For office use only	Team Control Number	For office use only
T1 _____	**2013133**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2020
MCM/ICM
Summary Sheet

E-commerce Review Data Mining Based on Structured and Unstructured Data

Summary

Reviews can effectively provide customers and businesses with effective information,but at the same time they can also cause information overload.Based on this,we propose a Review Usefulness Evaluation Model based on the multiple linear regression method.This review usefulness model based on ratings and reviews takes the relevant data of ratings and reviews as independent variables,and assumes that helpful-votes and total-votes can be objective.To reflect the usefulness of reviews, and use it as a dependent variable for ratings and reviews related variables for stepwise regression analysis.Our model fits real well,helps Sunshine Company screen out those reviews that are useful, and quickly and accurately obtain the users' attention to the product and emotional orientation.

We also use the LDA model for data mining of text-based reviews.By summarizing the extracted review topics,we have obtained some product features that customers are more concerned about.

Next,on the basis of thinking that star-rating can reflect the real experience of customers,we fit the star-rating time-based sequence and use it to represent product 's reputation.We performed a stationary test on the sequence,and judged that it meets the requirements of stationarity through its scatter plot of autocorrelation coefficients,and then used the moving average method and GPR (Gaussian Process Regression) to fit the trend of product 's reputation.On this basis,we continue to use the LDA model to mine data on the inflection points,rising periods,and falling periods of reviews on the product 's reputation curve to extract data that can predict the success or failure of the product.

Finally,we use the moving average method to study the relationship between specific star ratings and reviews,ganjiand use simple linear regression to get the relationship between specific quality descriptors of text-based reviews and rating levels.

Keywords: LDA; Review usefulness; Text analysis; MA; Topic extraction

Contents

1 Introduction

1.1 Background

Worldwide, e-commerce is booming, and many companies are using e-commerce as one of their main profit ways.As an important source of information searching for consumers,online customer reviews provide more information details about the product.Through online customer reviews, customers share their personal beliefs about and experiences with purchase decisions and evaluate the services or products after purchase.A well-written review contributes to facilitate a rational purchase decision,identify the advantages and weakness of products for developers,and identify what can be learned for new product development.

Therefore, for Sunshine Company to produce three new products, it is quite useful to obtain useful information from user rating and review data of previous products.This process will help their new products to gain consumer favor and occupy a favorable market position advantageous. Therefore, our team is committed to helping Sunshine Company identify useful information among numerous reviews and conduct data mining to provide advice and reasonable marketing strategies for their new products.

1.2 Problem Restatement

1.Process and clean the review data sets of the three existing products.Mine and analyze the data sets with mathematical theory and quantitative or qualitative analysis methods.Find out the key factors(such as the performance of the products)in the rating and review data that can help Sunshine Company to win the market.

2.Mine data, model for specific requirements of Sunshine Company:

- Select appropriate measurement indicators from the existing product rating and review data set for analysis;
- Conduct data mining and analysis of the data set to obtain the changes of the company's reputation in the online market over time;
- Combine the measurement indicators of ratings with the measurement indicators of text-based reviews to assist Sunshine Company in making production decisions;
- Judge whether the rating of others will directly affect one prerson's review, so that we can truly discover the useful reviews;
- Determine whether there is a certain relationship between some specific reviews and the fixed rating value. If there is a certain relationship, these reviews should be properly processed and analyzed.

2 Assumptions

- Except for outliers and missing values, the data we get is true and reliable.
- When the user reviews the product, he expresses his true experience of using the product, and there is no malicious fraud.
- The user can only access the online review information, and cannot obtain the real product experience and the effects of offline ratings and reviews before purchasing.
- Assume the review of unpurchased users is not convincing, and it is removed as abnormal data.
- Assume that products purchased below 15 are not of reference significance when comparing the average star rating of each product.
- Assume that the influence of users'ratings and reviews has a certain timeliness.
- Assume that the influence of members' reviews and ratings is different from that of ordinary users.
- In addition to ratings and reviews that affect users' choice of products, price is the main influencing factor for users to purchase products.
- The length of a review has a certain effect on the influence of the review.

3 List of Notation

Table 1: The List of Notation

Symbol	Meaning
D	Collection of word sequences for text reviews
d_i	I-th word sequence
W	Corpus base on the data set provided
N	Number of words in the corpus
T_i	I-th topic
w_i	Word
K	The number of topic
ϕ	Topic-word weight matrix
M	Number of word sequences

4 Review Usefulness Evaluation Model

4.1 Preparing the Data Set for Analysis

We found that there are some missing values and outliers in the data set, and some data are shifted. Therefore, we first eliminated outliers and deleted missing values to ensure the structure, integrity and correctness of the data.

4.2 Building the Model

In order to allow consumers to evaluate online reviews, so that high-quality reviews come first, and are easy for consumers to read, online retailers generally use "usefulness" as the basic way for consumers to rate reviews. Mudambi and Schuff define useful reviews as product reviews that facilitate the customer's buying decision process[2].

The usefulness of reviews can be used as a response to the quality of reviews. It can help consumers at multiple stages of their purchase decision. Therefore, for Sunshine Company to use ratings and reviews to assist the production and marketing of new products,high-quality review data is necessary. We designed this model to allow Sunshine Company to predict the usefulness of existing or newly generated reviews and make the most of useful reviews.

In recent years, many experts and scholars have researched the usefulness of reviews, mainly starting from the attributes of the reviewer, the description of the review itself, and the product type.

With reference to many literatures and the dataset of this question,Ghose et al[3]. Considered that too long reviews are less readable, and it takes more time to read, which increases the cost of reading. Its usefulness is low, but the research results are the opposite.Another study found that the prestige of one reviewer and the recognition of other consumers can greatly enhance the impact on other consumers[4][5].

