR	0.00000	0.00857	0.00688	0.00817	0.00053	0.00047

Appendix 2

Regression results of distribution lag model

1. Hair dryer

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.342472	0.066430	-5.155354	0.0000
PDL01	0.323795	0.040572	7.980680	0.0000
PDL02	-0.592878	0.040352	-14.69277	0.0000
PDL03	0.215540	0.028612	7.533150	0.0000

			0.64829
R-squared	0.902197	Mean dependent var	3
			0.22365
Adjusted R-squared	0.896444	S.D. dependent var	9
			-2.3550
S.E. of regression	0.071974	Akaike info criterion	86
			-2.2090
Sum squared resid	0.264190	Schwarz criterion	99
			-2.2986
Log likelihood	68.76488	Hannan-Quinn criter.	32
			1.71152
F-statistic	156.8187	Durbin-Watson stat	2
Prob(F-statistic)	0.000000		

Lag Distribution of TOTAL		Coefficient	Std. Error	t-Statistic
.	*|	1.13221	0.05595	20.2356
. *	|	0.32379	0.04057	7.98068
		-0.0535		-1.3294
*.	|	4	0.04027	8
*	|	0.00020	0.05539	0.00358
	Sum of Lags	1.4026 6	0.09243	15.175 1

2. Pacifier

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.05453 7	0.10941 9	-0.498424	0.6201
PDL01	0.217428	0.05220 1	4.165216	0.0001
PDL02	-0.41115 2	0.05242 4	-7.842873	0.0000
PDL03	0.163331	0.03664 1	4.457639	0.0000

			0.73679
R-squared	0.629296	Mean dependent var	9
			0.16242
Adjusted R-squared	0.609786	S.D. dependent var	3
			-1.6749
S.E. of regression	0.101461	Akaike info criterion	65

			-1.5365
Sum squared resid	0.586774	Schwarz criterion	47
			-1.6207
Log likelihood	55.08642	Hannan-Quinn criter.	17
			1.59103
F-statistic	32.25389	Durbin-Watson stat	8
Prob(F-statistic)	0.000000		

Lag Distribution of TOTAL		Coefficient	Std. Error	t-Statistic
.	*|	0.79191	0.08171	9.69200
. *	|	0.21743	0.05220	4.16522
*.	|	-0.0303 9	0.04982	-0.6099 9
.*	|	0.04845	0.06081	0.79669
	Sum of Lags	1.0273 9	0.14225	7.2225 3

3. Microwave oven

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.22638 8	0.07200 6	-3.144010	0.0030
PDL01	0.177103	0.05482 2	3.230494	0.0023
PDL02	-0.59749 4	0.05752 5	-10.38663	0.0000
PDL03	0.284116	0.04604 5	6.170329	0.0000

			0.52140
R-squared	0.835728	Mean dependent var	9
			0.22024
Adjusted R-squared	0.824528	S.D. dependent var	0
			-1.8488
S.E. of regression	0.092257	Akaike info criterion	16
			-1.6928
Sum squared resid	0.374502	Schwarz criterion	83
			-1.7898
Log likelihood	48.37159	Hannan-Quinn criter.	89
			1.49272
F-statistic	74.61626	Durbin-Watson stat	4

Prob(F-statistic)		0.000000		

Lag Distribution of TOTAL		Coefficient	Std. Error	t-Statistic
.	*|	1.05871	0.07634	13.8681
. *	|	0.17710	0.05482	3.23049
* .	|	-0.13628	0.05500	-2.47759
. *	|	0.11858	0.07611	1.55802
	Sum of Lags	1.21812	0.11498	10.5944

Appendix 3

Partial system clustering results

1.

