Problem Chosen	2020	Team Control Number
C	MCM/ICM Summary Sheet	2020230

A Wealth of Data

Summary

To assist Sunshine Company in developing a scientific and reasonable online sales strategy, our team shall establish a SO-PMI based Review Quantization Model (RQ) and the Comprehensive Rating Model of Product Reputation (CRPR). Based on the data provided and the two models, we analyzed the influence of star rating, reviews and helpfulness rating on the sales of goods and the relationship among them.

After data processing (including index deletion, meaningless data elimination, etc.), we introduce an RQ model to quantify the text review. First, after text segmentation and elimination of stop words, we build a dictionary of positive and negative words as seed words. Then, we use the revised formula to calculate the SO- PMI value of each word in a review and obtain the quantified review.

To study the popularity of products, we introduce a CRPR model. First, we define review values with the combination of quantified review and helpfulness ratings and define identity coefficient using vine and verified_purchase. Then, use the entropy weight method to determine the weight of the index, and obtain the reputation value after linear weighting. Taking the pacifier data as an example, and the weight are 0.49474 and 0.50526, respectively. Finally, the sensitivity analysis of the identity coefficient using hair_dryer data shows that both vine and verified_purchase have almost the same impact on reputation. Moreover, our correlation analysis indicates that reviews are more important than star ratings.

To investigate the change of reputation with respect to time, we use the least square method and obtain the functional relationship of hair_dryer, microwave and pacifier. RMSE are 0.14, 0.16, and 0.09 for hair dryer, microwave, and pacifier, respectively.By analyzing the fitting curve, we conclude that the reputation of hair_dryer and microwave is stable in the later period, while pacifier fluctuates greatly.

To find potential successful products based on the functional relationship between product reputation and time, we establish an LDA(Latent Dirichlet Allocation) based product prediction model. First, we determine the reputation inflection points of 3 data set according to the fitting curves. Take hair_dryer as example, we obtain the corresponding time period as: (2012.12-2013.1, 2015.3-4). Then, we use the LDA to extract keywords from the comment sets in these time periods and employ the PMI to calculate the semantic similarity between words, thereby predicting products success or failure by combining the semantic similarity and the sales volume. Taking hair_dryer set as an example, we find product_id =b003v264ww is a potential successful product.

Also, we conduct cross correlation analysis of star ratings and review values, which indicate that the rating will trigger more reviews. Based on the theme extraction and word frequency of each star levels review by LDA, we observe that some specific words in the comments will affect the star rating.

Keywords: SO-PMI; LDA; Least Squares; Correlation analysis; Sales strategy

Contents

1 Introduction

1.1 Background

With the continuous improvement of web technology and the rapid development of e-commerce technology, online shopping has become an important channel for netizens to purchase. In order to better understand consumers' subjective impression of products, many online shopping platforms provide online review mechanism for consumers. These review data play an important role in online shopping transactions. On the one hand, shopkeepers can use the data to deeply understand customers' needs, further improve the quality of goods and provide more comprehensive after-sales service. On the other hand, for potential consumers, due to the lack of direct access to commodity entities, the evaluation information from previous consumers greatly assists them in making purchases.

However, in the face of such complex review data, how to quickly and effectively identify the valuable comments as well as their contained consumer emotional tendency, and thus to get an accurate score of the goods is a crucial issue.

We are commissioned by Sunshine Company to study the provided data and thus to help them develop a scientific and reasonable online sales strategy.

1.2 Our Work

Based on our understanding of the problem, we set the following goals:

- Process the given data set (including field selection and processing of meaningless data) to determine which fields are useful for analyzing the online sales status of goods.
- Establish a review quantization model(RQ Model) to quantify the text comment into an intuitive value that can be calculated.
- On the basis of the review quantification model, the comprehensive rating model of product reputation (CRPR Model) was established by combining indicators such as star ratings.
- Establish a regression model of reputation and time to study the influence of time on the products reputation in the online market.
- Establish LDA Based Product Prediction Model to anticipate potential successes or failures.
- Combined with the above model for analysis, help Sunshine Company to develop a scientific and reasonable online sales strategy.

2 Preparation of the Model

2.1 Assumptions

Given that the actual situations are extremely complicated, we need to make some relatively idealistic assumptions to focus on the major issues. Our assump- tions come as following:

- Assumed that there is no malicious rating, such as the high rating of the store and the low rating of the competitors.
- Assume that each account corresponds to one person, and one person does not use multiple accounts.
- Assume one cannot illegally become a Vine voice user.
- Assuming that the additional impact of additional comments on comment quantification is ignored, it is directly calculated in the same way as other comments.

2.2 Notations

Table 1: The main symbols used in our paper

Symbol	Description
r	The review affective tendency value
T	The set of review texts
t_i	The i^{th} review text in T
L	The set of words
w_i	The i^{th} word in L
S	The set of SO-PMI value
p_i	The i^{th} SO-PMI value in S
h	Helpfulnes rating
w	Review value
v	Identity coefficient
γ	The product reputation value

2.3 Data Processing

In a large number of raw data, there are often some incomplete and abnormal data, which seriously affects the efficiency of modeling and the accuracy of conclusions. Therefore, it is very important to preprocess the data before using them.

2.3.1 Filtering of Data Fields and Eliminating of Dirty Data

Although users in different countries may have different opinions on the same product, considering that the **marketplace** in the provided data is US, it can be eliminated as redundant data. In addition, the data corresponding to **product_id**, **product_parent**, **product_title** and **roduct_category** have a lot of redundancy, while **product_id** is the unique ID of the product belonging to a comment. Therefore, we choose to keep **product_id** and delete others.

Whether a customer is invited to become an Amazon Vine Voice and whether he/she bought the product can have a huge impact on the usefulness his/her review. When **vine = N** and **verified_purchase = N** means that the customer is neither Amazon Vine member nor has purchased the product. These customers did not actually use the product but commented on it, which shows low credibility and usefulness of their comments. Thus, we decide to eliminate this category of consumer data.

2.3.2 Determination of Helpfulness Rating

In deciding whether to buy a product, consumers often look up reviews of previous purchases. We use **helpfulness_rating** to indicate the value of a review, which is defined by the proportion of **helpful_votes** in **total_votes**:

$$helpfulness\_rating = \frac{helpful\_votes}{total\_votes} \tag{1}$$

2.3.3 Merge Review Headline and Review Body

Since both the title and content of the comment have an impact on the subsequent text quantification, we consider that the two belong to a parallel relationship and combine them to improve the accuracy of text quantification.

Considering that there may be some emotional inconsistencies between the headline and the body, such as the headline expressing a positive evaluation while the body is not, we will further deal with this in the subsequent quantification modeling of the text.

After the above processing, the valid data fields we get are shown in the following table:

Table 2: Valid data fields

Field	Description
customer_id	The unique number that identifies a customer
review_id	The unique ID of a review
product_id	The unique ID of a product
star_rating	The 1-5 star rating of the review
helpfulness_rating	Defined by the proportion of helpful_votes in total_votes which shows the usefulness of a review
vine	Amazon Vine Voices flag
verified_purchase	The flag shows whether the reviewer bought the product
review	The combination of review_headline and review_body
review_date	The date the review was written

3 Models

In this paper, we propose the following models.