So this model is based on the these assumptions:

- Comment length has a positive impact on the usefulness of comments.
- The identity of the reviewer has a positive effect on the usefulness of the review when the member is a member.
- Buyer reviews have a positive effect on the usefulness of reviews.
- Use the rating star to represent the extremeness of the review, from 1 to 5 stars, 1 star to extreme bad reviews, 3 stars to neutral reviews, and 5 stars to praise.

Table 2: Variable Summary

Variable Name	Measure	Type
vine	whether the reviewer it is a member	Continuous
verified_purchase	whether the reviewer has purchased this product	Discrete
star_rating	rating	Discrete
review_length	review length	Continuous
review_days	number of days a review exists	Continuous
total_sell	total sales of a product	Continuous
helpful_votes	the number of helpfulness rating	Continuous

Our model uses vine(whether the reviewer it is a member),verified_purchase(whether the reviewer has purchased this product),star_rating (rating),review_length(review length),review_days(number of days a review exists),total_sell (total sales of a product) as independent variables,helpful_votes (the number of helpfulness rating) as the dependent variable (ie, the measure of the usefulness of reviews) to analyze the relationship between review usefulness and the above six independent variables.The Varible summary is shown in the Table 2 on page 4.The model is shown as in the Figure 1 on page 4.

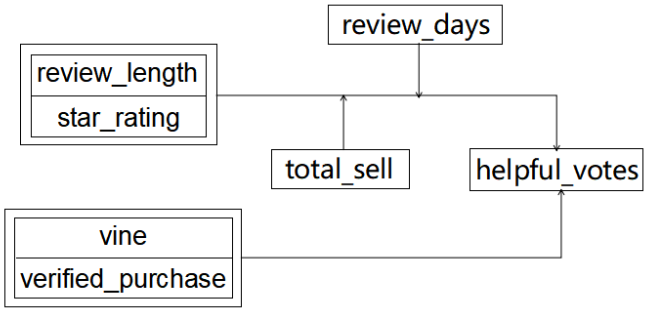

Figure 1: The Usefulness of Review Model

One of the characteristics of our model is to include total_sell (the number of products sold) to help Sunshine fully draw on best-selling products and avoid the same problems as slow-moving products.

4.3 Model Analysis

We first use correlation analysis to study the correlation between the respective variables and the dependent variables. Correlation analysis is a commonly used statistical method to study the correlation between random variables. The correlation coefficient r is used to indicate the degree of correlation between the two variables, and the value of r is calculated using sample data. The value ranges from -1 to 1. $r > 0$ indicates that there is a positive correlation between the two variables, otherwise it is a negative correlation; $r = 0$ indicates that there is no

correlation between the variables; |r|>0.8 indicates that there is a strong correlation between the variables, |r|<0.3 indicates that the correlation between variables is extremely weak and can be considered irrelevant.The correlation analysis result is shown in the Figure 2 on page 5.

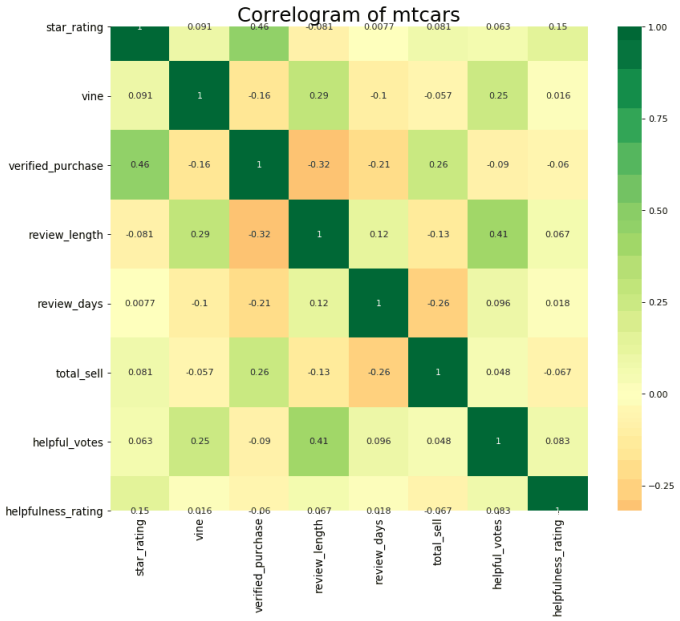

Figure 2: The Correlation of Variables.

The Figure 2 can show that there is no obvious correlation between the respective variables.Next, we performed a multiple linear regression analysis on the dependent and independent variables. The multiple linear regression analysis method uses the optimal linear combination of multiple independent variables to study the relationship with the dependent variable. The combination of multiple variables to predict or estimate the dependent variable is more effective and consistent than using only one variable for prediction.

We use the Stepwise Regression method. Stepwise regression is a commonly used method to eliminate multicollinearity between variables and obtain the "optimal" regression equation. At each step in the calculation process, the partial regression sum of squares (ie, contribution) of the variables that have been introduced into the equation is calculated, and then the significance test is performed on the least contributing variable at a given significance level. This variable remains in the regression equation, otherwise, if it is not significant, you need to delete the variable, and then calculate other unintroduced variables according to this step, until the variables in the regression equation cannot be eliminated without new Until the variables can be introduced, the stepwise regression process ends. The Table 3 on page 6 shows how well the model fits, and the significance values prove that our model performs well.

Table 3: Model Fitting Result

R^2	F	ρ	Square of error
0.0816	160.670	0.000	48.809

The ROC curve as shown in the Figure 3 on page 6 shows our model fits well in practice.