Microwave oven	10 clusters	9 clusters	8 clusters	7 clusters	6 clusters	5 clusters
1:109226352	1	1	1	1	1	1
2:147401377	2	2	2	2	2	2
3:149559260	3	3	3	3	3	3
4:155528792	2	2	2	2	2	2
5:166483932	4	4	4	4	4	2
6:168181302	4	4	4	4	4	2
7:215953885	2	2	2	2	2	2
8:242727854	5	5	4	4	4	2
9:295520151	4	4	4	4	4	2
10:30560899	1	1	1	1	1	1
11:30926741	1	1	1	1	1	1

12:311592014	6	6	5	5	5	4
13:313983847	7	7	6	6	6	5
14:379992322	4	4	4	4	4	2
15:392967251	2	2	2	2	2	2
16:423421857	1	1	1	1	1	1
17:454581724	7	7	6	6	6	5
18:459626087	1	1	1	1	1	1
19:464779766	4	4	4	4	4	2
20:486381187	5	5	4	4	4	2
21:494028413	2	2	2	2	2	2
22:494668275	5	5	4	4	4	2
23:522487135	6	6	5	5	5	4
24:523301568	1	1	1	1	1	1
25:539049610	2	2	2	2	2	2
26:542519500	5	5	4	4	4	2
27:542731946	2	2	2	2	2	2
28:544821753	2	2	2	2	2	2
29:550562680	8	8	7	7	5	4

2.

Pacifier	10 clusters	9 clusters	8 clusters	7 clusters	6 clusters	5 clusters
1:723849	1	1	1	1	1	1
2:1006724	2	2	2	2	2	2

3: 1398002	3	3	1	1	1	1
4: 1439995	3	3	1	1	1	1
5: 1448183	4	2	2	2	2	2
6: 1696639	3	3	1	1	1	1
7: 1892472	4	2	2	2	2	2
8: 2143250	5	4	3	3	3	3
9: 2332208	4	2	2	2	2	2
10: 2341622	4	2	2	2	2	2
11: 2775015	4	2	2	2	2	2
12: 3090006	6	5	4	4	3	3
13: 3146962	3	3	1	1	1	1
14: 3179934	5	4	3	3	3	3
15: 3729223	2	2	2	2	2	2
16: 3916839	4	2	2	2	2	2
17: 4145037	1	1	1	1	1	1
18: 4569674	4	2	2	2	2	2
19: 4649401	3	3	1	1	1	1
20: 4792175	2	2	2	2	2	2
21: 5180901	2	2	2	2	2	2
22: 5471085	2	2	2	2	2	2
23: 5645959	4	2	2	2	2	2
24: 5747909	4	2	2	2	2	2

25: 5749221	4	2	2	2	2	2
26: 5848633	4	2	2	2	2	2
27: 5981131	4	2	2	2	2	2
28: 6156155	2	2	2	2	2	2
29: 6744486	2	2	2	2	2	2
30: 6784496	2	2	2	2	2	2
31: 7090204	4	2	2	2	2	2

3.

Hair dryer	10 clusters	9 clusters	8 clusters	7 clusters	6 clusters
1: 423960	1	1	1	1	1
2: 4120409	2	2	2	1	1
3: 11468070	3	3	3	2	2
4: 12536427	2	2	2	1	1
5: 14552349	2	2	2	1	1
6: 16483457	1	1	1	1	1
7: 16983648	2	2	2	1	1
8: 21033180	2	2	2	1	1
9: 21750700	2	2	2	1	1
10: 26711891	1	1	1	1	1
11: 30965255	3	3	3	2	2
12: 44138644	4	4	4	3	3
13: 44703144	5	5	5	4	4

14: 45575190	1	1	1	1	1
15: 46450049	2	2	2	1	1
16: 46677591	2	2	2	1	1
17: 47684938	1	1	1	1	1
18: 50000317	2	2	2	1	1
19: 54378879	2	2	2	1	1
20: 54987170	4	4	4	3	3
21: 55445525	1	1	1	1	1
22: 55520986	2	2	2	1	1
23: 57056668	1	1	1	1	1
24: 61225676	4	4	4	3	3
25: 62808517	6	1	1	1	1
26: 64142513	2	2	2	1	1
27: 66014174	1	1	1	1	1
28: 66259499	1	1	1	1	1
29: 66279275	2	2	2	1	1
30: 68100320	1	1	1	1	1
31: 68816102	2	2	2	1	1
32: 71698270	4	4	4	3	3
33: 74202592	7	6	6	5	5
34: 74735317	1	1	1	1	1
35: 77898021	2	2	2	1	1

