

Main Document

Contents

Letter to the Marketing Director	2
1 Introduction	4
1.1 Background	4
1.2 Our work	4
2 Notation and Assumptions	4
3 Relationship between variables	5
3.1 Descriptive statistics	5
3.2 Reviews that triggered by a specific star rating	6
3.2.1 Research ideas	6
3.2.2 Model construction and solution	6
3.3 Relationship between <i>Star rating</i> and specific emotion words	8
4 Identify useful reviews	10
4.1 Research ideas	10
4.2 Model establishing	10
4.3 Model Result	12
5 GD, SGD and Neural Network	13
5.1 Logit model based on GD and SGD optimization methods	13
5.1.1 Model establishing	13
5.1.2 Model solving	13
5.2 Neural network model based on SGD optimization method	14
5.2.1 Model establishing	14
5.2.2 Model solving	15
5.2.3 Model results	16
5.2.4 Cross validation	16
6 Track the real utility of customers	17
6.1 Research ideas	17
6.2 Feature extraction	17
6.3 Construction of Customer Utility Function	18
6.4 Model Result	19
7 Product Reputation and Potential Successful Product	20
7.1 Model Establishing	20
7.2 Model Solving	20
7.3 Product Forecast	22
8 Strengths and Weaknessess	22
8.1 Strengths	22
8.2 Weaknessess	22
References	23
Appendix	24

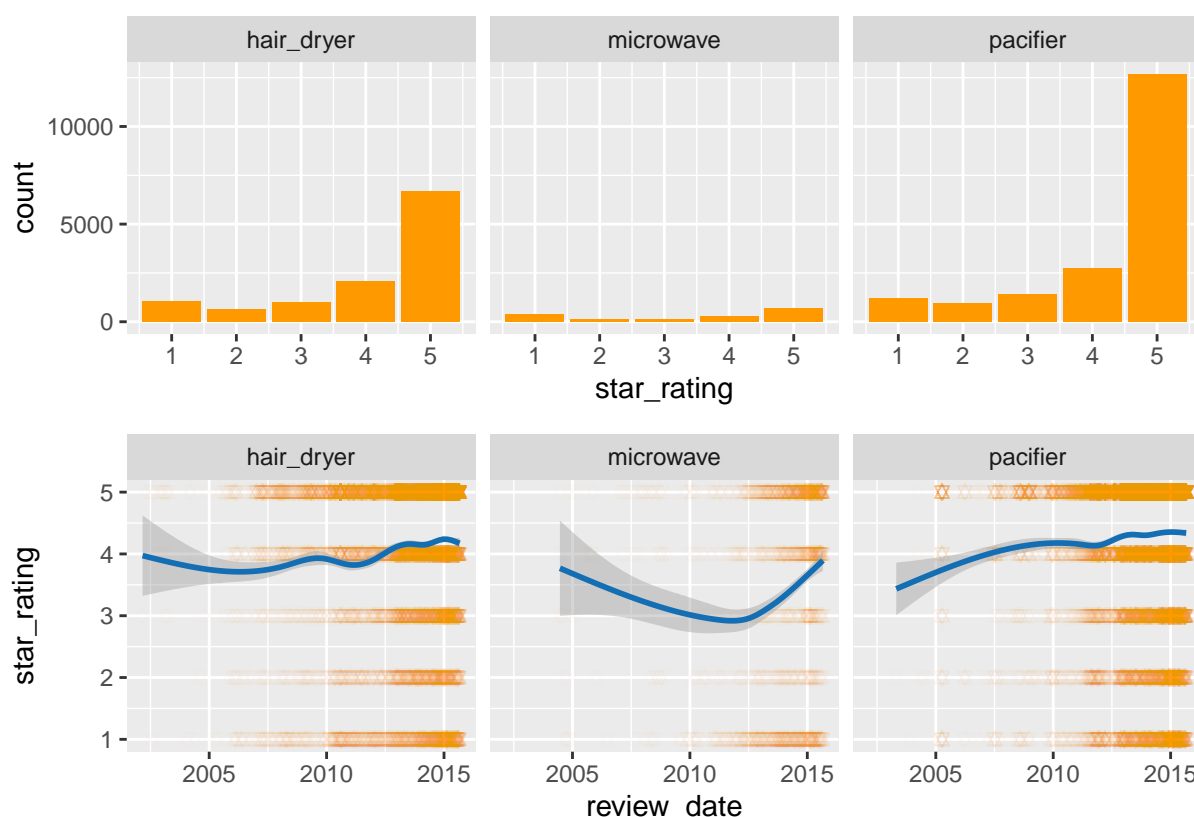


Figure 1: Relationship between star rating and review date

Letter to the Marketing Director

Dear Sunshine Company's Marketing Director:

Thanks for inviting us! We have performed an analysis based on the data you gave. The following are our analysis results and reasons.

- **You might consider push the hair dryer and pacifier to the market as soon as possible, but not microwave, which should be published later.**

Judging from the ratings, the number of five-stars in the hair dryer and pacifier customer ratings is much higher than other stars. But the microwave has the lowest five star rating ratio. Microwaves are search-type products, which have high requirements on rigid technology. Customers pay more attention to the rigid indicators of the product, and it is more difficult to get five star rating than the other two products.

As for time, the average score of hair dryer and pacifier has gradually increased with the passage of time, while the average score of microwave has gradually decreased first, reached a trough in 2012, and then gradually recovered. In general, the market performance of microwave ovens is not stable enough. The other types of products currently have a good development trend, and are gradually recognized by more and more customers. Therefore, hair dryer and pacifier should enter the market as soon as possible and occupy market share.

But we need to more time to improve quality of microwave and find a good time to participate.

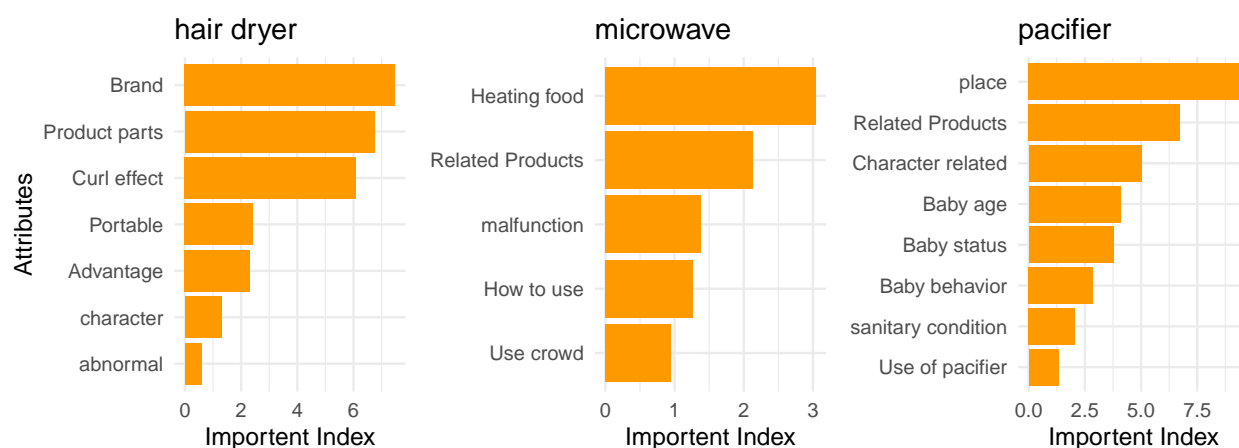


Figure 2: The most important attributes in the reviews of three commodities

- **Please keep an eye on improving product ratings in the early stages of market entry.**

In the early stages of product release, customers cannot get product information from other channels, and they are bound to pay close attention to customer reviews. Through analysis, we found that the a growth in the number of 4 and 5 star ratings will increase the proportion of positive emotion words in the four categories of customer reviews, as shown in the figure.

In the structural equation modeling, we find that an increase in positive sentiment words will significantly improve product reputation, which indicates the future success.

- **Enhance specific products attributes to meet the customers need.**

We find out the most important words about the three types of products through our model, indicating that customers are very concerned about these aspects. Doing well in these aspects will greatly improve the competitiveness of the product.