3.1 SO-PMI Based Review Quantization Model

Consider that the reviews in the dataset are in the form of text, which is not conducive to participating in the computation. Therefore, we adopt the improved SO-PMI model to quantify the comment text. The quantified value is called the review affective tendency value, which is denoted as r_i.

Model Establishment

For a given set of review text data T $= \{t_1, t_2, \cdots, t_n\}$, t represents the customer 's comments on a product. Because English text is composed of words, there is a space between the words, so for English text segmentation only need to be divided by the space.[2] The resulting word list is set as $L' = \{w_1, w_2, \cdots, w_n\}$, where w_i represent the i^{th} word in the review. Using the English stop word dictionary StopDict (D_s) collected from the Internet, filter w_i. If $w_i in D_s$, then delete w_i from L'. Keep repeating until get a new list L, which can be view as a collection of emotion words.

By using the co-occurrence between the pairs of emotion words, we can judge the polarity of emotion words.[3] SO-PMI is such an algorithm, which mainly introduces the PMI method to calculate the emotional tendency of words. The Pointwise Mutual Information(PMI) is mainly used to calculate the semantic similarity between words. The basic principle is to count the probability of two words appearing in the text at the same time. The greater the probability, the closer and higher the correlation is. The calculation formula of PMI values for word1 and word2 is as follows:

$$PMI(\text{ word } 1, \text{word } 2) = \log_2\left(\frac{P(\text{ word } 1\&\text{ word } 2)}{P(\text{ word } 1)P(\text{ word } 2)}\right) \tag{2}$$

Where, P(word1&word2) represents the probability of word1 and word2 appearing together, and P(word1) and P(word2) represent the probability of two words appearing separately. The greater the co-occurrence probability of two words in a small range of the data set, the greater the correlation degree. Conversely, the correlation degree is smaller. And the ratio of P(word1&word2) to P(word1) to P(word2) is a measure of the statistical independence of word1 and word2. Its value can be converted into three states:

- P(word1&word2) > 0: The two words are related; The larger the value, the stronger the correlation.
- P(word1&word2) = 0: The two words are statistically independent, unrelated and not mutually exclusive.
- P(word1&word2) < 0: The two words are unrelated and mutually exclusive.

In order to be more accurate, we choose the emotional words with obvious tendency from the English library of positive and negative words collected on the Internet as the seed words. Emotional words (also known as test words, i.e. $L = \{w_1, w_2, \cdots, w_n\}$) were traversed. Meanwhile, PMI was used to calculate the correlation strength between test words and seed words, so as to judge the polarity of test words. For example, words that are often associated with "good" tend to be positive, while words that are often associated with "bad" tend to be negative. The difference between the PMI value of the test words and the positive seed words minus the difference between the PMI value of the test words and the negative seed words is the semantic tendency, and the formula is as follows:

$$\begin{aligned} SO-PMI(word\_i) = & \sum_{Pword \in PwordSet} PMI(word\_i, Pword) \\ & - \sum_{Nword \in NwordSet} PMI(word\_i, Nword) \end{aligned} \tag{3}$$

Where, PwordSet and NwordSet represent the positive word set and the negative word set respectively. The result is the emotional disposition of the i^{th} word w_i in

L. Take 0 as the threshold value of the so-pmi algorithm, thus three states can be obtained:

- SO-PMI(word) > 0: Positive tendency, positive words.
- SO-PMI(word) = 0: Neutral tendency, neutral words.
- SO-PMI(word) < 0: Negative tendency, negative words.

Thus, the corresponding SO-PMI value p_i of each word w_i is obtained, denoting as $S' = \{p_1, p_2..., p_n\}$. Then, we can use the following formula to calculate the final review affective tendency value:

$$\begin{aligned} r &= \sum_{j=1}^{n} p_j \\ &= \sum_{j=1}^{n} SO - PMI(w_j) \end{aligned} \tag{4}$$

Model Modification

According to this formula, we can find that the calculated so-pmi value may be greater than 1 or less than -1, as shown in the following table.

Table 3: SO-PMI of hair_dryer(partial)

word	**SO-PMI**
beautiful	4.78821
good	7.26938
suck	-3.79023
normal	0.01217
great	6.98453
...	...

In other words, there is no definite upper or lower bound for the so-pmi value, which hinders subsequent modeling. Therefore, we make the following modifications to the model calculation process:

- **Step 1**: Input a review text T
- **Step 2**: Segment T by space to obtain a word list $L' = \{w_1, w_2, \cdots, w_n\}$, w_i represents the i^{th} word in the list
- **Step 3**: Use StopDict (D_s) to filter w_i, then obtain a new list L
- **Step 4**: Use PwordSet and NwordSet to traverse L, calculate the corresponding SO-PMI value of each word w_i, and a SO-PMI list $S' = \{s_1, s_2, \cdots, s_n\}$ is obtained
- **Step 5**: Calculate $p_{max} = \max(S')$ and $p_{min} = \min(S')$

- **Step 6**: Traverse S' and use the following formula to normalize p_i, then gain a new list S:
$$p_i = \frac{p_i - \min(S')}{\max(S') - \min(S')} \tag{5}$$

- **Step 7**: Calculate the mean value of list S to obtain r
$$\begin{aligned} r &= \sum_{j=1}^{n} p_j \\ &= \sum_{j=1}^{n} SO - PMI(w_j) \end{aligned} \tag{6}$$

- **Step 8**: Output mean r, which is the quantized value of review T

Through step 6, we normalize the value of SO-PMI so that the final result falls within the interval of [0,1].

3.2 The Comprehensive Rating Model of Product Reputation

3.2.1 The Determination of Model Indicators

1. **Star rating**: The numerical score is an overall evaluation of the product by consumers, indicating their satisfaction with the product. It also indirectly reflects consumers preference and the reputation of the product. Therefore, we take it as one of the indicators for the comprehensive evaluation of product reputation and denote it as s_i, which represents the average score of product i. s_i can be calculated by the following formula:
$$s_i = \frac{\sum_{j=1}^{n} s_{ij}}{n}, \quad s_i, s_{ij} \in \{1, 2, 3, 4, 5\} \tag{7}$$
Where, s_{ij} means the j^{th} star rating of product i and n is the total number of star rating for product i.

2. **Helpfulness rating**: Helpfulness rating h_i, determined by the ratio of helpful_votes to total_votes, is the auxiliary rating of a product review which has certain influence to the consumer purchase decision. The higher the helpfulness rating, the more useful the comment is:
$$h_i = \frac{\sum_{j=1}^{n} h_{ij}}{n}, \quad h_i, h_{ij} \in [0, 1] \tag{8}$$
Where, h_{ij} means the j^{th} helpfulness rating of product i and n is the total number of helpfulness rating for product i.