36: 80193353	8	7	7	6	2
37: 84440271	1	1	1	1	1
38: 91277457	2	2	2	1	1
39: 98133587	1	1	1	1	1
40: 99665579	2	2	2	1	1
41:10734 1965	1	1	1	1	1
42:10819 1918	1	1	1	1	1
43:10910 6777	1	1	1	1	1
44:11093 5305	2	2	2	1	1
45:11241 3045	9	8	5	4	4
46:11526 4052	2	2	2	1	1

Appendix 4

Part of the monthly score of review and star rating

1. Hair dryer

review_date	star_rating	compound
2002/3/31	3	0.9736
2002/4/30	5	0.7313
2002/5/31	3.829601	0.505325
2002/6/30	3.829601	0.505325
2002/7/31	5	0.9592
2002/8/31	3	0.5793
2002/9/30	3.829601	0.505325

2002/10/31	3.829601	0.505325
2002/11/30	4	0.9261
2002/12/31	5	-0.1297
2003/1/31	4.5	0.7576
2003/2/28	4	0.899
2003/3/31	4	0.88425
2003/4/30	5	0.8294
2003/5/31	3.829601	0.505325
2003/6/30	3.829601	0.505325
2003/7/31	3.829601	0.505325
2003/8/31	3.829601	0.505325
2003/9/30	3.829601	0.505325
2003/10/31	3.829601	0.505325
2003/11/30	3.829601	0.505325
2003/12/31	3.829601	0.505325
2004/1/31	2.5	0.08485
2004/2/29	3.829601	0.505325
2004/3/31	5	0.95955
2004/4/30	3.829601	0.505325
2004/5/31	1	-0.4767
2004/6/30	3.829601	0.505325
2004/7/31	3.829601	0.505325

2004/8/31	2	0.4588
2004/9/30	3.829601	0.505325
2004/10/31	3.829601	0.505325
2004/11/30	4	0.15315
2004/12/31	3.829601	0.505325
2005/1/31	3.829601	0.505325
2005/2/28	3.829601	0.505325
2005/3/31	3.829601	0.505325
2005/4/30	1	-0.4714
2005/5/31	1	-0.2966
2005/6/30	5	0.2008
2005/7/31	5	-0.4066
2005/8/31	3	0.8255
2005/9/30	4.333333	0.8087
2005/10/31	5	0.839467
2005/11/30	3.5	0.88995
2005/12/31	4.5	0.68685
2006/1/31	3.357143	0.667943
2006/2/28	2.875	0.453613
2006/3/31	3.309524	0.2146
2006/4/30	3	0.396333
2006/5/31	1.666667	0.2426

2006/6/30	2	0.37806
2006/7/31	3.333333	0.447167

2. Pacifier

review_date	star_rating	compound
2003/4/30	2	0.5835
2003/5/31	4.4	0.60842
2003/6/30	4.333333	0.770685
2003/7/31	3.375	0.742837
2003/8/31	4.111111	0.487467
2003/9/30	4.210547	0.571278
2003/10/31	4.55	0.751675
2003/11/30	3.605442	0.523395
2003/12/31	1	0.92
2004/1/31	4.137209	0.697009
2004/2/29	4.196297	0.703307
2004/3/31	4.146052	0.45704
2004/4/30	4.5	0.6737
2004/5/31	2	0.9186
2004/6/30	2	0.8108
2004/7/31	4.499116	0.653345
2004/8/31	4.455441	0.646819
2004/9/30	4.361652	0.581429

2004/10/31	4.452883	0.646562
2004/11/30	4.379498	0.662509
2004/12/31	4.344347	0.62298
2005/1/31	4.333011	0.593006
2005/2/28	5	0.9696
2005/3/31	3.534091	0.492098
2005/4/30	3.883333	0.639112
2005/5/31	4.391507	0.50169
2005/6/30	4.296748	0.49209
2005/7/31	4.434554	0.525116
2005/8/31	4.332839	0.495799
2005/9/30	4.344876	0.519694
2005/10/31	5	0.8472
2005/11/30	5	0.9689
2005/12/31	4.541667	0.724842
2006/1/31	4.6375	0.759315
2006/2/28	4	0.9406
2006/3/31	5	0.9494
2006/4/30	3.534091	0.492098
2006/5/31	2.7	0.75511
2006/6/30	3.666667	0.21575
2006/7/31	4.267442	0.540129