It can be seen from the figure that the first three concerns are as follows:

Hair dryer:

- brand: Customers are very concerned about the brand, so they need to enhance brand image and brand awareness
- product part: Focus on the quality of parts and improve product quality
- curl effect: Actively develop new technologies

microwave:

- heating food: Provides heating modes for different foods, making it easy for users to heat various foods
- related products :Co-organize promotional activities with manufacturers of related products such as refrigerators
- malfunction: Improve product quality and reduce the probability of failure

pacifier:

- place: Places for use of pacifiers: Customers are very concerned about whether pacifiers can be used by children such as newborn babies and teeth changing stages.

- related products: Can be sold with diapers and other products
- character related: Breastfeeding women pay close attention to the condition of pacifiers because they are about to wean their children. This can improve the quality of pacifiers and place advertisements in specific groups.

1 Introduction

1.1 Background

In recent years, online shopping has become an indispensable consumption mode in life. Online marketplaces such as Amazon have invested a considerable amount of resources in designing and improving the store interface, since online store interaction plays a significant role in the selection process. Heijden & Verhagen (n.d.)

Customers use online review systems to evaluate goods or services sold by merchants. The move towards a read-write Web and the ability to provide more relevant information to consumers on online marketplaces has been also demonstrated to influence the product choice process in electronic markets. Godes, Ofek, & Sarvary (n.d.)

Online review reflects the needs and preferences of customers and influences their purchase intentions. Due to the rapid development of online shopping, the number of online reviews has surged. How to get useful information from online reviews to help businesses make optimal decisions has become more and more important. For example, reviews closer to consensus may be considered more helpful by potential consumers than those exhibiting extremes of opinion. Sergeantanis & Economopoulos (n.d.) Nikolaos Korfiatis explored the qualitative or textual characteristics of online consumer reviews. Korfiatis & Poulos (n.d.)

This paper modeled and analyzed the variables in the review system, clarified the relationships between the variables, and screened out useful information, providing effective advice for Sunshine Company to enter the market.

1.2 Our work

Explore the relationship between variables through regression models.

Use regression models to study the impact of CLI and star rating on HR.

Use the \$ TF-DIF \$ model and the customer utility function to calculate the true utility of the customer and use the test set to check the validity of the model.

Use a hierarchical model to predict the product's reputation trend and find potentially successful and failing products.

Make recommendations for Sunshine Company based on the above analysis,

2 Notation and Assumptions

To establish our models, we made the following assumptions:

- **We assume that there is no difference between customers and they share the same preferences for the same characteristics of the product.** This allows us to extract user preferences based entirely on the dataset. While this assumption is not entirely realistic, we cannot consider the educational level of customers, different age

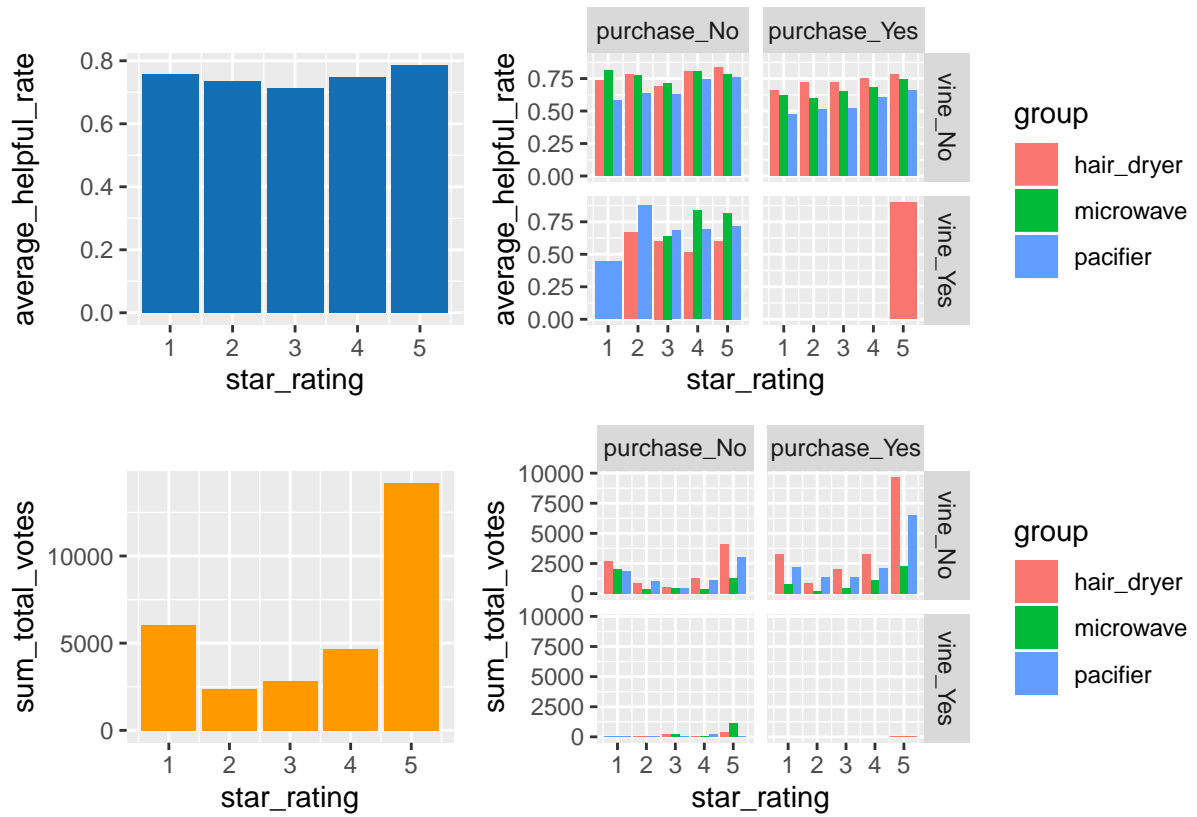


Figure 3: Relationship between star rating and helpful votes rates or total votes

groups, and other factors that cause differences in customer awareness and understanding of products in the data provided.

- **To further reduce external uncertainties, we do not consider the impact of changes in the market of alternative products and the macroeconomic cycle on these three types of commodities.** This allows our models to better study the characteristics of different products. Furthermore, this assumption simplifies our analysis and allows us to focus more on exploring the internal features of the dataset.

3 Relationship between variables

3.1 Descriptive statistics

From the figure 1 we can see: Because the review rating uses a discrete scale range from 1 to 5 star rating, many customers who want to express their indifference will choose 3 star rating (middle option). For example, when they analyzed online reviews of books on Amazon Lee, Trimi, & Yang (n.d.), they found that extreme ratings (1 star rating or 5 star rating) is more likely to resonate than the indifference rating (about 3 star rating). In other words, reviews that has more emotionally polar (only referring to either positive or negative aspects) are considered more helpful and have a greater impact on customers. This is also confirmed by the number of votes of each star rating. We found that although the star ratings are not much different in recognition (usefulness in voting), but in terms of the total number of votes,

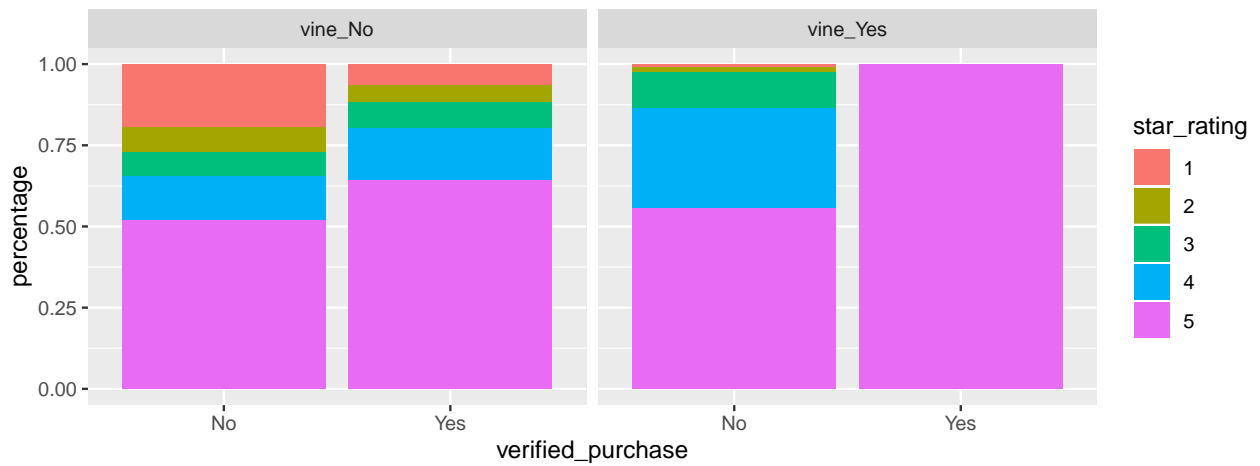


Figure 4: Percentage of star rating in different categories of customers

1 star rating and 5 star rating are obviously more, which shows that customers pay more attention to extreme reviews.

