| Problem Chosen | 2020 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2000905 |

# A study of factors that can develop products


## Summary

With the gradual improvement and development of online shopping platforms, different types of online comments, opinions and recommendations are considered to be the most influential communication channels between service providers and consumers as well as between consumers. User-generated content such as star ratings, reviews and help ratings can help companies gain insight into the market and make sound decisions.

**First:** theme extraction model. For the sake of seeking help star ratings, reviews, ratings, the relationship between the set meaningful measurement model, help company to understand the user requirements, this article embarks from the users' comments, the commentary subject extraction model based on the LDA is established, by using the model found three data on the theme of the user comments. The theme of the hair dryer is mainly distributed in price and blow-drying speed, the theme of the microwave oven is mainly distributed in size, price and heating time, the main theme of the baby nipple is distributed in the shape of the nipple and the baby's love degree. These factors can be taken into account when making improvements to the product.

**Second:** establishment of evaluation model. In order to help companies measure the market of commodities according to the data of rating and review, this paper establishes an evaluation model based on AHP. The weight of star is 0.7306, the weight of comment emotion score is 0.1884 and the weight of comment subjectivity score is 0.081. The evaluation model thus established is:

$$y = 0.7306x_1 + 0.1884x_2 + 0.181x_3$$

In addition, in order to verify the stability of the model, the data in the data set of the infant pacifier were verified with the model, and the model was found to be of good stability

**Third:** based on the above model, this paper analyzes the potential good and bad of commodities from four perspectives: time, text, rating and comment. We found that with the change of time, the reputation of goods is constantly improving, and began to stabilize in 2013. The improvement of product reputation will also have a positive impact on the sales of products. There is a positive correlation between stars and reviews.

Based on our conclusions, we wrote a letter to the director of marketing of sunshine company.



**Keywords: LDA , AHP , Product Reputation Model, Evaluation Model, Topic Extraction model**

# Contents

# 1. Introduction

## 1.1 Literature Review

As a way of evaluation and feedback, user-generated content has attracted more and more attention from scholars in recent years. It has been studied for years to explore the impact of online customer reviews on the business management of enterprises. For example, Victorhostudied online reviews in 2017 and found that there are three actions that managers must take in response to online reviews: acknowledging the problem, expressing feelings and thanking the reviewer.[1] Hatzivassiloglou et al. attempted to establish a lexicographical emotion dictionary in 1997, based on which they conducted text emotion analysis, and the accuracy of the judgment results reached 82%.[2]. Wiebe et al. (2001) distinguish between subjective and objective texts and carry out emotional analysis on subjective texts. Pang, et al. (2002) is introduced into data mining algorithm in the naive bayes and maximum entropy model and support vector machine (SVM) model to text sentiment analysis of movie reviews, in China, 2011 ZhaiLiKong in studying the impact of online reviews on consumer purchase intention, found that the more the number of online reviews, online reviews the intensity, the greater the impact on consumer purchase intention is, the greater the at the same time, the influence of the influence of the negative comments than positive comments, etc

In recent years, with the gradual improvement of big data means and methods, it has become a trend to use big data for text emotion research. However, there are still some problems in e-commerce online comments based on text emotion analysis: first, most of the studies only focus on users' online comments, without the comprehensive study of other user-generated content; second, the comprehensive analysis of time series is not integrated in the study of appropriate marketing strategies.

Therefore, this paper USES the data provided by sunshine company to improve the traditional model of text analysis, and establishes a text analysis model based on LDA to help sunshine company make better decisions by analyzing the market user feedback of commodities.

## 1.2 Problem Summary

Sunshine plans to launch and sell three new products in the online marketplace: microwave ovens, baby pacifiers and hairdryers. They hired our team as consultants to

identify key patterns, relationships, metrics, and parameters in past customer ratings and reviews related to competing products, and to develop mathematical models to develop a reasonable sales strategy to help sunshine succeed in three new online markets.

Using the three data sets provided by sunshine data center, we propose the following research questions:

1.How to find meaningful quantitative or qualitative models in star rating, comment, and help rating to determine the relationship between the three?

2. How can time-based measurement patterns in each dataset be used to determine changes in demand for products in the market?

3. Do specific star ratings incite more reviews? For example, are customers more likely to write some type of review after seeing a series of low star ratings?

4. How can text-based measures be combined with rate-based measures to best indicate potential success or failure of a product?

Are text-based comments closely related to ratings?

Finally, we wrote a letter to sunshine based on the results of our research, in which we introduced the team's research results and the specific reasons for the recommendation.

# 2. Assumptions

1. Assuming that all purchase behaviors are real and effective, and there is no false behavior to improve the sales volume and credit,

   Reason: this behavior is not helpful in determining the market movements of commodities

2. Assume that merchants and platforms don't ask customers to change or remove comments that are not favorable to them

   Reason: in order to be effective, all comments should be true and valid

3. Assume that there are no malicious bad reviews from competitors

   Reason: the malicious bad comments of competitors will cause the model to produce errors in the evaluation of goods

4. Assume that all user comments are unbiased

   Reason: biased comments are not meaningful to our research

# 3. Symbol Description

| Symbol | Meaning |
|---|---|
| u | user |
| i | product |
| $P_u$ | user preference |
| $ave_{star\_day}$ | the average daily emotional score |
| $sum_{star}$ | the total number of stars scored on that day |
| $sum_{pos}$ | the sum of subjective scores of the day |
| dailysales | daily sales |

Note: some explanations are given in the article.

# 4. Data

## 4.1 Data Sources

Our model is derived from three data files (hair_dryer.tsv, developer.tsv, pacifier.tsv) provided by the sunshine corporation data center. The data represent ratings and reviews from customers of microwave ovens, baby pacifiers and hair dryers sold on amazon's marketplace over the period shown in the data.

### 4.1.1 Data Cleaning

The supplied data set is partially missing and contains data for some unrelated products. We do the following to clean up the dataset.

-- delete and sunshine company of three types of products sold to unrelated product data (such as washing machines, refrigerators) .