2006/8/31	5	0.9691
2006/9/30	4.25	0.735613
2006/10/31	5	0.95905
2006/11/30	3	0.7349

3. Microwave oven

review_date	star_rating	compound
2004/6/30	4	0.705233
2004/7/31	3	-0.7992
2004/8/31	1	-0.4215
2004/9/30	1.333333	0.241867
2004/10/31	5	0.9925
2004/11/30	3	0.5063
2004/12/31	3	0.9818
2005/1/31	5	0.9854
2005/2/28	3	0.347794
2005/3/31	2.773333	0.169515
2005/4/30	2.865385	0.360717
2005/5/31	3	0.320475
2005/6/30	3	0.418019
2005/7/31	5	-0.5267
2005/8/31	2.773333	0.169515
2005/9/30	2.865385	0.360717

2005/10/31	4	0.874467
2005/11/30	5	0.9925
2005/12/31	2	-0.5707
2006/1/31	3.294118	0.253166
2006/2/28	3.857143	0.4814
2006/3/31	3.177083	0.290946
2006/4/30	3.15625	0.352928
2006/5/31	3.115385	0.602631
2006/6/30	1	-0.4215
2006/7/31	3.634921	0.319398
2006/8/31	3.384615	0.126777
2006/9/30	3.291667	0.510772
2006/10/31	3.533333	0.445201
2006/11/30	3.793333	0.313961
2006/12/31	3.378788	0.290235
2007/1/31	5	0.9868
2007/2/28	4.5	0.9294
2007/3/31	5	0.9459
2007/4/30	3.177083	0.290946
2007/5/31	3.15625	0.352928
2007/6/30	5	0.8613
2007/7/31	4	0.6808

2007/8/31	3	0.5866
2007/9/30	4	-0.5529
2007/10/31	2.364583	-0.08993
2007/11/30	4.5	0.7006
2007/12/31	5	0.97 7667
2008/1/31	2.364583	-0.08993
2008/2/29	3.666667	-0.28567
2008/3/31	3	0.6705
2008/4/30	2	-0.5106
2008/5/31	3.4	0.41524
2008/6/30	2.865385	0.360717
2008/7/31	2.666667	-0.13823
2008/8/31	3.666667	0.795633
2008/9/30	2.364583	-0.08993
2008/10/31	3.433333	0.172403
2008/11/30	2.364583	-0.08993
2008/12/31	4	0.6369
2009/1/31	3.1	0.22467
2009/2/28	3	0.214
2009/3/31	2.865385	0.360717
2009/4/30	5	0.9257
2009/5/31	2.25	0.089067