From the perspective of vine, vine's purchase of products after trial is almost all 5 star ratings, while the proportion of 5 star ratings in non-vine's are significantly lower than vine. Because customers are invited to become Amazon Vine Voices based on the trust that they have earned in the Amazon community for writing accurate and insightful reviews. Therefore, the 5 star rating will be more.

3.2 Reviews that triggered by a specific star rating

3.2.1 Research ideas

Actually, in Amazon Mall, most customers browse through the rating results of previous customers before they comment, and then write their own reviews. Each customer has a different feeling on the same product, and the rating will also be different. When the customer browses the previous reviews, it is likely to be affected by the number of ratings of 5 levels. Therefore, we quantify this impact by the proportion of 5 ratings of all previous reviews that the customer viewed. So, how to measure the customer's review characteristics? In the text analysis, the emotion of the text is a commonly used text feature Ho & E. (n.d.), so we use the total emotion of review body to indicate customer's review characteristics. Next, we use the function fitting to explore the relationship between the two variables.

3.2.2 Model construction and solution

As the model of three types of products are in exactly the same way, we use the hair dryer as an example.

Processing *star rating* results of previous customers:

- **Classify the products of the hair dryer.** At Amazon Mall, a product may be sold in different stores, and customers can only comment under the store where they buy a

product, so we define the same model products sold in the same store as one type. We classify the products in product id.

- **Screen every category of hair dryer.** There are 538 types of hair dryers. Since the customer who first reviewed did not have any previous scoring results as a reference, we first selected products with at least 2 reviews.
- **Calculate the rating results of the 2^{nd} to n^{th} customers browsing for each category of hair dryer.** In one category of hair dryer products, the i^{th} customer is the research object. Calculate the number and proportion of different ratings in $i - 1$ previous reviews.

The 1 star rating ratio is as following, and other star rating ratio calculations are the same:

$$HairRate1_i = \frac{HairSum1_i}{i - 1} \quad (1)$$

Gao Baojun found that in a system that does not classify ratings, because of fewer negative reviews and higher search costs, customers are more likely to see positive reviews, so positive reviews have a greater impact on customers. Because Amazon does not directly categorize its ratings, and the number of 4 and 5 star ratings of hair dryer are much higher than other ratings. At this time, searching for low star ratings from many reviews will spent a lot of time, so the impact of high ratings is significant. We add the 4 and 5 star rating ratio in the customer reviews as the praise rate rate. The praise rate of the i^{th} customer is:

$$rate_i = HairRate4_i + HairRate5_i \quad (2)$$

- Then calculate the total emotion in each review of 2^{nd} to n^{th} customers in each category.
- Finally, we use function fitting to find the relationship between the two variables. A linear regression was carried out with R. The difference praise rate rate of the previous customer's rating per customer viewing is an independent variable and the total emotion of that customer's review is a dependent variable.

$$\log(Sum_i) = Rate_i + C \quad (3)$$

The fitting results of the three products are shown in the plot.¹

The trend of the total emotion changes of the three products is different, due to the different types of products. In online shopping, products are classified as search-type products and experiential products. Search-type products can obtain information about quality before using. Experiential products must be used before assessing product quality. The impact of different product types of ratings varies. Pacifier is an experiential product. It must be used before evaluating product quality, so it is subjective and customers are more likely to be affected by other customers' ratings. As the praise rate increases, customers are significant influenced and more willing to express their feelings about a product. Microwave is a search-type product. Customers can accurately evaluate the quality through specific parameters before purchasing. It is so objective that customers are not easily affected by other customers'

¹See Appendix for specific data.

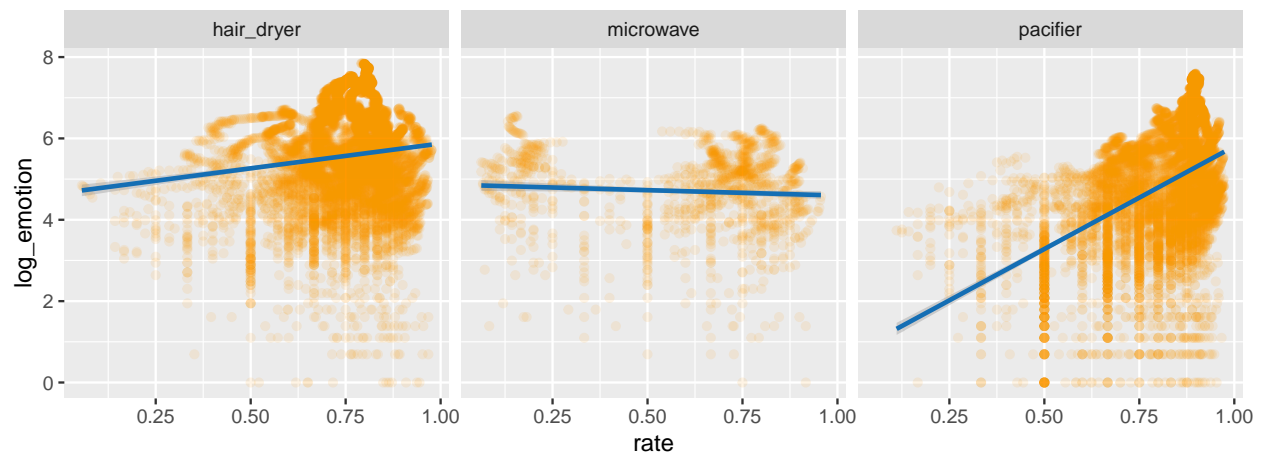


Figure 5: Trend of total emotion with praise rate

Table 1: Regression of Y on X for three products

	<i>Dependent variable:</i>		
	log_emotion		
	(1)	(2)	(3)
rate	1.222*** (0.102)	−0.262** (0.114)	5.048*** (0.108)
Constant	4.654*** (0.078)	4.859*** (0.075)	0.757*** (0.087)
Observations	9,978	1,303	8,651
R ²	0.014	0.004	0.202
Adjusted R ²	0.014	0.003	0.202
Residual Std. Error	1.230 (df = 9976)	1.025 (df = 1301)	1.383 (df = 8649)
F Statistic	143.482***	5.298**	2,192.962***

Note:

*p<0.1; **p<0.05; ***p<0.01

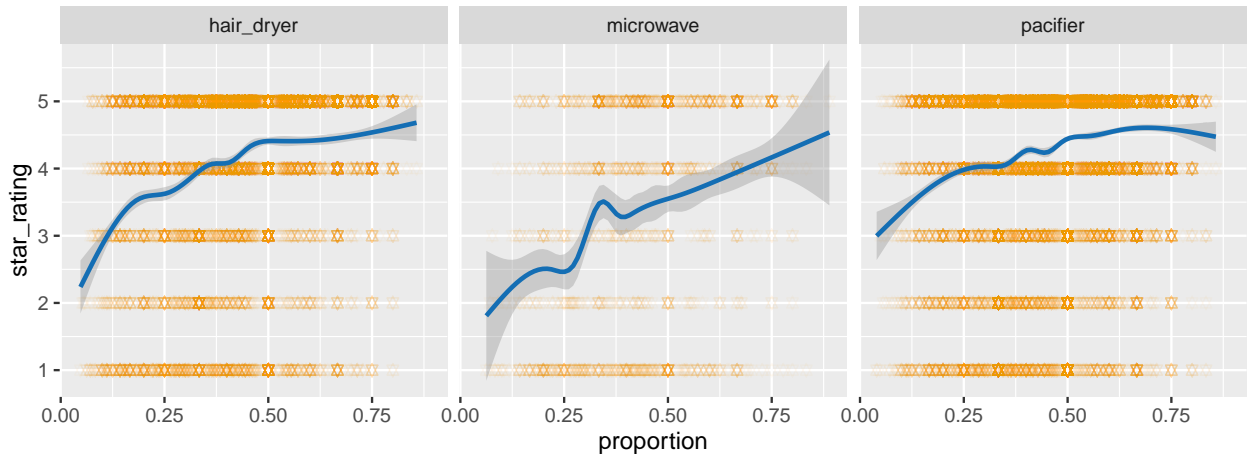


Figure 6: relationship between star rating and proportion of specific quality descriptors

ratings, so it is almost a horizontal straight line. The type of the hair dryer is between the microwave and pacifier, so the praise rate of other customers will have a certain impact on the total emotion of the review. Of course, the impact is lower than that of pacifier.