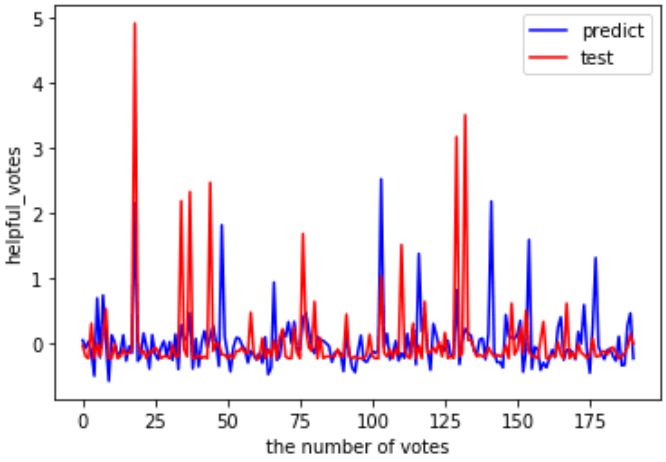

Figure 3: ROC

4.4 Model Validation

We can also find that the fitting of the model in some positions in the ROC curve is poor. We consider the possibility of outliers. After making a residual map, we remove the outliers and find that the effect of the fit is significantly improved.

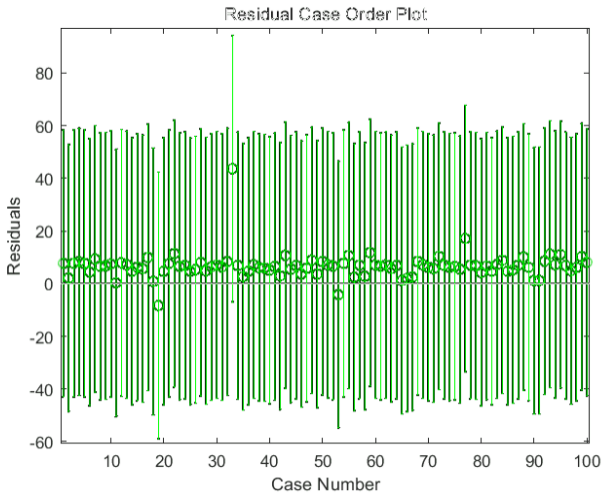

Figure 4: The Residual Map

4.5 Conclusion

Based on our model, Sunshine Company can use it to evaluate the usefulness of each review, which has played a large role in customer purchase decisions. Therefore, finding these useful reviews and analyzing them can accurately identify the user's needs, allow new products to win user trust and capture the market.Useful comments were selected, and we next analyzed the comments to identify useful content.

5 The LDA Topic Model

5.1 Preparing the Model

We choose LDA topic modeling. Like the name of this technical means, topic modeling is the process of automatically determining topics in a text target, while showing hidden semantics from the text corpus. The use of topic modeling is diverse, including: text aggregation, information retrieval from unorganized text, feature selection, and so on. The users' rating and review data for these three products is a typical semi-structured data, and it is completely appropriate to choose this method of subject modeling.

When we build this model,our goal is to extract some important and representative words from the reviews. These keywords will help us understand the customers' attitude to the product, as well as understanding the users' main concerns about the product, such as performance, price, style, etc., and then collecting score data for analysis, providing valuable opinions for Sunshine Company's production.

We treated these three rating and review data sets, first preprocessed them, removed outliers and missing values, ensured the correctness and completeness of the data, and then used the powerful text processing capabilities of the Python language to conduct the review text. deal with. Use the "nltk" and "pandas" libraries in the python language to delete punctuation numbers and delete stop words, and use the "spacy" library to reduce the form of the word to ensure that the form of the comment text is as uniform as possible. We counted and visualized the frequency of words and words appearing in the processed data, and visualized it:

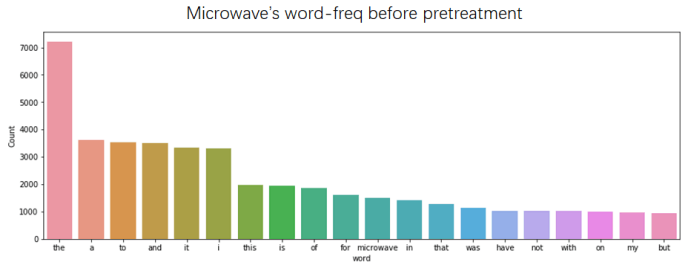

Figure 5: Term Frequency before Pretreatment

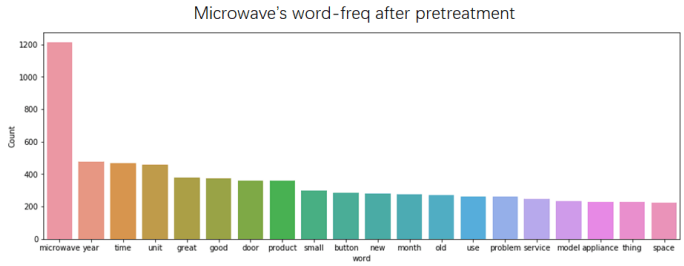

Figure 6: Term Frequency after Pretreatment

5.2 Building and Analysing the Model

After the text processing, we started to use LDA Topic Model. This methodology was chosen as it is simple and is often used in a variety of sciences for topic modeling text corpora, which can be used as a type of text summarization of a large set of documents. LDA produces a model of a corpus of documents, where the model assumes that each of the documents in the corpus are derived from a generative process where each document consists as a distribution of a finite set of topics, each topic is a multinomial distribution of the vocabulary of words in the corpus, and each word of the document is drawn from one topic in the generative process[6].

In this way, we can describe the creation of documents with a two-step process:

1.For each document m $\in$ M sample a topic proportions from θ_m Dirichlet distribution Dir(α).

2.For each word placeholder n in the document m:

a.Choose a topic $z_{m,n}$ randomly according to the sampled topic proportions θ_m.

b.Choose a word $w_{m,n}$ randomly from the multinomial distribution ϕ_k of the previously chosen topic $z_{m,n}$.