-- delete the comment without purchase behavior data. (for example varified_pucharse = NO but with a five-star high praise)

-- delete the item has nothing to do with the model of redundant. (product ID, for example, the origin of a product)

# 5. Part1

Based on the LDA model, this paper extracts the textual theme of user comments to help the company further understand the market trend and product development trend. After on the basis of the LDA model based on Logistic regression to the prediction of the score to the user, and compared with the actual score, verifies the validity of the LDA model.

## 5.1 Model Description

### 5.1.1 Model of LDA

Potential dilip clay (LDA) is a kind of probability distribution model generation model. LDA think every document $d$ is considered to conform to the K - dimensional topic distribution $\theta$, and each word in the document has the probability and the possibility of being subject $k$. LDA model by introducing the subject dimension between the text and word, dimension reduction on the vector space. In this article the LDA model used for the analysis of user reviews information, find the comments potential topics.

As shown in figure 5.1, a document can be regarded as an ordered sequence of N words, and a document set contains M documents. $\alpha$ and $\beta$ theta theme in the text of the distribution and theme distribution of the middle term of phi super parameters, subject to prior dirichlet distribution, including z theme.

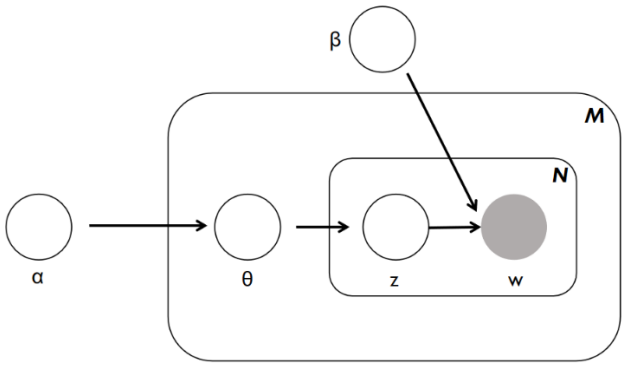


Figure 5.1 A Graphical Representation of the Model

The unstructured comment text contains the user's preferences and opinions on different subject dimensions of the product, and these potential topics can well reflect the potential factors of the user rating. In order to discover potential topics, the highly recognized text analysis and feature extraction tool, LDA model, was applied to comment text analysis.

This article will each user $u$ for goods $i$ give comments $d_{ui}$ as a document. Application of essays$\{d_{ui}\}$ the LDA model, $\theta_{ui}$ said $d_{ui}$ generated $K$ d theme distribution. User $u$ all comments collection defined as $D_u$, $D_i$ is commodity $i$ to obtain the set of all comments. Each user $u$ (or goods $i$) corresponding to the preference $p_u$ (or portraits$q_i$). Given a user $u$, define user preferences $p_u$ for

$$p'_{uj} = \frac{\sum_i \theta_{uij}}{|D_u|}, p_{uj} = \frac{p'_{uij}}{\sum_i p'_{uij}}, j \in \{1, \cdots, k\}. \quad (1)$$

The $p_u = (p_{u1}, p_{u2}, \cdots, p_{uk})$, $p_{uj}$ is user distribution of u in the first *j* a theme, $\theta_{uij}$ is $d_{ui}$ in the distribution of the first *j* a theme。Similar approach to defining goods *i's* patren $q_i$

$$q'_{uj} = \frac{\sum_i \theta_{uij}}{|D_u|}, q_{uj} = \frac{q'_{uij}}{\sum_i q'_{uij}}, j \in \{1, \cdots, k\}. \quad (2)$$

In simple terms, $p_u$ and$q_i$ user $u$ and commodity $i$ all comments theme distribution normalization results.

## 5.1.2 Logistic Regression Score Prediction Model

Given user $u$ and commodity $i$, to predict the user $u$ score for goods $i$ may give $\hat{r}_{ui}$, compare $\hat{r}_{ui}$ with the user for the goods from the actual score, to verify the effectiveness of the LDA model. When predicting the score, consider the topic dimension found in the comment text as a potential factor affecting the score. Score prediction model using linear regression and logistic regression model to establish score $\hat{r}_{ui}$ and review $d_{ui}$ theme distribution of theta $\theta_{ui}$ relationship. As a standard regression analysis model, linear regression has been widely applied in practice [5]. It is assumed that there is a linear relationship between dependent variables and independent variables, and the parameters are strict but easy to fit. The score prediction function of the multinomial linear regression method is defined as

$$\hat{r}_{ui} = W^T \theta_{ui} + \varepsilon_{ui} \quad (3)$$

In $W = (W_1, \cdots, W_k)$, $W_j$ is $j$th weight of topic, $\varepsilon_{ui}$ is error variance。

By taking the probability score as the value of the dependent variable, the relationship between an absolute dependent variable and K independent variables can be measured [6]. In other words logistic regression is to establish the relationship between prediction score and k-dimensional topic distribution. Hypothesis scoring $\hat{r}_{ui} \in 1,2 \cdots, N$,establish a multinomial logistic regression

$$P_r(\hat{r}_{ui} = n) = \frac{e^{\beta_n^T \theta_{ui}}}{1 + \sum_{n'=1}^{N-1} e^{\beta_{n'}^T \theta_{ui}}}, \quad (4)$$

$$P_r(\hat{r}_{ui} = N) = \frac{1}{1 + \sum_{n'=1}^{N-1} e^{\beta_{n'}^T \theta_{ui}}} \quad (5)$$

In $n = 1,2 \cdots, N-1$, $\beta_n = (\beta_{n1}, \beta_{n2}, \cdots, \beta_{nk})$ is a set of weight.

The score predicts the subject distribution $\hat{\theta}_{ui}$ of a given user $u$ and the reviewed item $i$, based on the user preference $p_u$ and the item portrait $q_i$

$$\theta'_{uij} = p_{uj} q_{ij}, \hat{\theta}_{uij} = \frac{\theta'_{uij}}{\sum_j \theta'_{uij}}, j \in \{1, \cdots, k\}. \quad (6)$$

## 5.2 Data Testing of Model

According to the results of LDA extraction, we extracted three main features from the comments for each data set, such as product quality characteristics, price characteristics, user experience characteristics, etc

In this paper, the mean square error (MSE) is used to measure the error between the prediction score and the actual score, and the number of times that the prediction score is consistent with the actual score is calculated. It is found that the consistency probability is about 0.7, and the experimental results are better.