3.3 Relationship between *Star rating* and specific emotion words

We explored whether descriptive words of specific nature in the review would be highly relevant to the rating. First of all, we will divide comments into words and find the content of word segmentation. And then we calculate the sum of joy, trust, surprise and anticipation emotions in each review. Furthermore, the ratio of the number of these four types of emotional words in each review to the total number of emotional words in that review is taken as independent variable, and the rating of each comment is taken as dependent variable. We draw figures of three data sets.

As shown in the figure, the smooth fitting curve has a clear trend. As the proportion of these four emotions increased, so did the rating. These four types of emotions are positive emotions, so the more such emotions in the reviews indicate that customers are satisfied with the product, and ratings are higher too.

We then perform linear regression on the three datasets respectively. As we can see, since the rating is integer, the R^2 of the model is quite small. But the coefficients obtained by fitting are very significant, and passed the F-test test.

4 Identify useful reviews

4.1 Research ideas

A text feature analysis of the review content is not limited to ratings and text length. Usually when analyzing reviews, it is difficult to analyze from the content, but we introduced CLI, total emotion, and product keyword analysis, and conducted a full analysis about text readability, emotional characteristics, and text keywords.

Table 2: Regression of star rating on proportion of specific quality descriptors for three products

	<i>Dependent variable:</i>		
	star_rating		
	(1)	(2)	(3)
proportion	2.108*** (0.087)	2.948*** (0.279)	1.615*** (0.063)
Constant	3.202*** (0.041)	2.061*** (0.131)	3.553*** (0.031)
Observations	7,988	1,195	14,540
R ²	0.069	0.086	0.043
Adjusted R ²	0.069	0.085	0.043
Residual Std. Error	1.246 (df = 7986)	1.594 (df = 1193)	1.149 (df = 14538)
F Statistic	590.479***	112.051***	657.771***

Note:

*p<0.1; **p<0.05; ***p<0.01

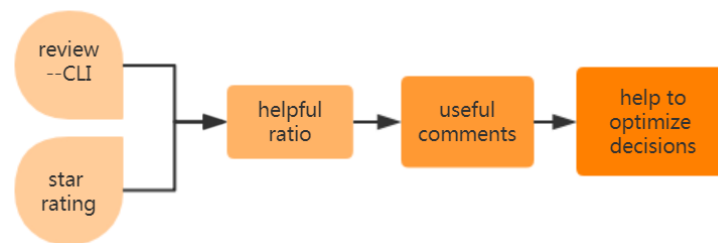


Figure 7: Identify useful comments

There is a lot of information in the online review system of online store, and the online store owners often want to analyze the information in the online review system to know more about the customers and adjust their products to promote the increase of sales volume. However, many reviews have no impact on customers, and we call it useless evaluation. Some reviews will have a great impact on customers, and then affect sales volume, called useful evaluation.

Only by finding useful evaluation and analyzing it can Sunshine Company accurately grasp customer preferences and make correct adjustment directions in time. When new products enter the market, reviews are published recently and need some time to be viewed by people, so their HRs are not accurate enough. However, in order to improve the product as soon as possible, without knowing the accurate HR, we should find useful comments accurately by analyzing some characteristics of the reviews. In this way, Sunshine Company can analyze and make optimal decisions as soon as possible.

4.2 Model establishing

In order to measure the usefulness of a review, we measure it by the helpful votes ratio of goods :

$$HR = \frac{HelpfulVotes}{TotalVotes} \quad (4)$$

Informativeness is a feature of reviews. We use the CLI to measure the amount of information contained in each review and accurately measure the information status of each review. The Coleman–Liau Index Coleman & Liau (n.d.) has been developed specifically for machine-based scoring. The following formula describes the Coleman–Liau index:

$$CLI = 5.89 \times \frac{characters}{words} + \frac{sentences}{word} - 15.8 \quad (5)$$

The CLI indicates the readability of the reviews. The smaller the CLI, the simpler and easier to understand the review, and the higher the CLI, the more complex the review. Combined with star ratings, through the relationship between these variables, useful reviews can be effectively identified to help Sunshine Company to make judgments. We assume that CLI and rating are not simple linear relation. So we introduce quadratic term in the regression model.

$$HR = \beta_1 CLI + \beta_2 CLI^2 + \beta_3 StarRating + \beta_4 StarRating^2 + C \quad (6)$$

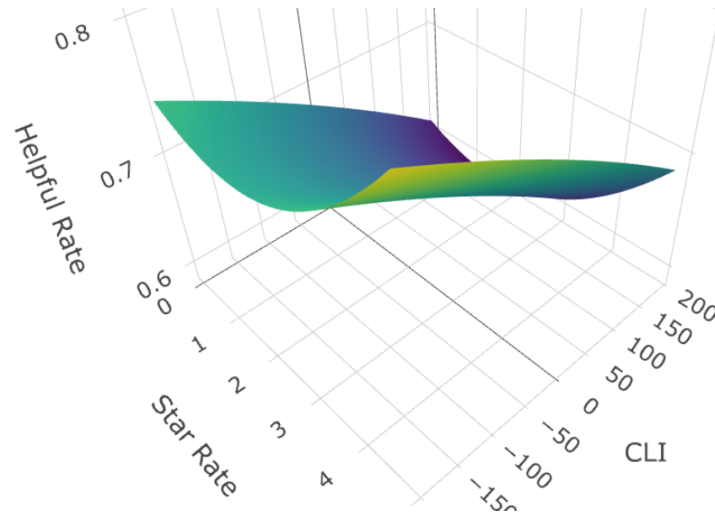


Figure 8: Relationship between StarRating, CLI and HR

4.3 Model Result

Table 3: Regression Coefficients for the regression model

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.68	0.03	24.75	0.00
CLI	0.00	0.00	-2.45	0.01
CLI_square	0.00	0.00	-3.40	0.00
star_rating	-0.03	0.02	-1.48	0.14
star_rating_square	0.01	0.00	3.03	0.00

The regression results are shown in the table. The quadratic terms of the CLI and StarRating's coefficients pass the hypothesis test and were significantly different from 0, indicating that the CLI and StarRating have a significant impact on HR. The absolute value of the quadratic term coefficient of StarRating is larger than that of CLI.

With the change of StarRating, HR will change more obviously. It can also be seen from the figure that the HR is larger at 1 and 5 star rating and smaller at 3 star rating, so the reviews corresponding to the 1 and 5 star ratings are more likely to be useful reviews.

With the increase of the CLI, the HR increases first and then decreases. When the review is too simple, it contains less information and the review is not very useful. When the review contains too complicated information, it is difficult for customers to process, which will also reduce the usefulness of the review. But the coefficient of the quadratic terms of the CLI is too small, the change trend is slow. Therefore, it is only when the reviews are extremely simple or extremely complex that they can lead to a significant decrease in the usefulness of the reviews.