In the above-mentioned process, the parameters α and β are vectors of hyper-parameters which determine the Dirichlet prior on θ as a set of topic distributions

for all documents and on ϕ as a set of word distributions in all topics. Typically, symmetric Dirichlet priors are used, where α_1 = α_2 = α_k = α, which define how probability distribution is concentrated into a single point. The entire process is depicted in Figure 7 on page 9.

The calculation process of our model is as follows,we first get Processed word sequence set:

$$D = \left(d_1, d_2, ..., d_M\right) \tag{1}$$

d_i is the sequence of words for each text comment:

$$d_i = \left(w_1, w_2, ..., w_n\right), 0 \leq i \leq M \tag{2}$$

Perform word frequency statistics on the word sequence set to obtain the corpus:

$$W = \left((w_1, p(w_1)), (w_2, p(w_2)), ..., (w_N, p(w_N))\right) \tag{3}$$

$p(w_i)$ is the prior probability statistics of the ith vocabulary in the corpus, which is defined as:

$$p(w) = \frac{\sum\limits_{d \in D} count(w \in d)}{\sum\limits_{w \in W} \sum\limits_{d \in D} count(w \in d)} \tag{4}$$

Then,apply LDA to extract topic,the goal is to generate Topic-Vocabulary Weight Matrix ϕ: Topic (T) refers to a list of words (w_i) that are semantically related to the topic and their weights $p(w_i \mid T)$,are vectors consisting of the conditional probability of each word under the topic. The words that are more closely related to the topic have a higher conditional probability. Otherwise the smaller it is. Represent it as a vector:

$$T = \left\{(w_1, p(w_1|T)), (w_2, p(w_2|T)), ..., (w_K, p(w_K|T))\right\} \tag{5}$$

The topic-vocabulary mixed distribution in the results of LDA topic extraction is ϕ:

$$\phi = (T_1, T_2, ..., T_K) = \begin{bmatrix} \phi_{11} & ... & \phi_{1N} \\ ... & ... & ... \\ \phi_{K1} & ... & \phi_{KN} \end{bmatrix} \tag{6}$$

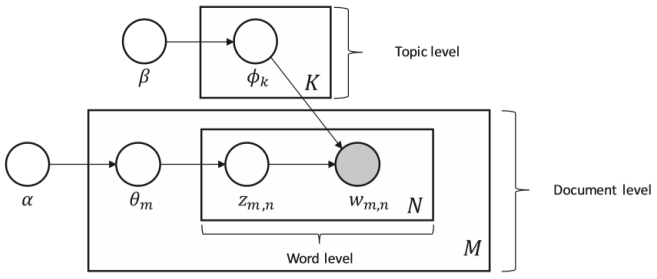

Figure 7: Latent dirichlet allocation model

We will make reasonable suggestions for Sunshine Company based on the n topics we get.Taking the microwave rating and review data set as an example, we show two tpoic of them here,in order to visualize our theme in a two-dimensional space, we use the pyLDAvis library:

Topic 1,which contains the words time, use, easy, etc. We summarize this topic as the ease of operation of microwave ovens,shown in the Figure 8 on page 10. Topic 2,which contains the words, problem, month, warranty, and repair. We summarize this topic as the reliability of microwave ovens,shown in the Figure 9 on page 10.

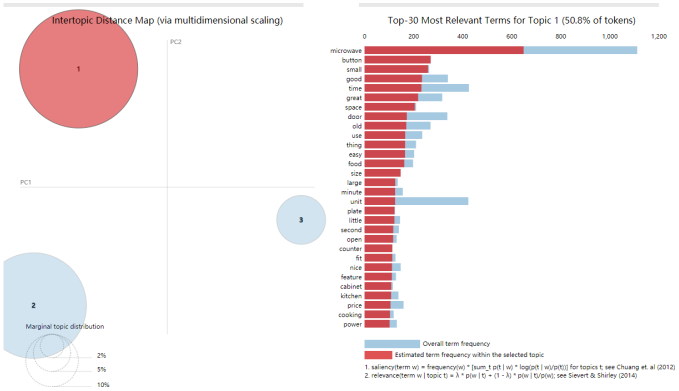

Figure 8: Microwave Topic 1

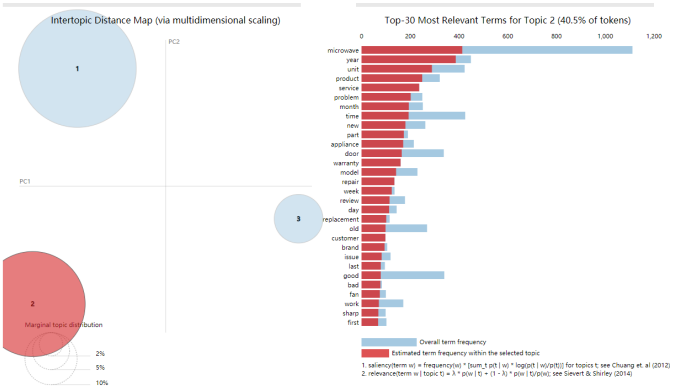

Figure 9: Microwave Topic 2

5.3 Conclusion

Using the LDA model, we performed subject extraction and analysis on the text of the review, helping Sunshine Company to develop and market new products from different perspectives.

6 Time-based Measures and Patterns

In order to explore the measure indicators and patterns of the condition of a product's reputation over time, we think the information of star-rating can be used as the measure indicators. By averaging daily star-rating, we can get a sequence of star-rating's changes over time. Next, we will carry on the detailed analysis proof to this sequence below.