3. **Review affective tendency value**: r_i, which is the quantified value of a product's reviews. Note that the bigger
$$r_i = \frac{\sum_{j=1}^{n} r_{ij}}{n}, \quad r_i, r_{ij} \in [-1, 1] \tag{9}$$
Where, r_{ij} means the j^{th} affective tendency value of a product i's review and n is the total number of review for product i. Note that r_{ij} is given by the SO-PMI model.

4. **Review value**: Considering that h_i and r_i are both related to the review. Therefore, we define a new indicator **review value** w_i' to simplify calculation, including the original quantified reviews affective tendency value and helpfulness_rating:

$$w_i = (1 + h_i)\, r_i \tag{10}$$

Note that the greater the value of w_i, the greater the degree of consumers' preference or aversion to the product.

5. **Identity coefficient**: Vine indicates whether users are invited to become Amazon Vine Voices, and vine comments are submitted independently and cannot be influenced, modified or edited by the seller. Amazon will not change or edit Vine reviews as long as it complies with its publishing policy. Vine Voices reviews have more impact than regular reviews. Thuswe define a new indicator identity coefficient v_{ij}. The tentative values are as follows:

$$V_{ij} = \begin{cases} 0.2 \quad \text{Vine } = N;\ \text{verified\_purchase } = Y \\ 0.3 \quad \text{Vine } = Y;\ \text{verified\_purchase} = N \\ 0.5 \quad \text{Vine} = Y; \text{verified\_purchase} = Y \end{cases} \qquad (\mathrm{i} = 1, 2 \cdots \mathrm{n}; \mathrm{j} = 1, 2 \cdots \mathrm{m}) \tag{11}$$

Where, v_{ij} means the j^{th} review's customer identity coefficient of product i.

3.2.2 Indicators Normalization

Due to certain differences in the properties, dimensions and other characteristics of each indicator, it is impossible to directly compare or synthesize between different indicators. In order to ensure the reliability of the results, we need to carry out numerical normalization on these indicators, so that the values of all indicators are in the same quantity level, so that the indicators of different units or different intervals can be comprehensively analyzed. In this paper, we apply **min-max normalization** to the original indicators to make the result fall into the interval of $[0, 1]$:

$$s_i = \frac{s_i - \min(S)}{\max(S) - \min(S)} = \frac{s_i - 1}{4} \tag{12}$$

Where, $s_i \in S = \{1, 2, 3, 4, 5\}$. Thus, $\min(S) = 1$ and $\max(S) = 5$

$$r_i = \frac{r_i - \min(R)}{\max(R) - \min(R)} = \frac{r_i + 1}{2} \tag{13}$$

Where, $r_i \in R = [-1, 1]$. Thus, $\min(R) = -1$ and $\max(R) = 1$

Based on the above derivation, we now give the **Comprehensive Rating Model of Product Reputation**:

$$\begin{aligned} \gamma &= \sum_{j=1}^{m} v_{ij} \cdot (\alpha s_i + \beta w_i) \\ &= \sum_{j=1}^{m} v_{ij} \cdot (\alpha \cdot s_i + (1 + h_i) \cdot r_i) \end{aligned} \tag{14}$$

Where, γ represents the final value of product reputation and all the indicators are normalized. Note that α and β respectively represent the weight of star rating and review value.

3.2.3 The Determination of Indicator Weight

In order to comprehensively consider subjectivity and objectivity, the entropy weight method is used to determine the weight. In Information Theory, entropy is a measure of uncertainty. The more information there is, the less uncertainty and entropy there is. According to the characteristics of entropy, we can judge the degree of randomness as well as disorder of an event by calculating the entropy value and can also judge the degree of dispersion of an indicator by using the entropy value. The greater the dispersion degree of the indicator, the greater the influence of the indicator on the comprehensive evaluation.

Therefore, we use information entropy to calculate the weight of each indicator and provide a basis for the comprehensive evaluation of multiple indicators. Assuming that there are n groups of data and m evaluation indicators, nŒm indicator matrix Z can be obtained, which is expressed as follows:

$$\mathrm{Z} = \begin{pmatrix} Z_{11} & \cdots & Z_m \\ \vdots & \ddots & \vdots \\ Z_{n1} & \cdots & Z_{nm} \end{pmatrix} \tag{15}$$

1. First, calculate the j^{th} index proportion of group i:

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}} \tag{16}$$

2. Then, calculate the proportion of the data of group j in the index i and the entropy of the index j:

$$\begin{gathered} p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}}, \quad i = 1, \cdots, n, j = 1, \cdots, m \\ e_j = -k \sum_{i=1}^{n} p_{ij} \ln(p_{ij}) \end{gathered} \tag{17}$$

   Where, $\mathrm{k} = \frac{1}{lnn} > 0$ and $e_j \geq 0$

3. Now, calculate the information entropy redundancy:

$$d_j = 1 - e_j \tag{18}$$

4. Finally, calculate the weight of each indicator, and then gain the weight of each indicator:

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{19}$$

According to the entropy weight method, the weights of the normalized indexes are assigned: 0.49474 and 0.50526.

3.3 Least Square Based Reputation-Time Relationship Model

The least square method seeks the best functional match of the data by minimizing the sum of squares of errors. The unknown data can be obtained simply by the least square method and the sum of the squared errors between the obtained data and the actual data is minimized. Therefore, we choose the least square method to fit the relationship between commodity reputation and time in each data set.

Suppose that the given function $\gamma = F(t)$, the value at the point t_i is γ_i Now calculate the following equation, where the polynomial is $p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^k$:

$$\sum_{i=1}^{k} \left(p\left(t_i\right) - \gamma_i\right)^2 = \min \tag{20}$$

In order to obtain the a value of the load condition, take the partial derivative of the left side with respect to a_i, then get k+1 equations:

$$\begin{aligned}
&-2 \sum_{i=1}^{n} \left[\gamma - \left(a_0 + a_1 t + \ldots + a_k t^k\right)\right] = 0 \\
&-2 \sum_{i=1}^{n} \left[\gamma - \left(a_0 + a_1 t + \ldots + a_k t^k\right)\right] t = 0 \\
&\cdots \\
&-2 \sum_{i=1}^{n} \left[\gamma - \left(a_0 + a_1 t + \ldots + a_k t^k\right)\right] t^k = 0
\end{aligned} \tag{21}$$

Rarrange the equation, then get:

$$\left\{\begin{array}{l}
na_0 + a_1 \cdot \sum_{i=1}^{k} t_i + \cdots + a_k \sum_{i=1}^{k} t_i^k = \sum_{i=1}^{k} \gamma_i \\
a_1 \cdot \sum_{i=1}^{k} t_i + a_2 \cdot \sum_{i=1}^{k} t_i^2 + \cdots + a_k \cdot \sum_{i=1}^{k} t_i^{k+1} = \sum_{i=1}^{k} t_i \cdot \gamma_i \\
a_1 \sum_{i=1}^{k} t_i^k + a_2 \cdot \sum_{i=1}^{k} t_i^{k+1} + \cdots + \sum_{i=1}^{k} a_{k-1}^k t_i^{2k} = \sum_{i=1}^{k} t_i^k \cdot \gamma_i
\end{array}\right. \tag{22}$$

Where time t is in units of one month.