# 6. Part2

## a. Model Extraction Based on LDA Theme

To according to the ratings and comments to determine can provide information of data measurement, this paper USES analytic hierarchy process (AHP) obtained the ratings and comments of weights, and then get the total score of the ratings and comments, the total score of the company can provide a kind of data measurement method to judge the merits of the merchandise so this article selects hair dryer data set as sample to establish the evaluation model based on AHP concrete implementation method is as follows

- **Step1: Data Processing**

In order to quantify users' comments on commodities, this paper first USES Python's textbolb to score the comments on each commodity, and obtains the subjective score and emotional score of the commodity. In order to exclude some extreme comments, our team calculated the daily sales volume of the products, and then obtained the daily average star score, the daily average subjective score and the daily average emotional score. The calculation formula is as follows:

$$ave_{\mathrm{star_d}} = \frac{\mathrm{sum_{star}}}{\mathrm{Dailysales}} \quad (7)$$

$$ave_{\mathrm{pos}} = \frac{\mathrm{sum_{pos}}}{\mathrm{Dailiysalse}} \quad (8)$$

$$ave_{\mathrm{emo}} = \frac{\mathrm{sum_{emo}}}{\mathrm{Dailiysalse}} \quad (9)$$

$ave_{star_d}$ represents the average daily star score,

$ave_{pos}$ represents the daily average subjective score,
$ave_{emo}$ represents the average daily emotional score ,
$sum_{star}$ represents the total number of stars scored on that day,
$sum_{pos}$ represents the sum of subjective scores of the day,
$sum_{emo}$ represents the total emotional score of the day, $Dailiysalse$ representative daily sales
See table for specific results 6.1

Table 6.1 Specific Results

| Average emotion score | Average subjective score | Average star score |
|---|---|---|
| 0.22 | 0.53 | 3.67 |
| 0.11 | 0.57 | 4.50 |
| 0.22 | 0.48 | 3.50 |
| 0.30 | 0.45 | 4.00 |
| 0.20 | 0.65 | 3.00 |
| 0.34 | 0.44 | 4.50 |
| 0.37 | 0.70 | 5.00 |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| 0.49 | 0.59 | 4.80 |
| 0.20 | 0.50 | 2.00 |
| 0.42 | 0.73 | 5.00 |
| 0.04 | 0.21 | 3.00 |
| 0.31 | 0.60 | 5.00 |
| 0.04 | 0.50 | 4.00 |
| 0.12 | 0.46 | 3.00 |

- **Step 2 Construct the judgment matrix A**
  $a_{ij}$ as shown in table 6.2

Table 6.2 The Value of the Criteria Table

| scale | Implication |
|---|---|
| 1 | Means that two factors are of equal importance |
| 3 | The former is slightly more important than the latter |
| 5 | The former is more important than the latter |
| 7 | The former is more important than the latter |
| 9 | The former is more important than the latter |
| 2,4,6,8 | Represents the intermediate value of the above ajia cent judgment |
| reciproca l | If the ratio of the importance ϕf factor |

Therefore,

$$A = \begin{pmatrix} 1 & 5 & 9 \\ 1/5 & 1 & 5 \\ 1/9 & 1/5 & 1 \end{pmatrix} \tag{10}$$

- **Step 3 Verify the consistency of the judgment matrix**
  **Step 3.1 calculate the consistency index CI**

$$CI = \frac{\lambda_{max} - n}{n-1} \tag{11}$$

Therefore,

$$CI = 0.0324$$

**Step 3.2 calculate the consistency ratio CR**

$$CR = \frac{CI}{RI} \quad (12)$$

It can be calculated from equation (12) $CR = 0.0624 < 0.1$, So this paper holds that the consistency of judgment matrix is acceptable in conclusion, we concluded that the weight of star 0.7306, comments emotional weight of 0.1884, comment on the weight of subjective score of 0.081.

Therefore, the model can be expressed as:

$$y = 0.7306x_1 + 0.1884x_2 + 0.181x_3 \quad (13)$$

Star score, including $x_1$ said $x$ emotional score, $x_2$said comments, $x_3$ said the subjectivity, y represent product reputation score.

In order to verify the reliability of the model, this paper selects a data set of infant pacifiers to verify the model, and finds that the model's sales evaluation of infant pacifiers is roughly consistent. Therefore, this model has certain reliability and can be applied to product evaluation.

## b. Product Reputation Model Based on AHP

In order to explore the changing trend of product reputation over time, this paper separately calculates the reputation of the annual sales data of baby pacifier microwave oven of hair dryer. Draw a line graph in Excel(Figure 6.1 ~ 6.3).