1.Analysis of stationarity

First, analyze the stationarity of the sequence. In the following three pictures, the autocorrelation coefficient quickly dwindled from 1 to near 0. Then as the order rises, it fluctuates slightly on the 0 axis. This basically meets the requirements of stationarity.Here are three graphs of autocorrelation coefficients:

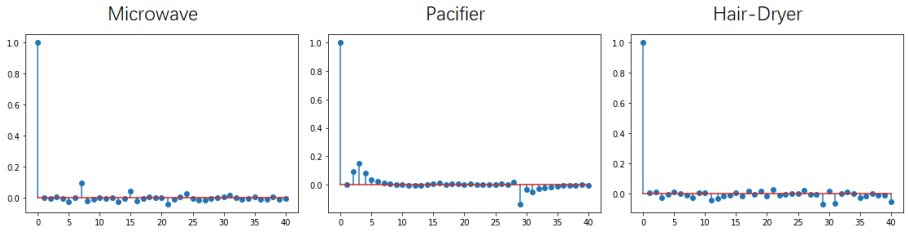

Figure 10: Autocorrelation Coefficients.

2.Moving average

Considering that the moving average method is beneficial to eliminate the impact of periodic fluctuations and random fluctuations on the time series values. At the same time, a sliding window of appropriate length can provide us a macro perspective to observe the trend of the product's reputation. Therefore, we decide to adopt moving average model to process the time series. First, we use the following formula to deal with moving average of the original sequence:

Table 4: Notation

S_i	Synthesis at time i
$\overline{starrating}_i$	Average star rating at time i
N_T	Long time window

$$S_i = \frac{\overline{starrating}_i + \overline{starrating}_{i-1} + ... + \overline{starrating}_{i-N_T+1}}{N_T} \tag{7}$$

In order to see the macro historical trend, we choose =365 day as the time window. The following three time series are obtained after processing:

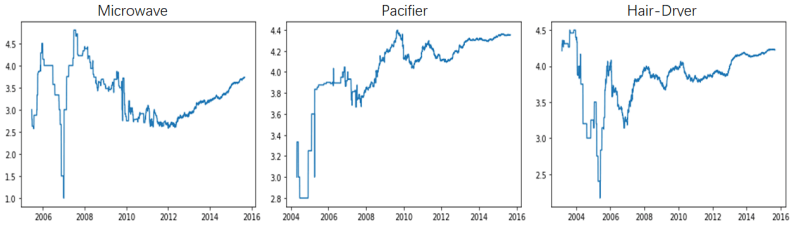

Figure 11: Reputation Change.

3.Use GPR to fit time series

Although the initial sequence obtained by moving average can intuitively reflect the variation trend of product replacement, in order to extract the inflection points of the curve, we still need to do further fitting to the sequence. Based on the observed values, we conclude that the fitted model needs to meet the following requirements:

(1)The fitted model can reflect the variation trend of sequence;

(2)The model can accommodate random noise. To meet the above requirements, we use Gaussian Process Regression (GPR) to fit this sequence, then we can find the timing when the sequence is at Peak and Troughs, as shown in the following figures:

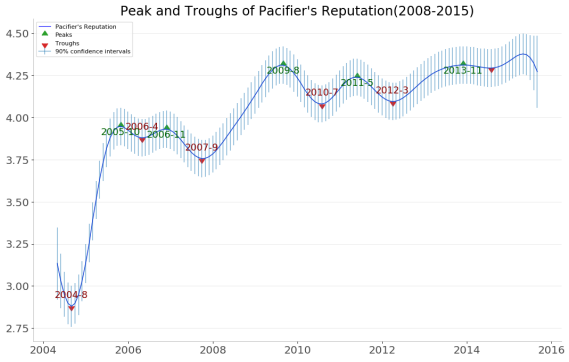

Figure 12: The Curve of Pacifier.

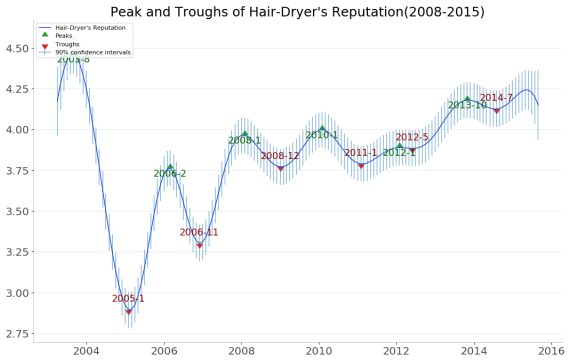

Figure 13: The Curve of Hair-Dryer.

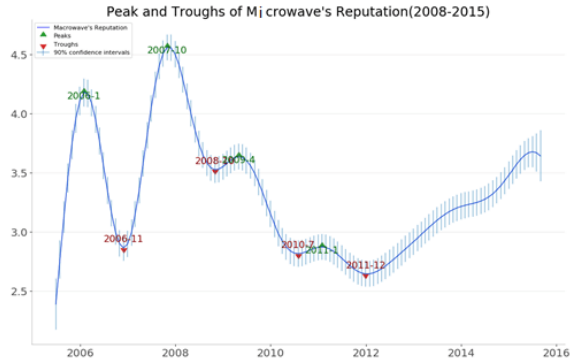

Figure 14: The Curve of Microwave.

Taking Microware as an example, we can see that its reputation declines after October 2007. Although there are fluctuations, the overall trend is decreasing. Then there is an inflection point in December 2011, and its reputation starts to rise and continues to the present.

7 Combinations of Text-Based Measure(s) and Ratings-Based Measures

From the previous product 's reputation trend curve, we can easily see the rising period, falling period and inflection point of each product 's reputation. By analyzing samples from several specific periods, we can derive the main characteristics that predict the success or failure of the product.