5 GD, SGD and Neural Network

5.1 Logit model based on GD and SGD optimization methods

5.1.1 Model establishing

Suppose we have independent training samples $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, $y = \{0, 1\}$. The probability that each observed sample (x_i, y_i) occurs is:

$$P(y_i, x_i) = P(y_i = 1|x_i)^{y_i} (1 - P(y_i = 1|x_i))^{1-y_i} \quad (7)$$

Then the likelihood function for our entire sample set, in which there are n independent samples:

$$L(\theta) = \prod P(y_i = 1|x_i)^{y_i} (1 - P(y_i = 1|x_i))^{1-y_i} \quad (8)$$

Then, the maximum likelihood function is called **cost function**.²

5.1.2 Model solving

The derivative of $L(\theta)$ with respect to θ is:

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n y_i x_i - \sum_{i=1}^n \frac{e^{\theta^T x_i}}{1 + e^{\theta^T x_i}} x_i = \sum_{i=1}^n (y_i - \sigma(\theta^T x_i)) x_i \quad (9)$$

For this function, we can **not** easily find the derivatives. So we use the properties of the gradient descent method Baobin & Yuxia (2011) , then:

$$\theta^{t+1} = \theta^t - \alpha \frac{\partial L(\theta)}{\partial \theta} = \theta^t - \alpha \sum_{i=1}^n (y_i - \sigma(\theta^T x_i)) x_j \quad (10)$$

We normalize each variable before training. Then, we set the learning rate to be gradually decreasing for training.

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n y_i x_i - \sum_{i=1}^n \frac{e^{\theta^T x_i}}{1 + e^{\theta^T x_i}} x_i = \sum_{i=1}^n (y_i - \sigma(\theta^T x_i)) x_i \quad (11)$$

²The key of applying maximum likelihood method is to find the value θ of the coefficient that maximizes the likelihood function in the model.

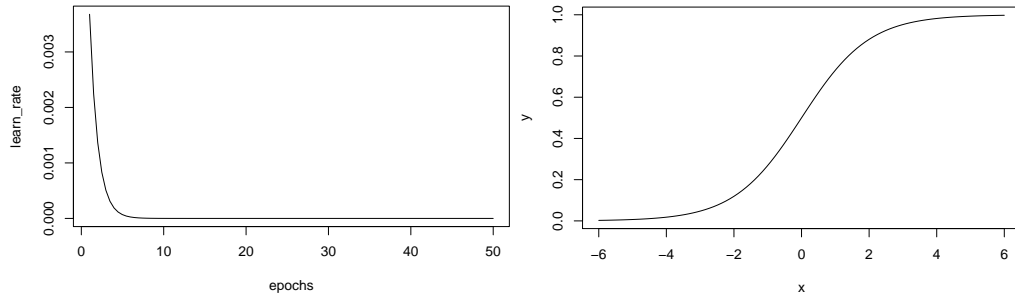


Figure 9: Decreasing learning rate function and Sigmoid function

5.2 Neural network model based on SGD optimization method

5.2.1 Model establishing

5.2.1.1 Neurons

A neuron is the basic unit of a neural network. Chua & Yang (1988) The neuron first takes input and then performs some mathematical operations before producing an output. Zhenquan & Dongsheng (1990)

For a two-input neuron, the output steps are: First multiply the two inputs by the weights. Add the two results and add an bias. Finally process them through the activation function to get the output.

$$y = f(x_1 \odot w_1 + x_2 \odot w_2 + b) \quad (12)$$

We choose the sigmoid function as the activation function of the neural network.

$$S(x) = \frac{1}{1 + e^{-x}} \quad (13)$$

$$S'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = S(x)(1 - S(x)) \quad (14)$$

5.2.1.2 Neural Networks

Considering the limitation of the computing resources, we build a simple neural network. It contains 2 inputs, a hidden layer with 2 neurons (h1 and h2) , and an output layer (o1) with 1 neuron. Pingfan (2005)

5.2.1.3 Feedforward and Loss

The process of passing the input of a neuron forward to obtain the output is called feedforward.

Before training a neural network, we need to have a standard definition so that we can make improvements. We use the mean square error (MSE) to define the loss:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{true} - y_{pred}) \quad (15)$$

We only use one sample at a time during training, so the loss function is:

$$\begin{aligned} MSE &= \frac{1}{1} \sum_{i=1}^1 (y_{true} - y_{pred})^2 \\ &= (y_{true} - y_{pred})^2 \\ &= (1 - y_{pred})^2 \end{aligned} \quad (16)$$

5.2.2 Model solving

5.2.2.1 Training neural network

The process of training a neural network is to minimize the loss. The lower the loss, the better the prediction result.

Since the predicted value is calculated by a series of network weights and biases, the loss function is actually a multivariate function depending on multiple weights and biases:

$$L(w_1, w_2, w_3, w_4, w_5, w_6, b_1, b_2, b_3) \quad (17)$$

Derivative chain rule: (take w_1 as an example)

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y_{pred}} * \frac{\partial y_{pred}}{\partial h_1} * \frac{\partial h_1}{\partial w_1} \quad (18)$$

This system using backward calculation of partial derivatives is called reverse conduction.

5.2.2.2 SGD optimization method

We use the stochastic gradient descent (SGD) optimization algorithm Gongpeng, Meng, & Changyong (2018) based on convolutional neural networks to gradually optimize the weight w and bias b so that the loss function will be slowly decreased, thereby improving our neural network. Take w_1 as an example:

$$w_1 \leftarrow w_1 - \eta \frac{\partial L}{\partial w_1} \quad (19)$$

5.2.2.3 Initialization of weight

The initialization of the weights and biases of the nodes in the neural network follows the standard normal distribution.

$$Weight_i, Bias_i \sim N(0, 1) \quad (20)$$

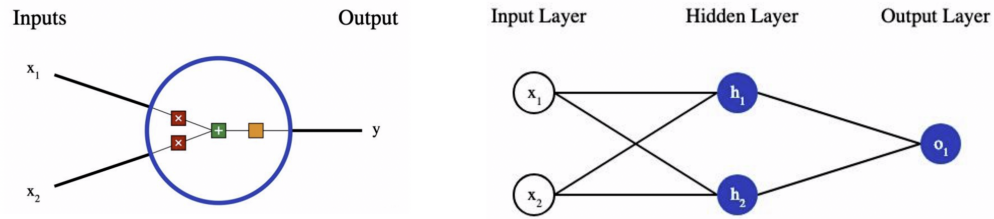


Figure 10: Schematic diagram of the built neural network

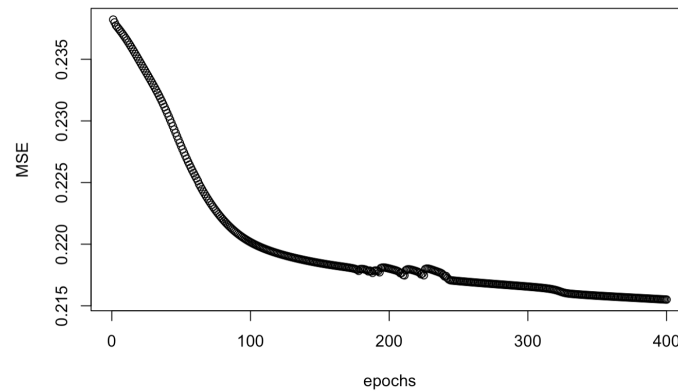


Figure 11: Variation of MSE with the number of iterations

5.2.3 Model results

From the results, as the number of iterations increases, the mean square error first decreases rapidly, and then stabilizes.

5.2.4 Cross validation

We use *total emotion* and *square of star rating* to predict *helpful rate*.