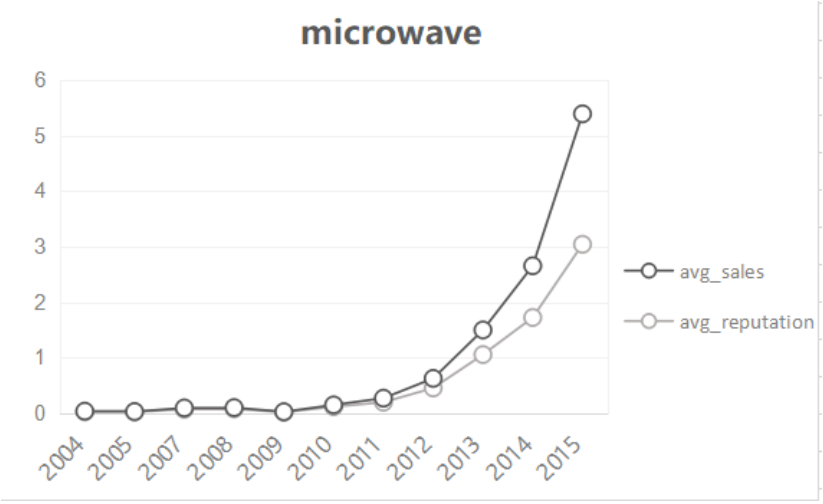


Figure 6.1 Microwave Average Product Reputation and Sales Change Chart

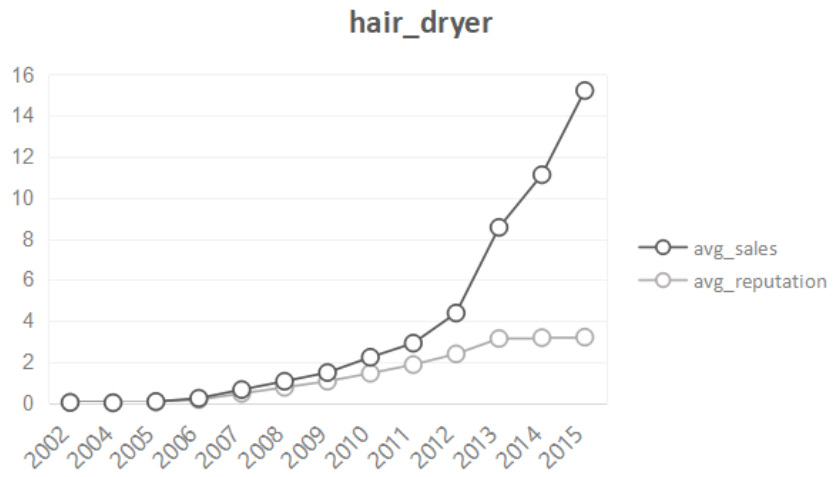


Figure 6.2 Hair dryer Average Product Reputation and Sales Change Chart

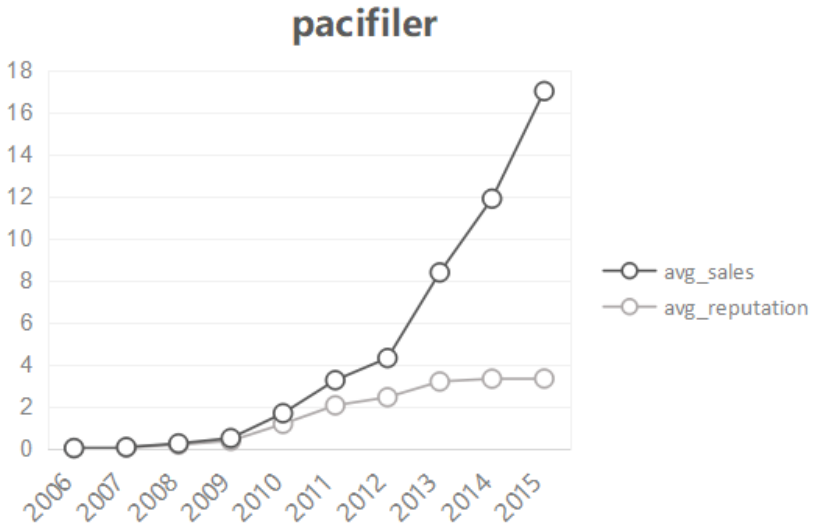


Figure 6.3 Pacifiler Average Product Reputation and Sales Change Chart

## c. Product Decision Optimization Model Based on Correlation Analysis

According to the time-based reputation evaluation model established in question b, we can get the average reputation growth rate of the three products, as shown in the figure below.

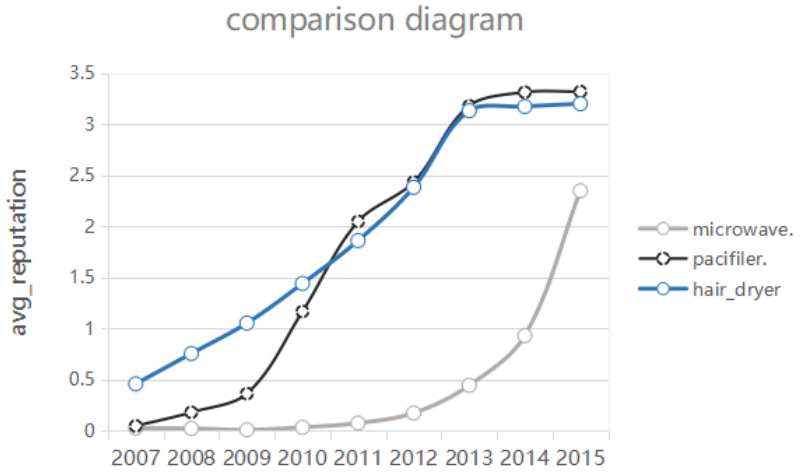


Figure 6.4 Product potential success (failure) rate chart

From this chart, we can see that the growth rate of reputation of microwave oven and pacifier is still rising. In other words, the sales of these two products are still good, which can be regarded as potential successful products. While the hair dryer has a slight decline or flat, which shows that the reputation of hair dryer may not grow or even decline, can be seen as a potential failure of the product.

To further analyze the reasons for the formation of potentially successful/filed products, In this paper, SPSS software was used to conduct binary correlation analysis on several indicators (average sentiment analysis score, average star score, reputation and average subjective score) of the three products, and three Pearson coefficient diagrams were obtained (as shown in the figure below).

**Correlations**

| | | Average emtion score | Average subjective score | Average star rating | Reputation |
|---|---|---|---|---|---|
| Average emtion score | Pearson Correlation | 1 | .419** | .315** | .366** |
| | Sig. (2-tailed) | | .000 | .000 | .000 |
| | N | 1669 | 1669 | 1669 | 1669 |
| Average subjective score | Pearson Correlation | .419** | 1 | .156** | .192** |
| | Sig. (2-tailed) | .000 | | .000 | .000 |
| | N | 1669 | 1669 | 1669 | 1669 |
| Average star rating | Pearson Correlation | .315** | .156** | 1 | .998** |
| | Sig. (2-tailed) | .000 | .000 | | .000 |
| | N | 1669 | 1669 | 1669 | 1669 |
| Reputation | Pearson Correlation | .366** | .192** | .998** | 1 |
| | Sig. (2-tailed) | .000 | .000 | .000 | |
| | N | 1669 | 1669 | 1669 | 1669 |

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6.5 Analysis Result of Pacifiler