As the star-rating and other structured data only can reflect a kind of particular information while text-based reviews can reflect more customer's feedback information, so we decide to still use the LDA model to data mining the comments in the rising period, falling period and inflection point of the above trend curve, and extract the topic from these reviews. The specific method is as follows:

1.Use LDA to analyze the positive comments of star-rating greater than 4 points during the rising period of product 's reputation trend curve. The increase in positive comments will cause the product 's reputation curve to rise, so during the rising period, we analyze the positive comments with a star-rating greater than 4 points.

2.Use LDA to analyze the negative comments of star-rating greater than 4 points during the rising period of product 's reputation trend curve. Contrary to 1, we analyze the negative comments with a star-rating of less than 2 points during this period.

3.Use LDA at the lowest point of the product 's reputation curve to analyze positive comments with star-rating greater than 4 points. At the lowest point of the curve, the reputation curve will show an upward trend. The reason for this

phenomenon is the increase in positive comments, so we analyze the positive comments with star-rating greater than 4 points in this period.

4.Use LDA at the lowest point of the product 's reputation curve to analyze negative comments with star-rating less than 2 points. Contrary to 3, we analyze the negative comments with a star-rating of less than 2 points during this period.

Take the microwave oven as an example: its reputation trend curve dropped to its lowest point in December 2011, and then began to rebound. Because during the moving average process of reputation curve, we choose 365 days as the sliding time window. This results in a delay of up to 365 days in the variation of reputation curve, so we use LDA to analyze the comments within one year before inflection point.The word frequency statistics of text reviews in this period are as follows:

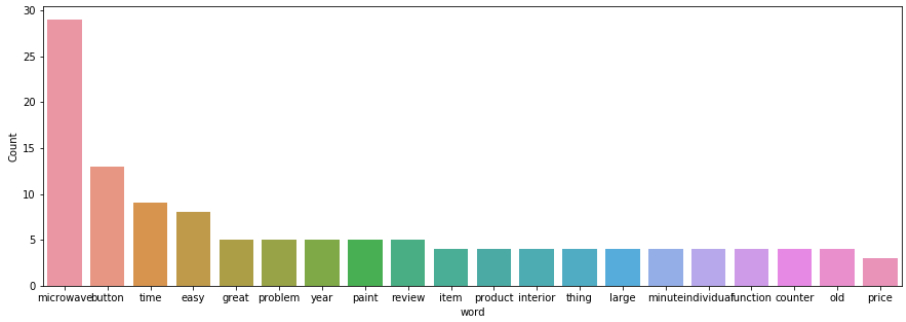

Figure 15: The Word Frequency in the Lowest Period.

From the bar chart, it can be seen from the above figure that the words such as button, time, easy, etc. appear more frequently during this period. It can be assumed that the usability of the product at this time is a potential factor that leads to the success of the product.

Continue to use LDA for further analysis. The topics obtained by the analysis are shown in the figure below:

Words / Topics	1	2	3	4	5	6	7	8	9	10
1	easy	time	first	part	sleek	use	great	install	item	problem
2	microwave	button	function	counter	individual	thing	large	review	bread	panel
3	button	minute	open	cook	easy	second	close	sensor	cooking	drawer
4	microwave	time	year	old	loud	price	happy	product	customer	store
5	paint	paper	hour	grit	microwave	easy	problem	door	fast	kid

Figure 16: The Topic1 Obtained

There are words such as easy, use, time, and install in Topic1, which can be determined to be related to the usability of the product. Topic3 contains words

such as button, minute, and easy. It can be determined that this topic is related to the convenience of the product. Based on the above, we determine that the convenience and usability of the product indicate that the product will be successful.

From this analysis, we can summarize the main characteristics that predict the success or failure of the product from the extracted topics.

Similarly, we analyze the rising period after the above lowest point:

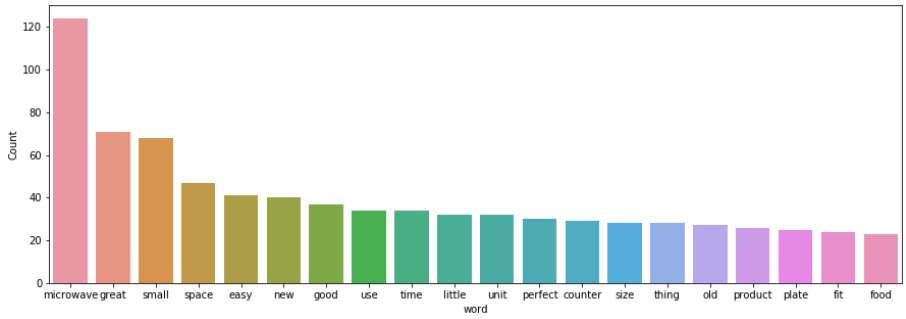

Figure 17: The Word Frequency in the Rising Period.

From the words 'small', 'space', 'size', we can make assumptions that the usability of the product at this period is a potential factor that leads to the success of the product.Further LDA analysis:

Words / Topics	1	2	3	4	5	6	7	8	9	10
1	microwave	time	perfect	cook	space	unit	small	counter	food	huge
2	microwave	good	small	kitchen	love	fit	space	new	size	product
3	microwave	small	minute	time	old	convection	easy	plate	new	stainless
4	great	microwave	easy	use	product	work	item	size	small	large
5	microwave	great	little	new	nice	plate	easy	space	paint	small

Figure 18: The Topic2 Obtained

Topic1 contains words such as space and small. Topic2 contains words such as small, kitchen, fit, and size. It can be determined that this topic is related to the size of the product. Based on the above, we can determine that the appropriate size characteristics of the product indicate that the product will be successful.Then analyze the negative comments of less than 2 points in the falling period:

Words / Topics	1	2	3	4	5	6	7	8	9	10
1	control	item	problem	pay	overcame	heated	correct	unplugged	adage	pad
2	house	noise	unit	door	microwave	smoking	new	poor	low	absolute
3	unit	poor	service	year	food	schedule	recommendation	warranty	turntable	other
4	sharp	service	microwave	unit	week	part	factory	garage	day	appointment
5	appliance	year	oven	repair	model	old	use	less	purchase	power

Figure 19: The Topic Obtained in the Failing Period.