We use python to achieve curve fitting and the results are as follows:

The expression of the reputation and time fitting function of hair_dryer is:

$$F(t) = 2.663^{-9} t^5 - 7.471^{-7} t^4 + 7.537^{-5} t^3 - 0.003 t^2 + 0.052 t + 2.001 \tag{23}$$

The corresponding curve is as follows: It can be seen from the figure that the reputation of hair_dryer in the early stage is on the rise, and it is gradually stable in the later stage, and the sales volume also increases gradually with time, so it can be seen that the success rate of hair_dryer is relatively high.

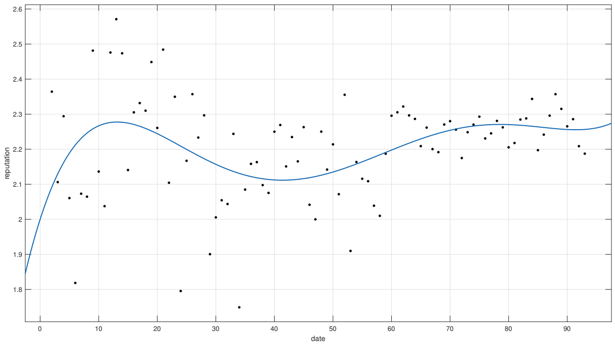


Figure 1: Fitting result of reputation value of hair dryer

The expression of the reputation and time fitting function of microwave is:

$$F(t) = -4.168^{-6}t^4 + 0.0003t^3 - 0.007t^2 + 0.055t + 2.076 \tag{24}$$

The corresponding curve is as follows: From the fitting curve of microwave, we can see that microwaves reputation is stable and sales volume is increasing with time. This shows that the chances of its success are also high.

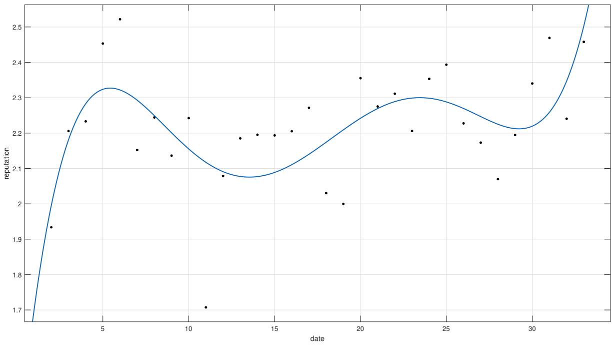


Figure 2: Fitting result of reputation value of microwave

The expression of the reputation and time fitting function of pacifier is:

$$F(t) = -3.812^{-07}t^4 + 4.171^{-5}t^3 - 0.00136t^2 + 0.01603t + 2.191 \tag{25}$$

The corresponding curve is as follows: According to the fitting curve of pacifier reputation, it can be seen that the reputation of Pacifier has been on the rise in the early period with the change of time, until it suddenly declines in the middle and late period. Therefore, later pacifier may fail.

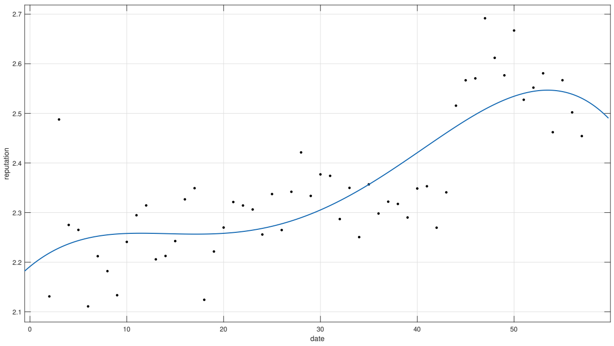


Figure 3: Fitting result of reputation value of pacifier

3.4 LDA Based Product Prediction Model

According to f1,f2 and f3, we determined the reputation inflection points of 3 data set and obtain the corresponding time period as: (2012.12-2013.1, 2015.3-4),(2014.10-2015.1, 2015.3-4), (2014.7-8, 2015.1-2). Now, were going to extract the reviews of 3 data sets in their respective time periods and use LDA to extract keywords from these reviews.

The so-called LDA (Latent Dirichlet Allocation) model is a topic model, which can give the topic of each document in the document set in the form of probability distribution, so as to extract their topic (distribution) by analyzing some documents. At the same time, it is a typical word bag model, that is, a document is composed of a group of words, and the relationship between words is not considered. A document can contain multiple topics, and each word in the document is generated corresponding to one of the topics. The specific steps are as follows:

1. Treat each comment in the known dataset as a text, and sample from the Dirichlet distribution α to generate an arbitrary potential topic distribution θ_k of text k.

2. Use sample from potential topic polynomial distribution θ_k to generate text k's n^{th} word topic $z_{k,n}$.

3. The word distribution ϕ_z corresponding to potential topic $z_{k,n}, n$ is generated by sampling from dirichlet distribution β.

4. Extract the word $w_{k,n}$ from the polynomial distribution ϕ_z of the word.

5. Repeat steps 2, 3 and 4 until all words are extracted.

6. Finally, the corresponding topic distribution of each text and the corresponding word distribution of each topic are obtained.

From this, the corresponding keywords and corresponding probabilities in each comment can be obtained, and finally, the keywords of all comments are integrated to obtain the keywords of the entire data set using k-means.

We select the hair_dryer data set as example and calculate the corresponding keywords and probabilities of all comments during the period from 2012.12-2013.1. The results were as follows:

$$[(0,'0.039*"hair" + 0.037*"dryer" + 0.016*"blow"'), \tag{26}$$

$$(1,'0.044*"hair" + 0.039*"dryer" + 0.010*"dry"'), \tag{27}$$

$$(2,'0.042*"dryer" + 0.039*"hair" + 0.022*"great"')] \tag{28}$$

The equivalent cloud image is shown as figure 4:

We use the key words from the dataset as the reference words during that time period, and re-iterate all the comments, namely the vocabulary list L as the test words. At the same time, the PMI formula is used to calculate the semantic similarity between reference words and test words. The higher the PMI value is, the closer the semantics of the two words are. When the key word of a product is very high with the key word of that period of time, if the reputation of the product shows a downward trend and the sales volume of the product also decreases, it indicates that the product is a potential failure product; otherwise, it is a potential success product. After calculation,

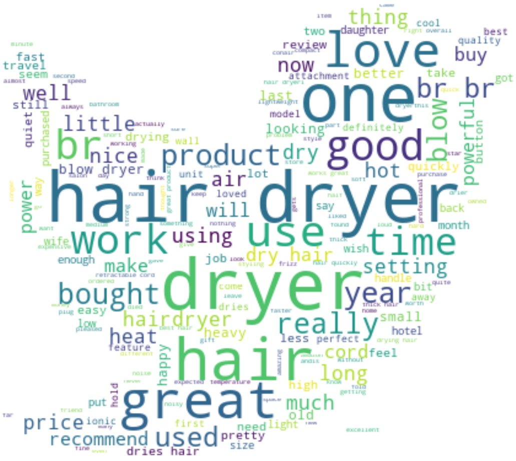


Figure 4: Cloud image of hair_dryer set

we obtain the distribution of product(id=b003v264ww)s keywords in hair_dryer data set as follows:

$$[(0,'0.047*"dryer" + 0.046*"hair" + 0.019*"one"'), \tag{29}$$

$$(1,'0.031*"hair" + 0.028*"dryer" + 0.011*"it"'), \tag{30}$$

$$(2,'0.017*"dryer" + 0.015*"hair" + 0.013*"one"')] \tag{31}$$

The equivalent cloud image is shown as figure 5:

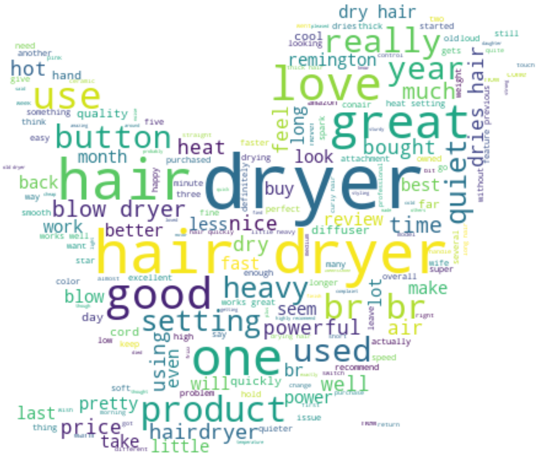


Figure 5: Cloud image of b003v264ww

The semantic similarity between the product review keywords and the data set keywords in the period 2012.12-2013.1 is as high as 70%, so we consider this product as a potential successful product.

3.5 Correlation Analysis

After the emotional tendency of each comment is quantified, the relationship curves of product star rating and time, review value and time, sales volume and time are calculated separately by month.

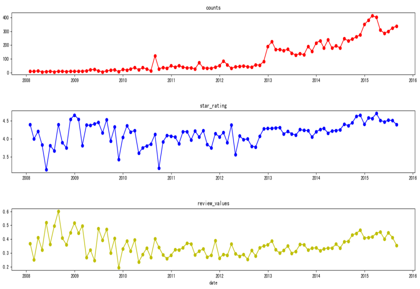


Figure 6: The relationship curves of hair dryer

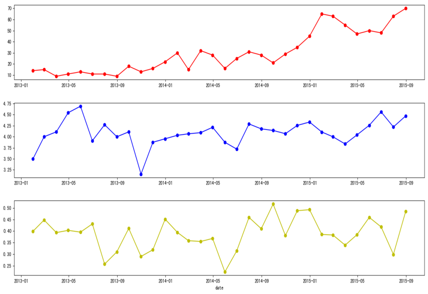


Figure 7: The relationship curves of microwave

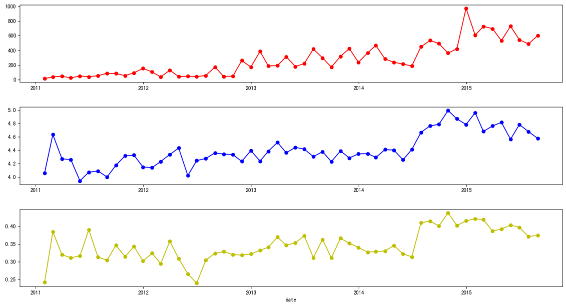


Figure 8: The relationship curves of pacifier

However, it is not very intuitive to determine which one of the star rating and review value has a greater impact on the sales volume through this curve. Therefore, on the basis of this curve, we respectively analyze the correlation between the star rating, review value and sales volume, and obtain the correlation heatmap:

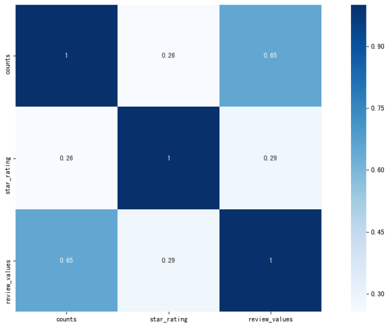


Figure 9: Correlation heatmap

From this we can get: the correlation between star rating and sales volume is weak, while the correlation between review value and sales volume is relatively strong. In addition, there is a strong correlation between star rating and review value, which indicates that the higher the star rating given to the product by the user, the more positive the comment emotional tendency is,that is, the higher the satisfaction with the product is. On the whole, there is a consistency between star rating and comment emotional tendency.

Cross-correlation analysis, also known as cross-correlation, refers to the degree of correlation between two time series at any two different times. Assuming that there are time series X_t and Y_t respectively, the correlation between X at time t and Y at time $t+n$ is n order cross-correlation. The specific formula is as follows:

$$ccf_n = f\left(X_t, Y_{t+n}\right) = r_{X,Y_{t+m}} = \frac{\sum\left(X_t - \bar{X}_t\right)\left(Y_{t+n} - \bar{Y}_{t+n}\right)}{\sum\left(X_t - \bar{X}_t\right)^2\left(Y_{t+n} - \bar{Y}_{t+n}\right)^2} \tag{32}$$

Where, function f is a function to calculate the correlation coefficient, and the value of the correlation number of lagging order n can be calculated by the above equation. Take the star rating and review value data of three data sets respectively, and use python for cross-correlation analysis. The effect is shown as follows:

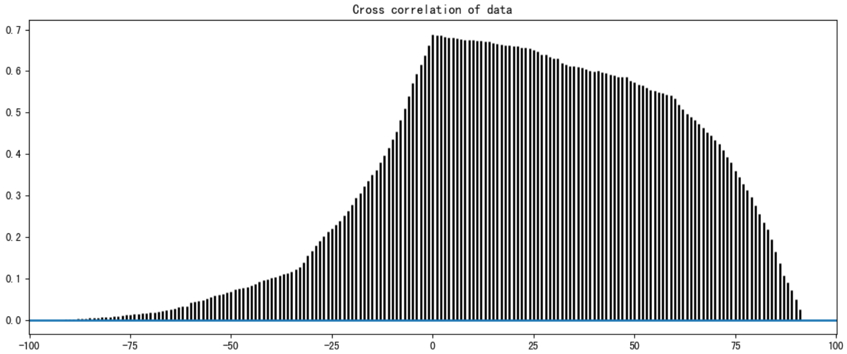


Figure 10: The hair_dryer's correlation diagram of star_rating and review value

For hair_dryer, when there is no delay (i.e., Lag=0), the correlation number reaches the maximum, which is close to 0.7. The two time series are positively correlated and moderately correlated. When the delay is 25 orders, the correlation number decreases when relative Lag=0, but it is still greater than 0.6. When the delay is 50 orders, the correlation number is still greater than 0.5, indicating that the two time series are moderately positive correlated, that is, the star rating will trigger more comments.