**Correlations**

| | | Average emtion score | Average subjective score | Average star rating | Reputation |
|---|---|---|---|---|---|
| Average emtion score | Pearson Correlation | 1 | .577** | .428** | .474** |
| | Sig. (2-tailed) | | .000 | .000 | .000 |
| | N | 646 | 646 | 646 | 646 |
| Average subjective score | Pearson Correlation | .577** | 1 | .248** | .284** |
| | Sig. (2-tailed) | .000 | | .000 | .000 |
| | N | 646 | 646 | 646 | 646 |
| Average star rating | Pearson Correlation | .428** | .248** | 1 | .999** |
| | Sig. (2-tailed) | .000 | .000 | | .000 |
| | N | 646 | 646 | 646 | 646 |
| Reputation | Pearson Correlation | .474** | .284** | .999** | 1 |
| | Sig. (2-tailed) | .000 | .000 | .000 | |
| | N | 646 | 646 | 646 | 646 |

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6.6 Analysis Result of Microwave

**Correlations**

| | | Reputation | Average star rating | Average emtion score | Average subjective score |
|---|---|---|---|---|---|
| Reputation | Pearson Correlation | 1 | .999** | .413** | .195** |
| | Sig. (2-tailed) | | .000 | .000 | .000 |
| | N | 1957 | 1957 | 1957 | 1957 |
| Average star rating | Pearson Correlation | .999** | 1 | .373** | .169** |
| | Sig. (2-tailed) | .000 | | .000 | .000 |
| | N | 1957 | 1957 | 1957 | 1957 |
| Average emtion score | Pearson Correlation | .413** | .373** | 1 | .397** |
| | Sig. (2-tailed) | .000 | .000 | | .000 |
| | N | 1957 | 1957 | 1957 | 1957 |
| Average subjective score | Pearson Correlation | .195** | .169** | .397** | 1 |
| | Sig. (2-tailed) | .000 | .000 | .000 | |
| | N | 1957 | 1957 | 1957 | 1957 |

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6.7 Analysis Result of Hair Dryer

It can be seen from the figure that the correlation between the four indexes of the hair dryer is lower than that between the microwave oven and the nipple, which proves that the hair dryer is a potential failure product.

## d. A Predictive Model Based on LDA and Specific Ratings for Comments

Based on the reputation evaluation model of time mode, this paper obtains the reputation change model of time. The phenomenon of following suit is that a particular star level causes more comments, often in a short time, caused by a review of a product of high quality or low quality. Negative/Positive comments correspond to specific stars(5 stars or 1 stars). This will also bring about rapid changes in the reputation of the product. As a result, we found three periods in which the product's reputation had the highest rate of change(Figure 6.8 ~ 6.10). LDA model was used to analyze the comments in this period, and the topic composed of several high-frequency words was obtained.