Topic1 contains words such as control, problem, and unplugged. Topic2 contains words such as noise, smoking, and low. From these words, we can determine that this topic is related to the size of the product. Based on the above, we can determine that poor quality characteristics of product indicate that the product will fail.

8 Do specific star ratings incite more reviews?

In order to investigate whether a particular star rating has a certain impact on subsequent reviews, we first make the following assumptions:

- Suppose that each customer will look at about 20 recent star ratings before purchasing a product. These 20 sets of data can be regarded as a window;
- Assume a rating of 1 or 2 stars is called a bad rating, and a rating of 4 or 5 stars is called a positive rating;
- Assume that the number of negative reviews is greater than or equal to 8 in a window, the window can be called a "low score window"; the number of positive reviews in a window is greater than or equal to 14, the window can be called " "High score window" (the average score of these three products is higher, and the number of high scores is much higher than the number of low scores);
- Assume that a customer's review exceeds 100 words, the review is considered to be a long review.

Investigate the impact of a series of low star ratings on the comment length of subsequent comments. Compare the probability of a long comment in the given dataset(Lc1), the probability of the next comment corresponding to the low score window as a long comment (Lc2), and the probability of the next comment corresponding to the sub-window as a long comment (Lc3):

$$Lc1 = \frac{Numberoftotallongcomments}{Numberoftotalcomments} \tag{8}$$

$$Lc2 = \frac{Numberoflongcommentsforlowscorewindow}{Numberoflowscorewindow} \tag{9}$$

$$Lc3 = \frac{Numberoflongcommentsforhighscorewindow}{Numberofhighscorewindow} \tag{10}$$

The final result (with 4 decimal places) is shown in the following table:

Table 5: Probability of Long Comments in Different Windows

	Microwave	Hair dryer	Pacifier
Lc1	0.2445	0.1423	0.1131
Lc2	0.3251	0.2125	0.2500
Lc3	0.1604	0.1339	0.1092

It can be seen from the table that Lc2 is significantly larger than Lc1 and Lc3 is slightly smaller than Lc1. From this we can conclude that a series of low star ratings will motivate customers to write longer reviews to a certain extent. They are positively correlated; The high star ratings have no such promoting effect, the two are negatively correlated, the degree of this negative correlation is weaker than the degree of positive correlation.

To further explore the impact of a series of low star ratings on subsequent reviews, we will conduct two sets of comparative analysis:

(1)Compare the negative rating rate (Nr1) in the given data set with the negative rating rate (Nr2) of the next star rating corresponding to the low score window:

$$Nr1 = \frac{Numberoftotalnegativereviews}{Numberoftotalreviews} \tag{11}$$

$$Nr2 = \frac{Numberofnegativereviewsforthelowscorewindow}{Numberofthelowscorewindow} \tag{12}$$

The final result (with 4 decimal places) is shown in the following table:

Table 6: Negative Rate Corresponding to Different Windows

	Microwave	Hair dryer	Pacifier
Nr1	0.3179	0.1459	0.1118
Nr2	0.3919	0.2000	0.1500

(2)The praise rate (Pr1) in the given data set is compared with the praise rate (Pr2) of the next star rating corresponding to the high score window:

$$Pr1 = \frac{Numberoftotalpraisereviews}{Numberoftotalreviews} \tag{13}$$

$$Pr2 = \frac{Numberofpraisereviewsforthehighscorewindow}{Numberofthehighscorewindow} \tag{14}$$

The final result (with 4 decimal places) is shown in the following table:

Table 7: Positive Rate Corresponding to Different Windows

	Microwave	Hair dryer	Pacifier
Pr1	0.6000	0.7672	0.8035
Pr2	0.6733	0.7758	0.8056

From these two tables, we can see that Nr2 is significantly larger than Nr1, but Pr2 is slightly larger than Pr1, which indicates that a series of positive or negative reviews will affect subsequent customer star ratings. The number of derogatory words and the corresponding ratings are negatively related, the number of commendatory words and the corresponding ratings are positively related, so a series of low star ratings will negatively affect subsequent customer reviews, while a series of high star ratings will have a positive impact, but the impact is weaker than the negative impact of a low star rating.

9 Explore the Relationship between Descriptors and Product Ratings

To explore the relationship between some specific descriptors and product ratings, we use python to preprocess and analyze the text, filter out the adjectives that appear in the reviews, and perform word frequency statistics on the adjectives. Select the word frequency first. Thirty adjectives were analyzed.

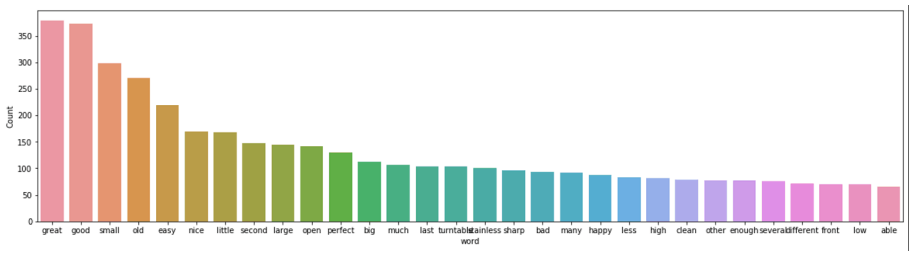

Figure 20: The Word Frequency First. Thirty

For each adjective, we adopt Bernoulli's method to express whether the adjective appears in each comment. If the adjective appears, f(x) is marked as 1 and if not, it is marked as 0.We use a linear regression method to explore the relationship between f(x) and rating for each adjective. We can see that when the adjective does not appear in a comment, the scoring results are roughly above and below the average. However, when an adjective appears in the comments, the result of the scoring obviously deviates from the average.The figure below uses "great","bad","disappointed" as examples to show our findings above.