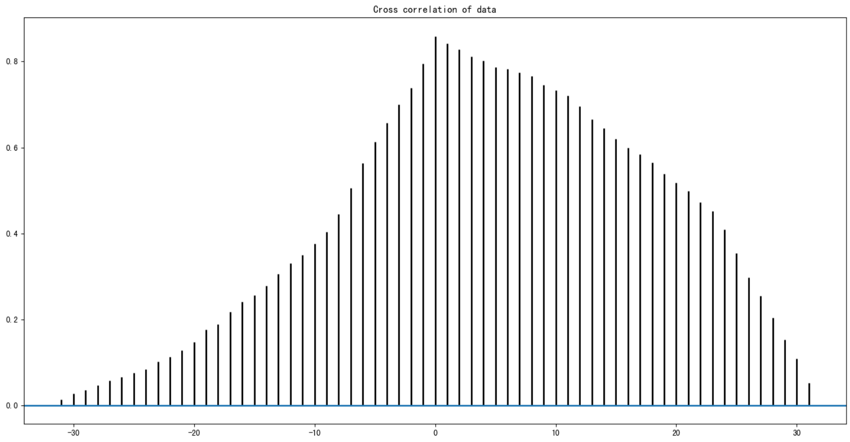


Figure 11: The microwave's correlation diagram of star_rating and review value

For microwave, when there is no delay (i.e., Lag=0), the correlation number is the largest, greater than 0.8, and the two time series are positively correlated and strongly correlated. When the delay is of order 20, the number of correlation is greater than 0.5, indicating a moderate positive correlation between the two time series.

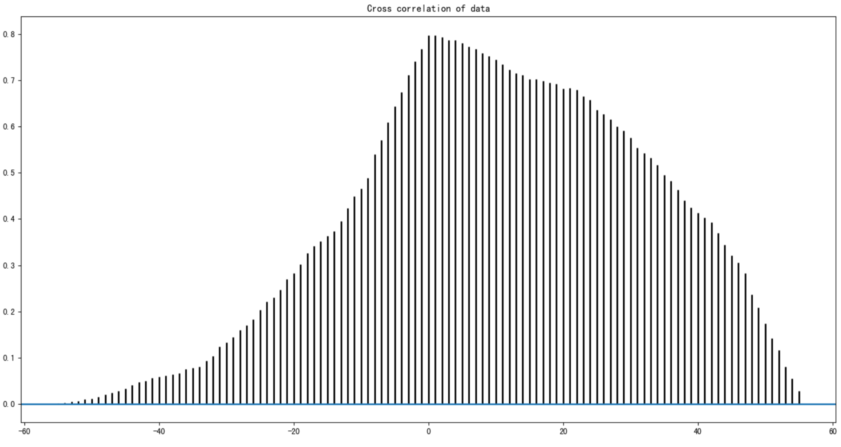


Figure 12: The pacifier′s correlation diagram of star_rating and review value

For pacifier, when there is no delay (i.e., Lag=0), the correlation number reaches the maximum, which is close to 0.8, and the two time series are positively correlated and moderately correlated. When the delay is of order 30, the correlation number is greater than 0.5, indicating a moderate positive correlation between the two time series.

To sum up, star_rating triggers more comments.

3.6 Review Keyword and Rating Level Correlation Analysis

We use LDA model to extract the theme and the word frequency statistics of each star level find that there are hair and dryer the subject word of each star rating. Through these we can see to the customer pays more attention to the practicality of hair dryer.

The following is the rating of the theme:

- **Level 1**:

$$[(0,'0.012*"money"+0.011*"dryer"+0.011*"product"'), \tag{33}$$
$$(1,'0.026*"hair"+0.022*"dryer"+0.013*"one"'), \tag{34}$$
$$(2,'0.038*"dryer"+0.024*"hair"+0.020*"one"')] \tag{35}$$

- **Level 2**:

$$[(0,'0.035*"dryer"+0.029*"hair"+0.014*"one"'), \tag{36}$$
$$(1,'0.021*"dryer"+0.017*"month"+0.015*"star"'), \tag{37}$$
$$(2,'0.032*"hair"+0.028*"dryer"+0.012*"work"')] \tag{38}$$

- **Level 3**:

$$[(0,'0.039*"hair"+0.036*"dryer"+0.016*"one"'), \tag{39}$$
$$(1,'0.025*"star"+0.025*"three"+0.013*"work"'), \tag{40}$$
$$(2,'0.024*"dryer"+0.016*"hair"+0.011*"cord"')] \tag{41}$$

- **Level 4**:

$$[(0,'0.026*"star"+0.021*"four"+0.019*"work"'), \tag{42}$$
$$(1,'0.016*"dryer"+0.010*"unit"+0.009*"hair"'), \tag{43}$$
$$(2,'0.049*"hair"+0.048*"dryer"+0.016*"good"')] \tag{44}$$

- **Level 5**:

$$[(0,'0.048*"great"+0.040*"star"+0.039*"five"'), \tag{45}$$
$$(1,'0.020*"one"+0.013*"year"+0.012*"hair"'), \tag{46}$$
$$(2,'0.082*"hair"+0.070*"dryer"+0.022*"love"')] \tag{47}$$

The word cloud map of each level is shown in the appendix.

There is a money keyword in the one-star review, what shows that the price of the product is difficult to be accepted by the customer who gives a star review. May be that the product life is short ,so the Two-star reviews contain the month keyword. Four - and five-star reviews include words such as great, love, and good, so the specific quality descriptors of text-based reviews strongly associated with rating levels.

4 Model Evaluation

4.1 Sensitivity Analysis

In order to further study the correlation and degree of the identity coefficient v on the comprehensive reputation score, we carried out sensitivity analysis on the model: Change the weight of different identities in the formula, and then see if it has a significant impact on the reputation score.

Table 4: The change of identity coefficient

group	vine	verified_purchase	weight
1	N	Y	0.1
	Y	N	0.3
	Y	Y	0.6
2	N	Y	0.3
	Y	N	0.3
	Y	Y	0.4

According to the above figures, we find that the change of internal value of identity coefficient has little impact on the comprehensive evaluation results of reputation.

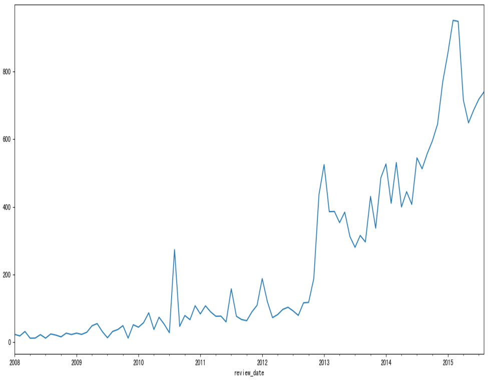


Figure 13: Group 1

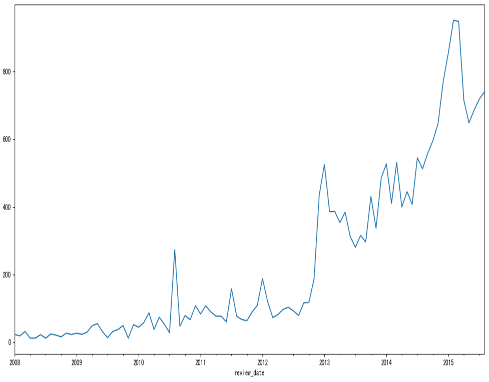


Figure 14: Group 2

4.2 Conclusions and Generalization

Our team is asked by sunshine as consultants to provide them with an online sales strategy and identify potentially important design features. After data processing and modeling, we successfully complet the task. First, we establish the SO-PMI based Review Quantization model to quantify the emotional tendency comments. After defining the concepts of review value and identity coefficient successively, we establish the Comprehensive Rating Model of Product Reputation. In addition, through correlation analysis, we find that the correlation between review value and sales volume is greater than that between rating and sales volume, indicating that reviews are more important than ratings.