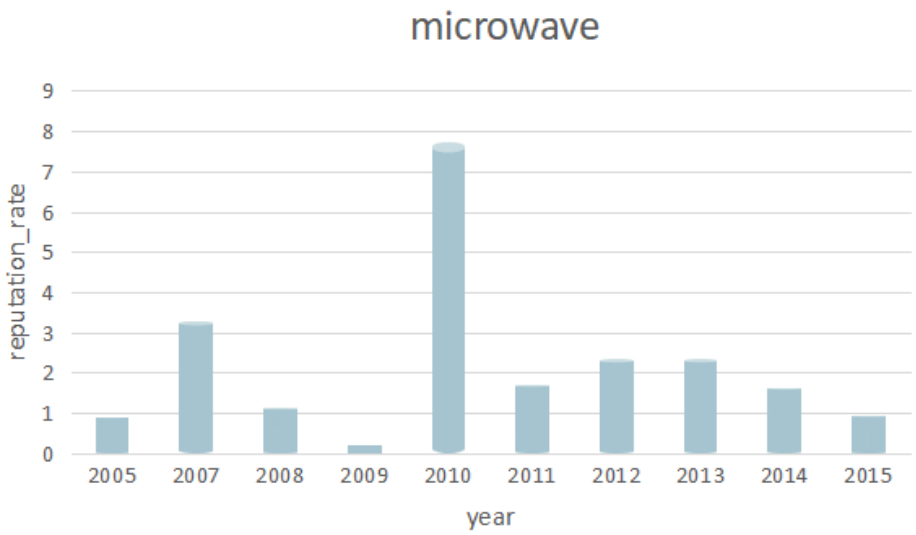


Figure 6.8 Growth Rate

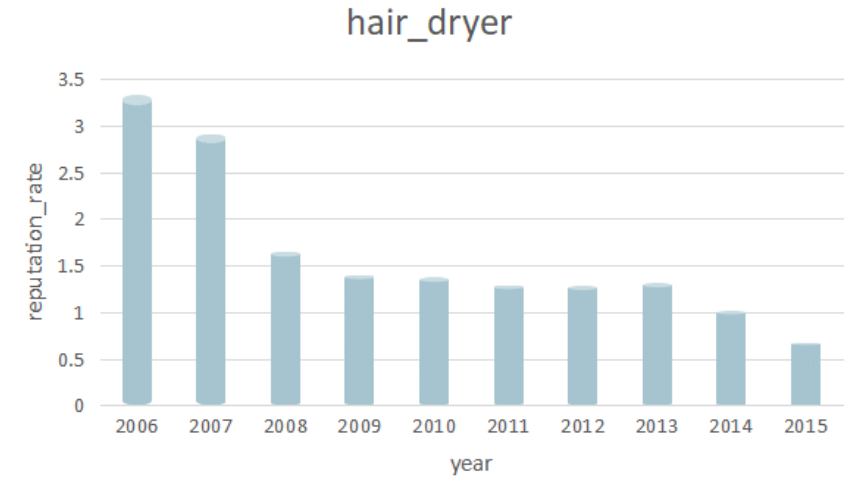


Figure 6.9 Growth Rate

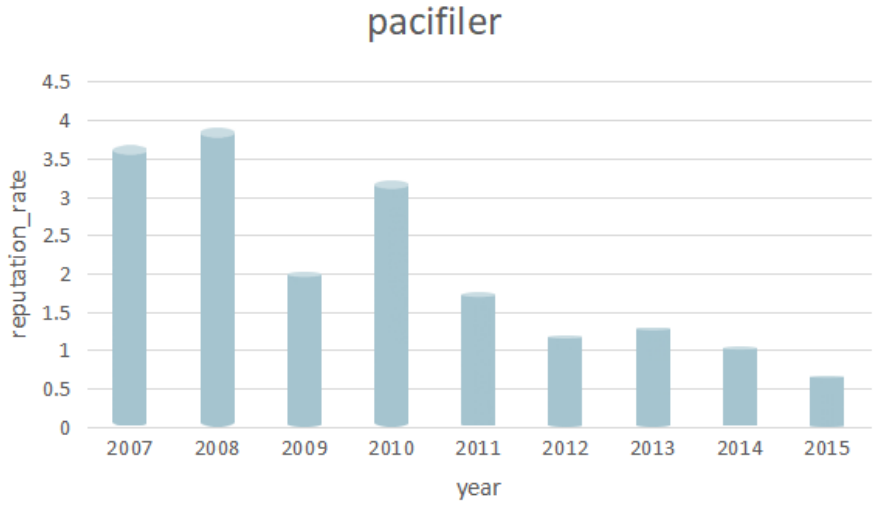


Figure 6.10 Growth Rate

From figure 6.8 ~ 6.10, it can be seen that the time periods in which the change rates of reputation of microwave oven, hair dryer and baby pacifier fluctuated greatly were 2010, 2006 and 2008, respectively.

After selecting the corresponding year comment with large fluctuations, the topic of the year comment is obtained through LDA model analysis (as shown in the following table).

Table 6.3 Analysis Result of LDA

| Positive themes | Negative theme |
|---|---|
| powerful | heat |
| price | speed |
| light | blow |
| | ***Hair_dryer*** |

Table 6.4 Analysis Result of LDA

| Positive themes | Negative theme |
|---|---|
| space | time |
| counter | door |
| compact | button |
| | ***Microwave*** |

Table 6.5 Analysis Result of LDA

| Positive themes | Negative theme |
|---|---|
| cute | color |
| gift | pink |
| clean | holes |
| | ***Pacifier*** |

From the above table shows a word description such as color, price, product quality than were higher than in other words, can be described product quality assessment will lead to copycat phenomenon, which will be more consumers to produce more similar comments, so that raise/lower the time period is the reputation of our products, or a particular star will cause further comment.

## e. The Relationship Between Text and Rating Levels

In order to find the relevance of specific quality descriptors (such as enthusiastic disappointed and others) to rating levels for textual comments. First, divide the user's

comments into positive and negative comments. Second, LDA model was used to extract the subject emotion words of each part, and the corresponding comment stars and help levels of the comment contents of a certain subject emotion words were counted. Finally, in order to make the evaluation more objective, we use the average value of corresponding indexes to eliminate the errors caused by users' subjective factors. The calculation formula is as follows:

$$ave_{star_i} = \frac{\sum_1^{n_i} S}{n_i} \quad (14)$$

$$ave_{help} = \frac{\sum_1^{n_i} H}{n_i} \quad (15)$$

Where $ave_{star_i}$ represents the average score of theme $i$;

S represents the star rating for a comment containing topic $i$;

$n_i$ represents the number of times the topic $i$ appears in the comments;

$ave_{help}$ represents the help average score;

$H$ represents the corresponding help score for a comment containing theme $i$;

To sum up, we can get the average score of each product in the following table:

Table 6.6 Analysis Results of Hair Dryer

| Positive themes | Average star | Average help level |
|---|---|---|
| price | 4.339 | 0.297 |
| light | 4.366 | 0.3 |
| power | 4.266 | 0.306 |
| powerful | 4.37 | 0.312 |
| fast | 4.52 | 0.33 |

| Negative theme | Average star | Average help level |
|---|---|---|
| cord | 4.13 | 0.39 |
| heat | 4.015 | 0.36 |
| speed | 4.09 | 0.35 |
| blow | 4.15 | 0.32 |
| handle | 4.05 | 0.35 |
| | *avg_start:4.186* | *avg_help:0.243* |

Table 6.7 Analysis Results of Microwave

| Positive themes | Average star | Average help level |
|---|---|---|
| space | 4.462 | 0.56 |
| price | 4.265 | 0.38 |
| counter | 4.368 | 0.54 |
| size | 4.291 | 0.47 |
| compact | 4.53 | 0.45 |

| Negative theme | Average star | Average help level |
|---|---|---|
| time | 3.789 | 0.57 |
| door | 3.32 | 0.53 |
| button | 3.673 | 0.58 |
| install | 3.91 | 0.59 |
| model | 3.777 | 0.616 |
| | *avg_start:4.025* | *avg_help:0.423* |

Table 6.8 Analysis Results of Pacilier

| Positive themes | Average star | Average help level |
|---|---|---|
| cute | 4.582 | 0.117 |
| gift | 4.695 | 0.139 |
| clean | 4.483 | 0.188 |
| wubbanub | 4.66 | 0.121 |
| Love | 4.75 | 0.126 |

| Negative theme | Average star | Average help level |
|---|---|---|
| nipple | 4.063 | 0.225 |
| color | 3.915 | 0.182 |
| pink | 3.61 | 0.22 |
| mouth | 4.286 | 0.19 |
| holes | 3.86 | 0.21 |
| | *avg_start:4.41* | *avg_help:0.13* |

From the above table can be concluded that: 1) The average help level of users' negative rating is higher than that of the average help level of users' positive rating, thus it can be seen that negative rating are more helpful to users. 2) The average score of users' positive rating is higher than the average score, negative rating score is below the average rating score. Thus, text-based user comments are positively correlated with rating levels.

# 7. Part 3: A Letter

Mr. Marketing Director of Sunshine Company:

We feel very honored to be your sales consultant. We did a lot of digging into the data and made recommendations based on the problems the company wanted to solve.