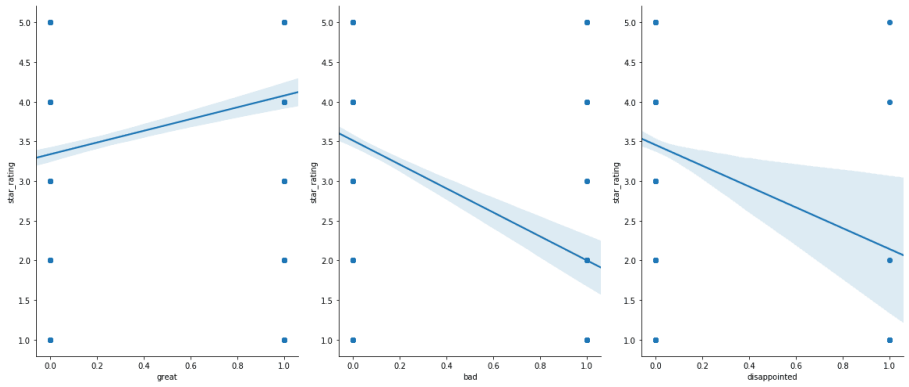

Figure 21: "Great","Bad"and"Disappointed"

10 Strengths and Weaknesses

10.1 Strengths

- After establishing the model, check the situation of outliers through the residual graph, remove some outliers, and then visualize the model to find that the effect is much better;
- LDA was used to extract the topics, and then the optimized topics were analyzed multiple times. Finally, five topics were obtained. After comparing and analyzing these topics with the original data, it was found that the two had a high degree of agreement;
- Because the data is sorted by time, the sample data is enough, the size of the sliding window is negligible compared to the sample, but at the same time it is not too small to reduce the accuracy of the results. These conditions greatly enhance the result's accuracy after the use of the sliding window;
- The smoothing is performed on the initial sequence obtained by using the moving average method. Through observation of the sequence noise, it is found that the noise obeys the Gaussian distribution. Therefore, fitting the sequence using GPR can extract the inflection points of the curve well.

10.2 Weaknesses and Extensions

- It is assumed that the star rating can reflect the objective thinking of the customer, and does not consider the behavior such as scalping;
- The moving average method is used, resulting in unpredictable fluctuations that will lead to higher or lower future curves;

- When using the LDA model to extract topics, it is assumed that the topics are independent of each other, and the correlation between the topics is not considered.

11 The Letter to Sunshine

To: Marketing Director of Sunshine Company

From: Team #2013133

Date: Mar 10th, 2020

Subject: Information of concern, Product characteristics, Sales strategy

Honorable Marketing Director of Sunshine Company,

Your company provided us with data on ratings and reviews of microwave, hair dryers, and pacifiers sold on the Amazon market within a certain period of time. Our team established a corresponding model to analyze these data, and finally excavated the information that customers care about, the main features of the product, and future sales strategies.

Information of concern: Our team has established a usefulness of review model, which extracts the topics that customers care about from the information given in the table, and then comprehensively analyzes these topics to get 5 topics that customers care about most: usability, durability, structure, experience, fitness. Later, we also extracted adjectives in the reviews of different products, and found that when the adjective did not appear in a review, the scoring results were roughly around the average, but when an adjective appeared in the reviews, the scoring results significantly deviated from the average. These adjectives are obviously what customers care about.

Product characteristics: Select the top 15 brands of each product for analysis and find that the average star ratings of products which are the same type but different brands are not much different, and the star ratings are high; the product reputation is analyzed using the sliding review method. It was found that the reputation of the microwave was in a downward trend between 2009 and 2012, reached its lowest point in 2012, and has been increasing since then. Using LDA to investigate the reasons for the microwave oven's decline or rise, finally we summarize the potential factors affecting the success / failure of the product from the extracted topics. While the reputation of the pacifier and hair dryer have been increasing between 2006 and 2015.

Sales strategy: sellers should pay more attention to the comments of members, because after our analysis, we find that the members' comments were liked more often which indicates it is easier to get the approval of other users; before investing in these products, your company had better pay attention to the adjectives and themes extracted by our team from the data of the three products, es-

pecially from the microwave, and carefully analyze the meaning of these words, so as to ensure the quality of the product and improve customer satisfaction and reputation; in our research, we found that a series of low-star reviews will cause customers to have bad emotions, thereby reducing the desire to shop, so the level of the star also has a certain impact on sales.

The above is the main content of our research. I hope these contents can provide some useful information to your company.

Thanks

12 Reference

1 Mudambi S M, Schuff D. What Makes A Helpful Online Review?A Study of Customer Reviews on Amazon.Com [J].MIS Quarterly,2010,34(1):185-200.

2 Ghose A,Ipeirotis P G.Designing Novel Review Ranking Systems:Predicting The Usefulness and Impact of Reviews[C].Proceedings Of The Ninth International Conference on Electronic Commerce, New York, NY, USA: Association Computing Machinery(ACM).2007:303—310.

3 Gruen T, Osmonbekov T, Czaplewski A. EWOM: The Impact of Customer-to-Customer online Know-How Exchange on Customer Value and Loyalty[J]. Journal of Business Research,2006,59(4):449-456.

4 Lerner J, Pathak P, Tirole I. The Dynamics of Open-Source Contributors[J]. The American Economic Review,2006,92(2):114-118.

5 Kyle Porter,Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling[J].Digital Investigation,2018,26(4):87-97.

6 Miha Pavlinek,Vili Podgorelec.Text classification method based on self-training and LDA topic models[J].Expert Systems with Applications,2017,80,(3):83-93.