Using systematic sampling, we divided the data set into three groups for cross-validation.³

Table 4: Time Consuming (s)

Method	Cross Validation 1	Cross Validation 2	Cross Validation 3
logit_GD	9.47	9.41	8.39
logit_SGD	4.50	4.46	4.11
Neural Network	170.13	144.55	152.11

Table 5: Correct Rate

Method	Cross Validation 1	Cross Validation 2	Cross Validation 3
logit_GD	0.72	0.73	0.73
logit_SGD	0.72	0.72	0.72
Neural Network	0.74	0.72	0.74

³Time consuming differences due to computer hardware levels and temperature conditions. For reference only

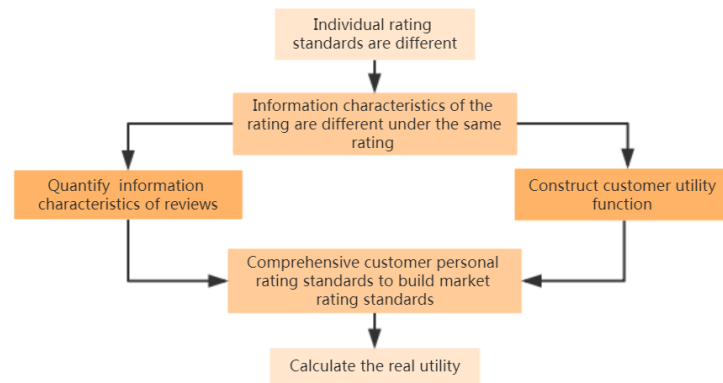


Figure 12: research idea of tracking the real utility

6 Track the real utility of customers

6.1 Research ideas

Rating reflects the utility of the customer. However, rating standards vary from person to person. For example, some people think that 3 stars represent average satisfaction, while others think that 3 stars can show their strong dissatisfaction. It can be seen that customers with different utility share the same rating based on different scoring criteria, and the information indicated by ratings and reviews is not completely consistent. Rating may cause deviations in the utility measurement due to the different personal rating standards of the customers and the discrete scale of scoring. Therefore, by establishing a suitable model and using the information in reviews to calculate their real utility, Sunshine Company can analyze it. Our model's principle is to quantify the information contained in the review and return to a score. And with the multiple regression, we can comprehensively consider the rating standards of all customers and build a market preference patterns. And then recalculate the corresponding score of each review. Finally, we match score with customer rating to ensure the correctness of our model. And the score reflect customers true attitude toward the product.

We will first build a model that extracts product features based on customer reviews of the product, and then build a market utility measurement model based on the extracted features.

6.2 Feature extraction

In order to take into account all the factors affecting the utility in the review, we use tf-idf model to extract the features of the review. It's a numerical statistic that is intended to reflect how important a word is to a review in our dataset. And then we can extract the feature matrix.

$Feature \in \mathbb{R}^{m \times n}$ can be summarized as follows:

$$\text{tfidf}(t, d, D) = TF \cdot IDF \cdot I_{d \in nrc} \quad (21)$$

Now, we will look at what each of these terms represent.

TF is an abbreviation for "Term Frequency". TF indicates how often t appears in d . Here we assume that "review_body" is the detailed discription of "review_head", thus we extract review features only in "review_body". Meanwhile, the "review" mentioned in the following refers to "review_body" by default.

A larger TF means more occurences of t in the review d .

IDF is the abbreviation for "Inverse document frequency". It's a measure of how much information the word provides in a review. We calculate IDF as:

$$IDF_{i,j} = \log \frac{|D|}{|\{j, t_i \in j\}|} \quad (22)$$

D shows a set including all reviews in a data set. N indicates the total number of reviews included by D . From the formula we can see that when a word in one review appears more often in other reviews, IDF will become smaller, which means that the word is more common.

$TF \cdot IDF$ measures the weight of each word in a review, so that each reviews can be represented by a vector.

This is an illustrative function. $I_{d \in nrc}$ is given by:

$$I_{d \in nrc} = \begin{cases} 1 & , d \in nrc \\ 0 & , d \notin nrc \end{cases} \quad (23)$$

nrc refers to "NRC Emotion Lexicon". It's a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Here we assume that sentimental words are the key factor that affects the customer's rating, which will help us avoid the interference of meaningless words on review features. It makes sense because the evaluation of the product can not avoid using the emotional words such as "good" or "bad".

In fact, as we solve our model, we find that this assumption does not result in the component elements of the vector representing the reviews features to be all 0, which means that all effective reviews carry at least one word in nrc . We also verified this in the other two data sets *hair_dryer.tsv* and *microwave.tsv*

6.3 Construction of Customer Utility Function

We believe that the utility of a product depends on the customer's personal preference and the characteristics of the product. Thus, our model for customer utility is summarized as:

$$u(s, c) = score(Content(s), ContentBasedProfile(c)) \quad (24)$$

$Content(s)$ is a matrix describing product characteristics based on content. That is $Freature \in \mathbb{R}^{m \times n}$, calculated in the previous model. $ContentBasedProfile(c)$ is a measure of customer preference. Based on our customer homogeneity assumption that customer preference

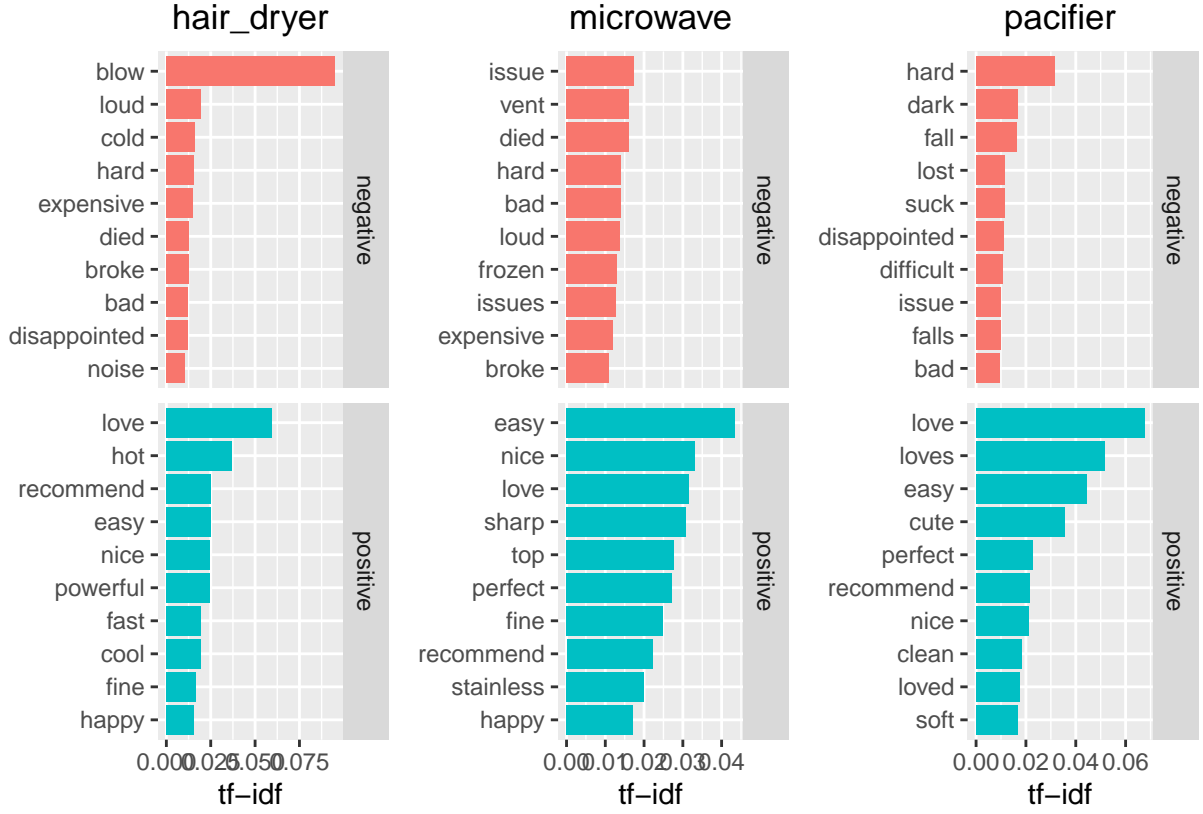


Figure 13: The most important words in the review of three products

patterns represents market preference patterns. So we define the vector representing market preference as $Preference \in \mathbb{R}^{n \times 1}$. Customer utility can be further expressed as:

$$U = Feature \cdot Preference \quad (25)$$

We consider the star rating given by customers in the data set as a reflection of the customer's utility. The significance of this is to let our model learn the characteristics of different ratings from reviews. In this way, both U and $Feature$ are known, and the customer's customer preference, which is also referred to the market preference, can be quantified by the least square method.