First of all, we established a model based on feature extraction of LDA, extracted from user reviews several main distribution, the theme of the blower are mainly distributed in the theme of the price and speed of dry, microwave ovens are mainly distributed in the theme of the size, price, and the heating time, the main theme of the pacifier around the pacifier morphology and the baby's liking. Therefore, the company can focus on these factors when improving the product, so as to better improve the performance of the product and improve user satisfaction.

Second, the analysis of the ratings and comments will help us better understand the product market, but emotions is often difficult to objectively make an accurate judgment, so we use textbolb quantify the passion of the user comments, and it is concluded that the weight of star 0.7306, comments emotional weight of 0.1884, comment on the weight of subjective score of 0.081. Based on this, the evaluation model based on AHP is established, and its mathematical expression is

$$y = 0.7306x_1 + 0.1884x_2 + 0.181x_3$$

Where y is the total score for the good,$x_1$ is Star score,$x_2$ is comment sentiment score,$x_3$ is comment on subjective scores. Companies can evaluate the quality of the commodity market according to the y-value of the model.

Third, our annual sales and reputation through the statistics of each product, found the reputation of the hair dryer in 2002-2013 showed a trend of growth, the reputation since 2013 to maintain high levels of blower and tends to be stable, the visible hair dryer has a good reputation on the market, thus has a larger development potential, in the same way, the pacifier, microwave ovens in 2013 after the reputation has been maintained at a higher level. So these three kinds of products have great development potential.

Fourth, through the study of reviews and star ratings, we found that a higher reputation can bring a higher sales volume to the product, so the company can expand its popularity by advertising to achieve a higher sales volume. In addition, certain star ratings lead users to similar comments

Finally, we through the LDA model for user reviews of subject headings, and calculate the keywords of the average star score and help ratings by calculating results we found that the text-based user comments and rating levels were positively correlated, and negative evaluation for more help to the user, this has to do with the results obtained from the literature (the second) is the same, so we hope that the companies are trying to improve products focusing on the user's bad review, to understand user needs, in order to obtain better market

These are the results of our team's research. I hope our research can help your company achieve greater success! Have a good day!

Best wish

# 8. Conclusion

## 8.1 Model Strengths:

8.1.1 Topic extraction based on LDA model can accurately judge the needs of users.

8.1.2 by calculating the average score of each index, the extreme of individual users can be greatly reduced, so as to make more objective evaluation.

8.1.3 users' comments are quantified so that their subjective expression can be measured in the form of data.

## 8.2 Model Weaknesses and Limiting Assumptions:

8.2.1 the weight determined by AHP has strong subjectivity.

8.2.2 LDA model is limited by the sample space and has certain limitations.

# 9. References

[1].VictorHo.AchievingServiceRecoverythroughRespondingtoNagativeOnlineReviews[J].Discourse&Communication,2017,11(1)
[2].Hatzivassiloglou V,Mckeown K R.Predicting the Semantic Orientation of Adjectives[C].Proceedings of the Acl,1997
[3].翟丽孔．网店在线评论对消费者购买意愿的影响研究[D]．大连：东北财经大学 2011．
[4].https://zhuanlan.zhihu.com/p/31470216
[5] BAMMANNK. Statistical models; theory and practice[J]. Biometrics, 2006(62);943.
[6] BISHOP CM. Pattern recognition and machine learning[M]. New York: Springer, 2006.

# 10. Appendices

## APH:

```
A=input('A=');
[n,n]=size(A);
x=ones(n,100);
y=ones(n,100);
m=zeros(1,100);
m(1)=max(x(:,1));
y(:,1)=x(:,1);
x(:,2)=A*y(:,1);
m(2)=max(x(:,2));
y(:,2)=x(:,2)/m(2);
p=0.0001;i=2;k=abs(m(2)-m(1));
while  k>p
  i=i+1;
  x(:,i)=A*y(:,i-1);
  m(i)=max(x(:,i));
  y(:,i)=x(:,i)/m(i);
  k=abs(m(i)-m(i-1));
end
a=sum(y(:,i));
w=y(:,i)/a;
t=m(i);
disp(w);
CI=(t-n)/(n-1);RI=[0 0 0.52 0.89 1.12 1.26 1.36 1.41 1.46 1.49 1.52 1.54 1.56 1.58 1.59];
CR=CI/RI(n);
if CR<0.10
    disp('Accept!');
    disp('CI=');disp(CI);
    disp('CR=');disp(CR);
end
```