2009/6/30	5	0.49495
2009/7/31	1	-0.8516
2009/8/31	1	0

Appendix 5

1. Code of LDA model

```
import pandas as pd
from snownlp import SnowNLP
import re
from gensim import corpora,models
#from nltk.tokenize import word_tokenize
data = pd.DataFrame(test2)
print(type(data))
#data_null = data.drop_duplicates()
#data_null.to_csv(&apos;C:/Users/lenovo/Desktop/comments_null.csv&apos;)
#data_null_comments = data_null[&apos;contents&apos;]
#data_null_comments.to_csv(&apos;C:/Users/lenovo/Desktop/contents.txt&apos;,index=False,encoding=&apos;utf-8&apos;)
#data_len = data_null_comments[data_null_comments.str.len()>4]
#print(data_len)
#data_len.to_csv(&apos;contents.txt&apos;,index=False,encoding=&apos;utf-8&apos;)
coms = []
coms = data[0].apply(lambda x:SnowNLP(x).sentiments)
data_post = data[coms>=0.01]
data_neg = data[coms<0.01]
print(data_post)
print(data_neg)
data_post[0].to_csv(&apos;C:/Users/lenovo/Desktop/comments_positive.txt&apos;,encoding=&apos;utf-8&apos;,header=None)
data_neg[0].to_csv(&apos;C:/Users/lenovo/Desktop/comments_negative.txt&apos;,encoding=&apos;utf-8-sig&apos;,header=None)
with open(&apos;C:/Users/lenovo/Desktop/comments_positive.txt&apos;,encoding=&apos;utf-8&apos;) as fn1:
    string_data1 = fn1.read()
    pattern = re.compile(u&apos;\t|\n|\.|-|——|：|！|、|，|,|。|;|\)|\(|\?|"&apos;)
    string_data1 = re.sub(pattern, &apos;&apos;, string_data1)
    print(string_data1)
    fp = open(&apos;C:/Users/lenovo/Desktop/comments_post.txt&apos;,&apos;a&apos;,encoding=&apos;utf8&apos;)
```

```
    fp.write(string_data1 + &apos;\n&apos;)
    fp.close()
with open(&apos;C:/Users/lenovo/Desktop/comments_nagative.txt&apos;,encoding=&apos;utf-8&a
pos;) as fn2:
    string_data2 = fn2.read()
    pattern = re.compile(u&apos;\t|\n|\.|-|——|：|！|、|，|,|。|;|\)|\(|\?|"&apos;)
    string_data2 = re.sub(pattern, &apos;&apos;, string_data2)
    print(string_data2)
    fp = open(&apos;C:/Users/lenovo/Desktop/comments_neg.txt&apos;,&apos;a&apos;,encoding
=&apos;utf8&apos;)
    fp.write(string_data2 + &apos;\n&apos;)
    fp.close()
data1 = pd.read_csv(&apos;C:/Users/lenovo/Desktop/comments_post.txt&apos;,encoding=&apos;utf
-8&apos;,header=None)
data2 = pd.read_csv(&apos;C:/Users/lenovo/Desktop/comments_neg.txt&apos;,encoding=&apos;utf
-8&apos;,header=None)
#mycut = lambda s: &apos; &apos;.join(word_tokenize(s))
#data1 = data1[0].apply(mycut)
#data2 = data2[0].apply(mycut)
#mycut = lambda s: &apos; &apos;.join(word_tokenize(s))
data1 = data1[0]
data2 = data2[0]
data1.to_csv(&apos;C:/Users/lenovo/Desktop/comments_post_cut.txt&apos;,index=False,header=Fa
lse,encoding=&apos;utf-8&apos;)
data2.to_csv(&apos;C:/Users/lenovo/Desktop/comments_neg_cut.txt&apos;,index=False,header=Fal
se,encoding=&apos;utf-8&apos;)
print(data2)
post = pd.read_csv(&apos;C:/Users/lenovo/Desktop/comments_post_cut.txt&apos;,encoding=&apos
;utf-8&apos;,header=None,error_bad_lines=False)
neg = pd.read_csv(&apos;C:/Users/lenovo/Desktop/comments_neg_cut.txt&apos;,encoding=&apos;
utf-8&apos;,header=None,error_bad_lines=False)
stop = pd.read_csv(&apos;C:/Users/lenovo/Desktop/stoplist.txt&apos;,encoding=&apos;utf-8&apos;
,header=None,sep=&apos;tipdm&apos;,engine=&apos;python&apos;)
stop = [&apos; &apos;,&apos;&apos;] + list(stop[0])
post[1] = post[0].apply(lambda s: s.split(&apos; &apos;))
post[2] = post[1].apply(lambda x: [i for i in x if i not in stop])
neg[1] = neg[0].apply(lambda s: s.split(&apos; &apos;))
neg[2] = neg[1].apply(lambda x: [i for i in x if i not in stop])
post_dict = corpora.Dictionary(post[2])
post_corpus = [post_dict.doc2bow(i) for i in post[2]]
post_lda = models.LdaModel(post_corpus, num_topics=4, id2word=post_dict)
for i in range(3):
    print(post_lda.print_topic(i))
print(&apos)
```

```
neg_dict = corpora.Dictionary(neg[2])
neg_corpus = [neg_dict.doc2bow(i) for i in neg[2]]
neg_lda = models.LdaModel(neg_corpus, num_topics=4, id2word=neg_dict)
for i in range(3):
   print(neg_lda.print_topic(i))
```