6.4 Model Result

We split the data into training and testing sets according to a 4: 1 ratio. Take "pacifier.tsv" as an example. Before extracting the reviews feature, we clean the data set by dropping two rows with null value. Then we calculate $Feature$ matrix. The shape of the $Feature$ matrix is (18937, 16413), of which there are 2366 words in nrc . Delete the rows with all 0. The shape of the matrix has not changed. This validates our hypothesis that sentimental words vocabulary is the key factor affecting ratings. The key words we have selected through feature extraction are shown in the figure.

Next, we use the least squares regression on the training set to obtain the market preference, and the value of RMSE is 1.27.

To further explore the effectiveness of the model, we predict the real utility of customers based on the characteristics of the reviews on the test set, and calculate the matching degree of the model on the test set. As the star rating set is considered to be a discrete variable, we think the error is acceptable within 1 star. The matching degree is 85.35% on test set.

In dataset *hair_dryer* and *microwave*, it is also found that the *Feature* matrix has no row elements that are all 0. The RMSE of the model is 1.50 and 2.14, and the matching degrees are 85.35% and 77.09%, respectively.

The real utility obtained by our model is close to the ratings of most customers, which shows that the market rating standards constructed by this model are indeed a comprehensive processing and averaging result of all customer rating standards, and can represent the overall rating standards of customers. At the same time, about 20% of customers don't match the real utility, indicating that this customer's rating standard is seriously deviating from the market rating standard, and the rating is inaccurate. It better reflects the relationship between product characteristics and the user's real utility.

7 Product Reputation and Potential Successful Product

7.1 Model Establishing

We want to measure reputation, which is difficult since it can not be seen or captured using a single method. We decide to measure this latent variable "reputation" by two new variables. These new variables have the advantage of being more clearly defined. And then, we add another two factors to be the regressors.

7.2 Model Solving

Table 6: Latent variables and Measurements

latent	measurement	explanation
reputation	dif_star	Increase of mean star rating every year
	dif_review	Increase of number of reviews every year
review	emotion	Total emotion
	CLI	Readability of text
	positive	Number of positive emotion words
mark	star	Actual ratings in the dataset
	score	Quantitative real utility from reviews
type of customer	vine	Yes / No Amazon vine member
	verified	Yes / No Verified purchase

Using factor analysis, we try to solve a problem: there are variables that we want to measure, but we can not directly observe them, which are called latent variables. These

can typically be rather complex variables, such as quality of life or general mental ability. Structural equation model (SEM) is a tool that can be used to try to measure these latent variables. We try to explain the variation in the unexplained, latent variables (factors), using observed, correlated variables.

Measurement part: - Factor model for the endogenous variables - Factor model for the exogenous variables

Structural part: - Model implied by theory

Factor model for the endogenous variables - y as a function of latent variables η :

$$y_q = \alpha_{q1\eta_1}^y + \alpha_{q2\eta_2}^y + \dots \alpha_{qs\eta_s}^y + \epsilon_q$$

Factor model for the endogenous variables - x as a function of latent variables ξ :

$$y_q = \alpha_{p1\xi_1}^x + \alpha_{p2\xi_2}^x + \dots \alpha_{ps\xi_s}^x + \epsilon_p$$

Interpretation for factor loadings - standardize $V(y_i) = 1$ and $V(\eta) = 1$, then the OLS estimator

$$\hat{\alpha} = (Cov(y, \eta)/V(\eta)) = \rho_{y,\eta} \quad (26)$$

and

$$V(\eta) = \hat{\alpha}^2 V(\eta) + V(\eta) \quad (27)$$

$$1 = \hat{\alpha}^2 + V(\eta) \quad (28)$$

$\hat{\alpha}^2$ is the proportion of the variance of y originating from the factor η and $1 - \hat{\alpha}^2$ from the error.

The structural part consists of a regression model for each dependent latent variable as a function of: - Other dependent latent variables, endogenous, η - Latent exogenous variables, ξ

$$\eta_s = \beta_{s1}\eta_1 + \beta_{s2}\eta_2 + \dots + \beta_{ss-1}\eta_{s-1} + \gamma_{s1}\xi_1 + \gamma_{s2}\xi_2 + \gamma_{sr}\xi_r + \zeta_s \quad (29)$$

It can be seen from the figure that the circle represents the factor and the square represents the measurement index. The number on the line is the regression coefficient. (The number is 0 because the coefficient is small and not shown intact.) . It can be seen from the figure that the more positive emotion words in the review, the higher the product's reputation. The higher the customer's true utility in the rating, the higher the customer's satisfaction with the product and the higher the product's reputation. Among the customer types, the more customers of the verified type indicate that many customers are willing to purchase the product, which is a recognition of the product quality and an increase in reduction. The decrease of positive sentiment words in the product reviews, the decrease of customer's true utility, and the decrease number of customers of the verified type indicate that the reduction of such products will gradually decrease.

7.3 Product Forecast

From the above analysis, it can be known that the CLI in each type of product reviews, the number of positive sentiment words, the true utility, and proportion of customers belonging

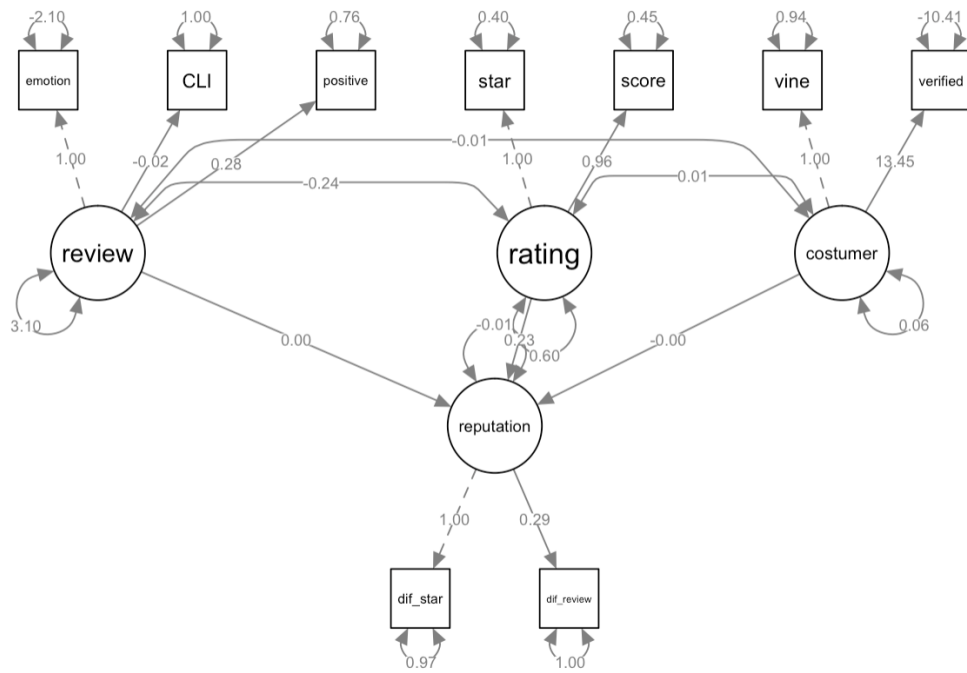


Figure 14: Structrual Equation Model Outcome

to the verified category have a significant impact on the product 's reputation .The product 's reputation continues to increase.Have development potential. It shows that this product has development potential and is a potential successful product. Otherwise, it is a potential failed product. Therefore, attention should be paid to the changes in these four indicators. It is predicted whether the number of reviews of the product will increase or decrease in the next year, and based on this, it is judged whether the product's reputation is increasing or decreasing. At the same time, we believe that products that increase in the next year can be regarded as potential successful products, and products that decrease in the next year can be regarded as potential failure products.

8 Strengths and Weaknessess

8.1 Strengths

1. **The analysis of the review content is not limited to ratings and text length.** We introduced CLI, total emotion, and product keyword analysis, and conducted a full analysis in terms of text readability, emotional characteristics, and text keywords.
2. **The models are rich and the effects are significant.** We not only introduce basic models such as regression, structural equation models, etc., but also construct our own models to calculate the real utility, and analyze the data more accurately from multiple angles.