## LDA:

```
#-*- coding : utf-8-*-
# coding:unicode_escape
from snownlp import SnowNLP
import pandas as pd
import re
import jieba
import collections
```

```
from gensim import corpora, models

def reputation_data():
    data_null = data.drop_duplicates()
    print(data_null)
    data_null.to_csv('comments_null.csv')
    data_null_comments = data_null['review_body']
    data_null_comments.to_csv('contents.txt', index=False, encoding='utf-8')
    print(len(data_null_comments))

    data_len = data_null_comments[data_null_comments.str.len() > 4]
    print(data_len)
    data_len.to_csv('contents.txt', index=False, encoding='utf-8')



def feel_analyze():
    data = pd.read_csv('contents.txt',encoding='utf-8', header=None)
    coms = data[0].apply(lambda x: SnowNLP(x).sentiments)
    data_post = data[coms >= 0.5]
    data_neg = data[coms < 0.6]
    # print(data_post)
    # print(data_neg)
    data_post.to_csv('post_result.txt', encoding='utf-8', header=None)
    data_neg.to_csv('neg_result.txt', encoding='utf-8-sig', header=None)

def remove_sample():
    with open('post_result.txt', encoding='utf-8') as fn1:
        string_data1 = fn1.read()
    # r = "|\.|-|——|：|!|、|，|'|’|,|。|;|\)|\(|\?"
    pattern = re.compile(u'\t|\.|-|—|：|！|、|，|,|。|;|!|\)|\(|\?')
    string_data1 = re.sub(pattern, '', string_data1)
    # print(string_data1)
```

```
    fp = open('comments_post.txt', 'a', encoding='utf8')
    fp.write(string_data1 + '/n')
    fp.close()

    with open('neg_result.txt',encoding='utf-8') as fn2:
        string_data2 = fn2.read()
    #  re.compile(u'"|\.|-|——|：|!|、|，|,|。|;|\)|\(|\?"')
    pattern = re.compile(u'\t|\.|-|—|：|！|、|，|,|。|;|!|\)|\(|\?')
    string_data2 = re.sub(pattern, '', string_data2)
    # print(string_data2)

    fp = open('comments_neg.txt','a',encoding='utf8')
    fp.write(string_data2+'/n')
    fp.close()

def cut_words():
    data1 = pd.read_csv('comments_post.txt',encoding='utf-8',header=None, delimiter='\t')
    data2 = pd.read_csv('comments_neg.txt',encoding='utf-8',header=None, delimiter='\t')

    mycut = lambda s: ' '.join(jieba.cut(s))
    data1 = data1[0].apply(mycut)
    data2 = data2[0].apply(mycut)

    data1.to_csv('comments_post_cut.txt',index=False,header=False,encoding='utf-8')
    data2.to_csv('comments_neg_cut.txt',index=False,header=False,encoding='utf-8')
    print(data2)

def word_statistic():
    with open("comments_neg_cut.txt",encoding="utf-8") as fn:
        string_data = fn.read()
        word_counts = collections.Counter(string_data)
        word_counts_top10 = word_counts.most_common(10)
        for w, c in word_counts_top10:
            print(w, c)
```

```
    with open("comments_post_cut.txt",encoding="utf-8") as fn:
        string_data = fn.read()
        word_counts = collections.Counter(string_data)
        word_counts_top10 = word_counts.most_common(10)
    for w, c in word_counts_top10:
            print(w, c)

def LDA():
    post = pd.read_csv('comments_post_cut.txt',encoding='utf-
8',header=None,error_bad_lines=False)
    neg = pd.read_csv('comments_neg_cut.txt',encoding='utf-
8',header=None,error_bad_lines=False)
    stop  =  pd.read_csv('stoplist.txt',encoding='utf-8',header=None,sep='tipdm',engine='python')

    stop = [' ', ''] + list(stop[0])

    post[1] = post[0].apply(lambda s: s.split(' '))
    post[2] = post[1].apply(lambda x: [i for i in x if i not in stop])
    neg[1] = neg[0].apply(lambda s: s.split(' '))
    neg[2] = neg[1].apply(lambda x: [i for i in x if i not in stop])

    print("positive topic analysis.....")
    post_dict = corpora.Dictionary(post[2])
    post_corpus = [post_dict.doc2bow(i) for i in post[2]]
    post_lda = models.LdaModel(post_corpus, num_topics=2, id2word=post_dict)
    for i in range(2):
        print(post_lda.print_topic(i))

    # print('first')

    print("Negative topic analysis.....")
    neg_dict = corpora.Dictionary(neg[2])
    neg_corpus = [neg_dict.doc2bow(i) for i in neg[2]]
    neg_lda = models.LdaModel(neg_corpus, num_topics=2, id2word=neg_dict)
```

```
    for i in range(2):
        print(neg_lda.print_topic(i))


if __name__ == '__main__':
    data = pd.read_csv('pacifier.tsv', sep='\t', encoding='unicode_escape')
    print(len(data))
    coms = []
    reputation_data()
    feel_analyze()
    remove_sample()
    cut_words()
    word_statistic()
    LDA()


#-*- coding : utf-8-*-
# coding:unicode_escape
from textblob import TextBlob
import pandas as pd
import matplotlib.pyplot as plt
from textblob import TextBlob
'''
# text = "text"
# blob = TextBlob(text)
# c = blob.sentiment  # Sentiment(polarity=0.15000000000000002, subjectivity=1.0)
# blob.sentences  # [Sente    nce("I am happy today."), Sentence("I feel sad today.")]
# c = blob.sentiment  # Sentiment(polarity=0.15000000000000002, subjectivity=1.0)
# print(c)
# polarity
# subjectivity
'''

def get_one_motion():
```

```
    test1 = "Amazing addition to the nursery!"
    blob1 = TextBlob(test1)
    print(blob1.sentiment)

def get_avg_motion():
    for i in range(len(data)):
        test1 = data.iloc[i, 1]
        blob1 = TextBlob(test1)
        p, s = blob1.sentiment
        p = float('%.3f' % p)
        s = float('%.3f' % s)
        polarity.append(p)
        subjectivity.append(s)
        print("polarity = %.2lf, subjectivity = %.2lf" % (polarity[i], subjectivity[i]))

def motion_count():
    pos_count = 0
    nag_count = 0
    for i in range(len(data)):
        if polarity[i] > 0:
            pos_count = pos_count+1
        else:
            nag_count = nag_count+1
    print('pos_count = %d, nag_count = %d' % (pos_count, nag_count))

def show():
    plt.bar(range(len(subjectivity)), subjectivity)
    plt.show()

def write_data():
    dataframe = pd.DataFrame({"polarity": polarity, "subjectivity":subjectivity})
    dataframe.to_csv("hair_result.tsv", index=False, sep='\t')
```

```
if __name__ == '__main__':
    data = pd.read_csv('microwave1.tsv', sep='\t', encoding='unicode_escape')
    polarity = []
    subjectivity = []
    get_avg_motion()
    # get_one_motion()
    motion_count()
    write_data()
```

## Sentiment analysis:

```
#-*- coding : utf-8-*-
# coding:unicode_escape
import pandas as pd
import matplotlib.pyplot as plt
from textblob import TextBlob

def get_avg_motion():
    for i in range(len(data)):
        test1 = data.iloc[i, 1]
        blob1 = TextBlob(test1)
        p, s = blob1.sentiment
        p = float('%.3f' % p)
        s = float('%.3f' % s)
        polarity.append(p)
        subjectivity.append(s)
        print("polarity = %.2lf, subjectivity = %.2lf" % (polarity[i], subjectivity[i]))

def write_data():
    dataframe = pd.DataFrame({"polarity": polarity, "subjectivity":subjectivity})
    dataframe.to_csv("hair_result.tsv", index=False, sep='\t')

if __name__ == '__main__':
    data = pd.read_csv('microwave1.tsv', sep='\t', encoding='unicode_escape')
    polarity = []
```

```
subjectivity = []
get_avg_motion()
write_data()
```