2. Code of word segmentation and NLTK model

```
      import pandas as pd
import seaborn as sns
import numpy as np
from nltk.tokenize import word_tokenize,sent_tokenize
import matplotlib.pyplot as plt
pacifier=pd.read_csv(&apos;C:/Users/lenovo/Desktop/pacifier.tsv&apos;, sep=&apos;\t&apos;)
b = pacifier[&apos;review_body&apos;].duplicated()
pacifier[&apos;marketplace&apos;] = pacifier[&apos;marketplace&apos;].str.lower()
pacifier[&apos;product_category&apos;] = pacifier[&apos;product_category&apos;].str.lower()
pacifier[&apos;vine&apos;] = pacifier[&apos;vine&apos;].str.lower()
pacifier[&apos;verified_purchase&apos;] = pacifier[&apos;verified_purchase&apos;].str.lower()
text_a_count = []
for i in pacifier[&apos;review_body&apos;]:
   if isinstance(i,str):
      print(i,type(i))
      print(len(i))
   elif isinstance(i,float):
      print(i,type(i))
      i = str(i)
   text_a_count.append(len(i))
#plt.figure(figsize = (8,6))
u = pacifier[&apos;helpful_votes&apos;]
x = np.array(text_a_count)
y = np.array(u)
p = pd.DataFrame({&apos;review_body_length&apos;:x,&apos;helpful_votes&apos;:y})
#p = p.drop_duplicates([&apos;helpful_votes&apos;])
#color = [&apos;r&apos;,&apos;y&apos;,&apos;k&apos;,&apos;g&apos;,&apos;m&apos;]
#plt.scatter(p[&apos;review_body_length&apos;],p[&apos;helpful_votes&apos;],c=color,marker=&
apos;>&apos;)
#plt.legend()
#plt.show()
v = pacifier.describe()
      import pandas as pd
import nltk
import numpy as np
from nltk.tokenize import word_tokenize,sent_tokenize
pacifier=pd.read_excel(&apos;C:/Users/lenovo/Desktop
```

```
/pacifier_lower.xlsx&apos;, sep=&apos;\t&apos;)
b = pacifier[&apos;review_body&apos;].duplicated()
test11 = pacifier.groupby(&apos;product_parent&apos;).sum()
pacifier[&apos;marketplace&apos;] = pacifier[&apos;marketplace&apos;].str.lower()
pacifier[&apos;product_category&apos;] = pacifier[&apos;product_category&apos;].str.lower()
pacifier[&apos;vine&apos;] =pacifier[&apos;vine&apos;].str.lower()
pacifier[&apos;verified_purchase&apos;] = pacifier[&apos;verified_purchase&apos;].str.lower()
test2 = pacifier[&apos;review_body&apos;].tolist()
text_a_count = []
text_b_count = []
for sent in test2:
    words = word_tokenize(sent)
        text_b_count.append(words)
    for word in words:
        if word.isalpha() == False:
            words.remove(word)
    text_a_count.append(len(words))
    counts1 = {}
    for word in words:
        counts1[word] = counts1.get(word,0) + 1
        items = list(counts1.items())
        items.sort(key=lambda x:x[1], reverse=True)
    for i in range(len(items)):
        word, count1 = items[i]
        print ("{0:<10}{1:>5}".format(word, count1))
        from collections import Counter
import pandas as pd
import numpy as np
c={"counts" : text_a_count,
  "split" : text_b_count}
data=pd.DataFrame(c)
w = Counter(text_a_count)
print(w)
w1 = set(text_a_count)
print(text_a_count)
w2 = w.most_common(len(w))
m = []
for i in range(305):
    m.append(list(w2[i]))
df = pd.DataFrame(m, columns=[&apos;number&apos;, &apos;count&apos;])
color = [&apos;r&apos;,&apos;y&apos;,&apos;k&apos;,&apos;g&apos;,&apos;m&apos;]
plt.scatter(df[&apos;number&apos;],df[&apos;count&apos;],c=color,marker=&apos;>&apos;)
plt.legend()
plt.show()
```