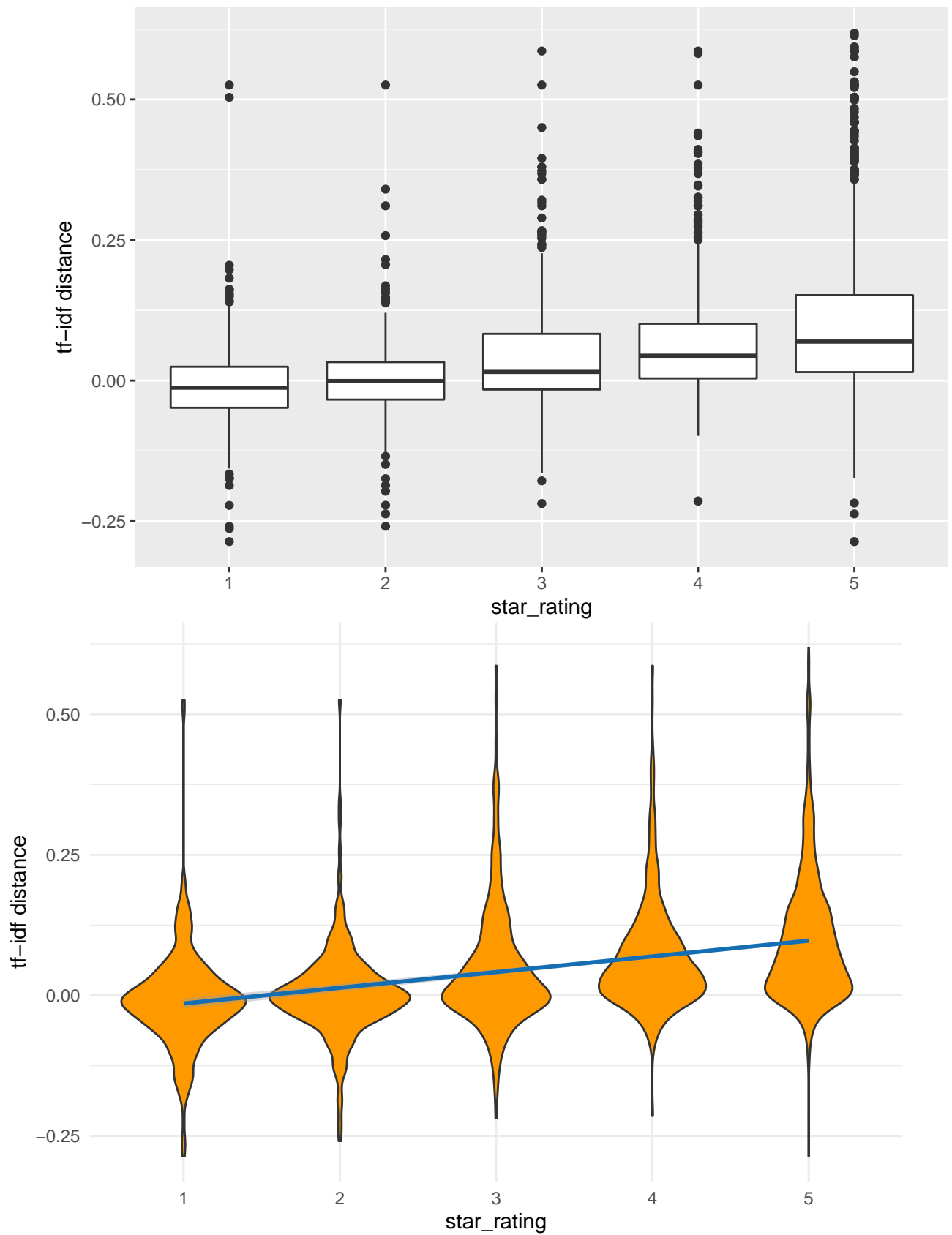
8.2 Weaknessess

1. In the analysis of the data, the control variables of the regression were not considered comprehensively. Multiple regression models should be tried to select better results.

References

- Baobin, W., & Yuxia, W. 2011. Some properties of stochastic gradient descent. *Journal of Mathematics*, 31(6): 1041–1044.
- Chua, L. O., & Yang, L. 1988. Cellular neural networks:Theory. In *IEEE international workshop on cellular neural networks & their applications*.
- Coleman, M., & Liao, T. L. n.d. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2): 283–284.
- Godes, D., Ofek, E., & Sarvary, M. n.d. Products vs. Advertising: Media competition and the relative source of firm profits. *Social Science Electronic Publishing*.
- Gongpeng, W., Meng, D., & Changyong, N. 2018. Stochastic gradient descent algorithm based on convolutional neural network. *Computer Engineering and Design*, 39(2): 441–445.
- Heijden, H. V. D., & Verhagen, T. n.d. Online store image: Conceptual foundations and empirical measurement. *Information & Management*, 41(5): 609–617.
- Ho, & E., D. G. n.d. Systemic text analysis in the esl writing classroom: Does it work? *Relc Journal A Journal of Language Teaching & Research*, 40(3): 333–359.
- Korfatis, N., & Poulos, M. n.d. Using online consumer reviews as a source for demographic recommendations: A case study using online travel reviews. *Expert Systems with Applications*, 40(14): 5507–5515.
- Lee, S. G., Trimi, S., & Yang, C.-G. n.d. Perceived usefulness factors of online reviews: A study of amazon.com. *Journal of Computer Information Systems*, 1–9.
- Pingfan, Y. 2005. *Artificial neural networks and simulated evolutionary computation (2nd edition)*.
- Sergentanis, T. N., & Economopoulos, K. P. n.d. *Re: Jiang et al.: meta-analysis of association between tp53 arg72pro polymorphism and bladder cancer risk (urology 2010;76:765)*, 77(1): 0–260.
- Zhenquan, Z., & Dongsheng, W. 1990. Neural networks and neural computers: Lecture 2 learning algorithms for neural networks. *Application of Electronic Technology*, (5): 38–41.

Appendix



lavaan 0.6-3 ended normally after 553 iterations

```

##
## Optimization method NLMINB
## Number of free parameters 24
##
## Number of observations 18113
##
## Estimator ML
## Model Fit Test Statistic 1815.000
## Degrees of freedom 21
## P-value (Chi-square) 0.000
##
## Parameter Estimates:
##
## Information Expected
## Information saturated (h1) model Structured
## Standard Errors Standard
##
## Latent Variables:
## Estimate Std.Err z-value P(>|z|)
## reputation =~
## dif_star 1.000
## dif_review 0.293 0.050 5.872 0.000
## review =~
## emotion 1.000
## CLI -0.022 0.002 -8.823 0.000
## positive 0.280 0.020 13.913 0.000
## rating =~
## star 1.000
## score 0.957 0.017 57.816 0.000
## costumer =~
## vine 1.000
## verified 13.447 8.826 1.524 0.128
##
## Regressions:
## Estimate Std.Err z-value P(>|z|)
## reputation ~
## review 0.002 0.001 1.707 0.088
## rating 0.235 0.012 20.381 0.000
## costumer -0.002 0.005 -0.344 0.731
##
## Covariances:
## Estimate Std.Err z-value P(>|z|)
## review ~~
## rating -0.238 0.007 -34.295 0.000
## costumer -0.006 0.004 -1.428 0.153
## rating ~~
## costumer 0.008 0.006 1.428 0.153
##

```

Variances:

##		Estimate	Std.Err	z-value	P(> z)
##	.dif_star	0.973	0.027	35.933	0.000
##	.dif_review	0.998	0.011	93.231	0.000
##	.emotion	-2.098	0.218	-9.608	0.000
##	.CLI	0.998	0.010	95.277	0.000
##	.positive	0.756	0.019	40.168	0.000
##	.star	0.404	0.010	39.221	0.000
##	.score	0.453	0.010	46.142	0.000
##	.vine	0.937	0.042	22.052	0.000
##	.verified	-10.407	7.474	-1.392	0.164
##	.reputation	-0.005	0.025	-0.208	0.836
##	review	3.098	0.216	14.318	0.000
##	rating	0.596	0.013	44.419	0.000
##	costumer	0.063	0.041	1.520	0.128