Then, we used the least square method to fit the reputation and time, and obtained the functional relationship of hair_dryer, microwave and pacifier. RMSE is about 0.14, 0.16 and 0.09 respectively, indicating a good fitting effect. By analyzing the fitting curve, we get that the reputation of hair_dryer and microwave is stable in the later period, while pacifier fluctuates greatly. Also, we use LDA to extract keywords from the comment sets in these time periods. Then, use PMI to calculate the semantic similarity between words, and predict products success or failure by combining the semantic similarity and the sales volume. Take hair_dryer set as example, we find product_id =b003v264ww is a potential successful product.

In addition, we conduct cross correlation analysis of star ratings and review values, which indicate that the rating will trigger more reviews. Based on the theme extraction and word frequency statistics of each star levels review by LDA, we find that some specific words in the comments will affect the star rating.

Finally, we made a new product sales plan for Sunshine Company based on the above analysis.

All the models established in this paper are applicable to the online shopping retail platform with online comment and rating functions, so as to facilitate the company of the system to track and analyze the user data and make timely response or planning plans. By converting star rating into thumb up number and forwarding number, it can be applied to the improvement of the content produced by bloggers on social network platforms such as Facebook, Twitter.

4.3 Strengths and Weaknesses

Strengths:

- For such a typical big data problem, we carried out data processing, including the selection of indicators and the processing of meaningless data. Through this step, the quality of data is greatly improved, which makes the problem solving more efficient and convenient.
- We normalized the data of different indexes to enhance the accuracy of the model results.
- We compared the results of the reputation comprehensive score model with the sales volume to verify the reliability of the model.
- For the indicator of identity coefficient v, we conducted sensitivity analysis on the reputation comprehensive score results to further determine the impact of vine and verified_purchase on reputation.
- We modified the SO-PMI model to make the results bounded

Weaknesses:

- In the review quantization model, the existence of negative words(such as "no", "not") in the text is not considered, so the quantization results have certain errors.
- In the process of data processing, text data suspected of containing advertisements were not removed.

5 A Letter to the Marketing Director of Sunshine Company

Dear Marketing Director of Sunshine Company,

We are honored to inform you our achievement after performing data analysis and modeling.

First, on the basis of the comment quantization model based on SO-PMI and the comprehensive evaluation model of product reputation, through correlation analysis, we find that the correlation coefficient between review value and sales volume of the three data sets is greater than that between rating and sales volume, indicating that reviews are more important than ratings.

Second, the sensitivity analysis of the identity coefficient shows that the change of its internal value has no obvious influence on reputation, which indirectly proves the accuracy of the model.

In addition, we use the least square method to fit reputation and time respectively, and obtain the relationship of multiple linear regression functions. The result shows that the reputation of hair_dryer and microwave was stable in the later stage, while pacifier fluctuated greatly. Then, according to the established product prediction model based on LDA, the product with product id of b003v264ww is a potential successful product by calculating the hair_dryer data set.

Moreover, we conduct the cross correlation analysis of star ratings and review value of three data sets. The result shows that the number of correlation between the two time series is greater than 0.5 when they are at least delayed to the 20th order, indicating that the rating will trigger more comments.

Finally, we combine the word cloud map of all data sets with the analysis of modeling results to provide an online sales strategy for your company:

- When a customer is shopping, reviews have a greater influence on the reference meaning of his purchase decision than ratings. So, we suggest using reviews as the measurement of the company's main data tracking.
- The reputation of hair_dryer and microwave was stable in the later stage, while pacifier fluctuated greatly. Your company can predict the future trend of each product according to the product prediction model established by us.
- According to the word cloud map, we find that users prefer folding style hairdryers with strong wind and suitable for traveling, conscious of the size of the microwave and prefer a cute pacifier. All these are important bases for your company to optimize the new product features.
- It is also worth noting that ratings attract more reviews.

We hope our work will be helpful to your company to improve your online sales strategy in the future. Thank you

Best regards,

Team #2020230

References

[1] Gao YifanYu WenzheChao Pingfuet alAnalyzing reviews for rating prediction and item recommendationJ.Journal of East China Normal University (Natural Science) .20153:80-90.

[2] Shi LinglingComprehensive Product Score Model Research Based on Mining Online Comments[D], Hangzhou Dianzi University,2016

[3] PD.Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[J]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002(12), 417-424.

[4] HU Zhong-kai,ZHENG Xiao-lin ,WU Ya-feng ,CHEN De-ren,Product recommendation algorithm based on users reviews mining [J]. Journal of zhejiang university (engineering), 2013,47(08):1475-1485.

[5] SU Qiao ,XU Xingyong, CHEN Guangquan, FU Tengfei, LIU Wenquan, Frequency and Hysteresis of Groundwater Levels Influenced by Tides [J]. Ocean Development and Management,2018,35(10):79-83.

[6] TAN YunzhiHANG MinLIU YiqunMA Shaoping, Collaborative Recommendation Framework Based on Ratings and Textual Reviews,[J].Pattern Recognition and artificial Intelligence ,2016,29(04):359-366.

[7] Hou Yinxiu,Li Weiqing,Wang Weijun,Zhang Tingting, Personalized Book Recommendation Based on User Preferences and Commodity, [J]. Data analysis and knowledge discovery ,2017,1(08):9-17.

Appendices

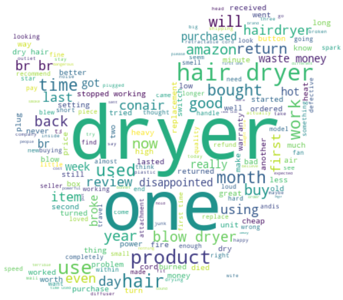


Figure 15: star level 1

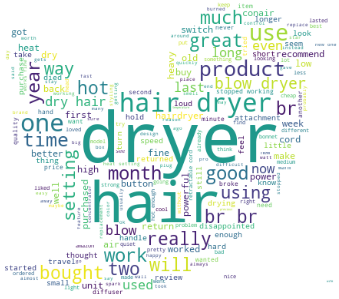


Figure 16: star level 2

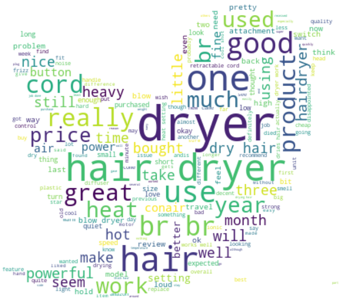


Figure 17: pstar level 3

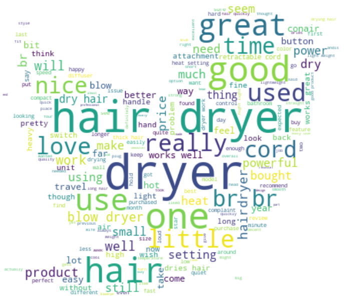


Figure 18: star level 4

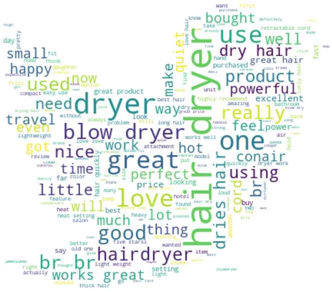
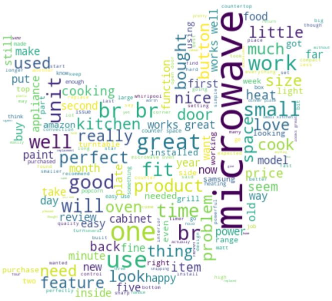


Figure 19: star level 5



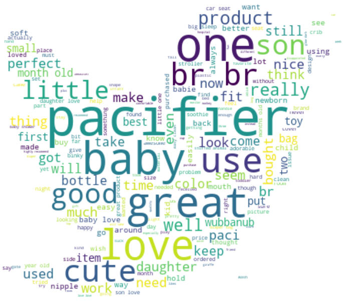


Figure 20: microwave cloud map

Figure 21: pacifier cloud map

```
#data3_Sentiment_analysis.py
from textblob import TextBlob
import pandas as pd

df1_new = pd.read_csv('Problem C_1/df1_new.csv', index_col=0)
comments = df1_new['review']
review = []
review = comments.apply(lambda x: TextBlob(x).sentiment)
review = pd.DataFrame(review)
review.columns = ['sentiment']
review['sentiment'].apply(pd.Series)
review[['polarity', 'subjectivity']] = review['sentiment'].apply(pd.Series)
df1_new_Sa = pd.concat([df1_new, review], axis=1)
df1_new_Sa['helpful_polarity'] = df1_new_Sa['polarity']*df1_new_Sa['
   helpfulness_rating']
df1_new_Sa = df1_new_Sa.drop(['sentiment', 'subjectivity', 'polarity', '
   helpfulness_rating'], axis=1)
df1_new_Sa.to_csv('Problem C_1/df1_new_Sa.csv')

#data2_visualization.py
import seaborn as sns
import PIL
import numpy as np
import pandas as pd
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score
from wordcloud import STOPWORDS, WordCloud
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

# read csv
```

```
df1 = pd.read_csv(’Problem C_1/hair_dryer.csv’, encoding=’utf-8’, delimiter
    ="\t")

# delete (df2[’verified_purchase’] == ’n’) & (df2[’vine’] == ’n’)
df1 = df1[~((df1[’verified_purchase’] == ’n’) & (df1[’vine’] == ’n’))]
df1 = df1.drop([’marketplace’, ’product_title’, ’product_category’, ’
    product_parent’], axis=1)

# add df[’helpfulness_rating’]
df1.insert(4, ’helpfulness_rating’, df1[’helpful_votes’].astype(’float’) /
    df1[’total_votes’].astype(’float’))

df1[’helpfulness_rating’] += 1

# date before 2008
df1[’review_date’] = pd.to_datetime(df1[’review_date’])
df1 = df1.sort_values(by=’review_date’)
df1[’year’] = df1[’review_date’].dt.year
groups_y = df1.groupby([’year’]).count()
# df2.index = pd.DatetimeIndex(df2["review_date"])
df1_new = df1[df1[’review_date’] >= ’2008-1-1’].drop([’year’], axis=1)

# merge review
df = df1_new[’review_body’].str.cat(df1_new[’review_headline’], sep=’ ’)
df1_new.insert(7, ’review’, df)

# df1_new_cnt_sum.plot(kind=’line’)
df1_reputation_sum = pd.read_csv(’Problem C_1/df1_reputation_sum.csv’,
    index_col=0).reset_index()
df1_reputation_sum[’review_date’] = pd.to_datetime(df1_reputation_sum[’
    review_date’])
df1_new_cnt_sum = df1_new_cnt_sum.reset_index()
df1_new_M = pd.merge(df1_reputation_sum, df1_new_cnt_sum, how=’left’, on=’
    review_date’)
# pearson
print(df1_new_M[[’star_rating’, ’helpful_polarity’, ’counts’]].corr())
#df1_new_M.corr().to_csv(r’E:\Myfiles’)
# word_cloud
text = ("".join(i for i in df1_new[’review’]))
stopwords = set(STOPWORDS)
image1 = PIL.Image.open(’Problem C/1.png’)
MASK = np.array(image1)
wordcloud = WordCloud(max_words=200, stopwords=stopwords, background_color=
    ’white’, mask=MASK).generate(text)
image = wordcloud.to_image()
image.show()
wordcloud.to_file(r’E:\Myfiles\’)
#LDA1.py
df = pd.read_csv(’Problem C_1/df1_new.csv’, index_col=0)
df.index = pd.DatetimeIndex(df["review_date"])
df_s = df[’2013-1’:’2013-2’]
df_f = df[’2015-3’:’2015-4’]
df_pick = df[df[’product_id’] == ’b003v264ww’]
df_1 = df[df[’star_rating’] == 1]
df_2 = df[df[’star_rating’] == 2]
df_3 = df[df[’star_rating’] == 3]
df_4 = df[df[’star_rating’] == 4]
df_5 = df[df[’star_rating’] == 5]
```

```
'''def clean(doc):
    """
    @param doc:
    @return:
    """
    stop_free = " ".join([i for i in doc.lower().split() if i not in stop])
    punc_free = ''.join(ch for ch in stop_free if ch not in exclude)
    normalized = " ".join(lemma.lemmatize(word) for word in punc_free.split
        ())
    return normalized


# Set the stop word
stop = set(stopwords.words('english'))
exclude = set(string.punctuation)
lemma = WordNetLemmatizer()

#  import a doc
doc_complete = df_1['review']
doc_clean = [clean(doc).split() for doc in doc_complete]

# get a dict
dictionary = corpora.Dictionary(doc_clean)
doc_term_matrix = [dictionary.doc2bow(doc) for doc in doc_clean]

# LDA model
Lda = gensim.models.ldamodel.LdaModel
lda_model = Lda(doc_term_matrix, num_topics=3, id2word=dictionary, passes
   =50)

# print
print(lda_model.print_topics(num_topics=3, num_words=3))'''

# word_cloud
text = ("".join(i for i in df_5['review']))
stopwords = set(STOPWORDS)
image1 = PIL.Image.open('Problem C/1.png')
MASK = np.array(image1)
wordcloud = WordCloud(max_words=200, stopwords=stopwords, background_color=
   'white', mask=MASK).generate(text)
image = wordcloud.to_image()
# image.show()
wordcloud.to_file(r'E:\Myfiles')
```