Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2004898

Reviews are a treasure! How to dig it?

Summary

With the development of technology, more and more users enjoy the convenience brought by online shopping. To help Sunshine Company better prepares its products for sale online, we analyzed the review data on Amazon's website, and the relevant conclusions are as follows.

First, we set up an **LDA model**, which converts review data in the form of strings into numerical data that symbolizes customer likes and dislikes for later analysis. In order to find the interesting internal relationship among star rating, review and helpfulness rating, we use **k-means** algorithm to divide customer review into four categories: fuzzy poor review, clear poor review, fuzzy good review and clear good review, so as to provide more targeted analysis for Sunshine Company.

Secondly, in order to evaluate the reputation of a product quantitatively, we carefully select the relevant indicators in the data set, define the **reputation index(RI)**, and complete the construction of the RI evaluation model with the help of **EWM-TOPSIS** model. Since the product review data can be viewed as a time series, we build a product reputation prediction model based on **ARIMA** time series to determine whether its reputation is on the rise or down trend.

Next, we use **SVM** model to find out the potential successful or failing products, and find that the potential successful products perform very well in each secondary index of RI.

In addition, according to **Matthew effect**, we speculate that the higher a product's existing star rating, the more likely customers are to give it a better review. Take into account this, we make **correlation analysis**, and the results show that our conjecture is correct. Moreover, we also use **Apriori** algorithm to mine association rules and find that there is a strong association between star rating and specific quality descriptors. For example, for hair dryer, high star rating is often accompanied by "hot" and "powerful", while low star rating is accompanied by "long time" and "waste money". This provides a valuable reference for Sunshine Company to understand customers' preferences.

Finally, as a creative way, the results of **sensitivity analysis** demonstrate that our model is robust and reliable. Based on the analysis of the whole paper, we suggest that Sunshine Company should not only pay attention to the product quality, but also pay attention to the after-sales of the product, and respond to the customer reviews in time to achieve success in the online sales market.

Keywords: shopping reviews, LDA, evaluation model, ARIMA, Matthew effect

Content

1 Introduction....1
1.1 Problem background....1
1.2 Literature review....1
1.3 Our work....1
2 Preparation of the Models....2
2.1 Assumptions....2
2.2 Notations....2
3 Data processing and overview....3
3.1 Dislocation value processing....3
3.2 Missing value processing....3
3.3 Star rating overview....3
3.4 Review overview....4
3.5 Helpfulness rating overview....4
4 Part1: Data pattern analysis....5
4.1 LDA theme model....5
4.2 Text-value conversion model....7
4.3 Model of Star Rating, Review and Helpfulness Rating based on K-means....8
5 Part2: Problem analysis based on "reputation index"....9
5.1 The construction of reputation index....9
5.2 Prediction of changes in reputation index....12
5.3 Discovery of potentially successful or failing products....14
5.4 The influence of star ratings on subsequent reviews....15
5.5 The association between descriptors and star levels....17
6 Sensitivity analysis....20
7 Strengths and Weaknesses....20
Letter....21
References....23

1 Introduction

1.1 Problem background

With the development of e-commerce and the improvement of the level of logistics services, more and more people have chosen online shopping as a fast and convenient way of shopping. To help customers better understand the products they want to buy and improve customer satisfaction with online transactions, many e-commerce sites have launched online customer review systems, such as Amazon.

As the number of reviews grows, so does the information they contain. While this information helps consumers make decisions, it can also provide merchants with a lot of valuable information. How to mine valuable information from a large number of reviews and provide a basis for merchants to implement corresponding sales strategies is becoming a challenge.

1.2 Literature review

Through consulting materials, we find that there has been a certain amount of research on the issue of product rating and evaluation on the e-commerce platform. For example, Mudambi and Schuff[1] divided products into search products and experience products, and found that the influence of review stars on review usefulness was restricted by product types. Wang junkui[2] included some other attributes of the product into the evaluation usefulness model and extended the evaluation usefulness model. Cao et al.[3] used the method of text mining to study the influence of basic features, cooperative style and grammatical features of comments on the number of votes available for comments, and concluded that compared with other features, syntactic features have more influence on the number of votes available for comments.

However, most of the literature don't study the internal characteristics and interrelationships among various types of review data in shopping websites. Also, they don't use reviews to fully evaluate the product.

1.3 Our work

Based on the analysis of the problem, we propose the model framework shown in figure 1, which is mainly composed of the following parts:

- **Data processing:** by observing the data, we carry out the repositioning and deletion of misplaced data and abnormal data.
- **Dataset analysis:** LDA model and k-means algorithm are used to find interesting relationships and laws between different indicators.
- **Reputation definition:** select indicators from the provided data set, construct evaluation model with EWM and TOPSIS, and define "reputation index" to provide valuable information for Sunshine Company to track.

• **Reputation prediction:** the historical evaluation data of a product is regarded as a time series, and the ARIMA time series model is constructed to determine whether a product's reputation is rising or falling by predicting the changes of reputation index.

• **Potential success/failure:** use SVM model to classify and identify potential success or failure products.

• **Incitement:** conduct a correlation analysis to see if past comments have any potential impact on new ones.

• **Descriptors and stars:** the Apriori algorithm is used for correlation analysis to mine the correlation relationship between quality descriptors and star levels.

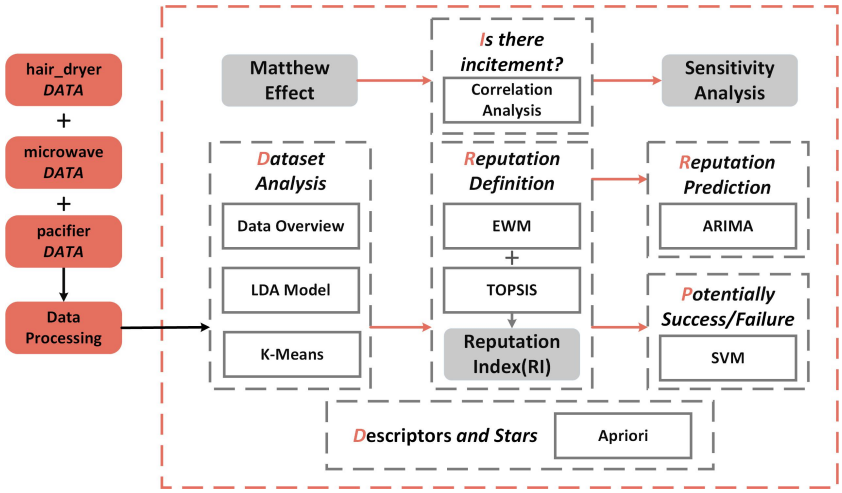

Fig 1: Model framework

2 Preparation of the Models

2.1 Assumptions

• **Assumption 1:** assume that a customer will read the reviews carefully before making a purchase.

Reason: this assumption is made to ensure that product reviews have an impact on customer shopping decisions.

• **Assumption 2:** assume that the pre-processed data is reliable.

Reason: this assumption is made to ensure the accuracy of the model solution.

• **Assumption 3:** assume that customer reviews are credible.

Reason: malicious reviews, such as comments made by customers to discredit a product, can have a significant impact on our model results. But we cannot tell whether a review is intentional, so we make this assumption.

2.2 Notations

The symbols used in this paper are shown in table 1:

Table 1: Notations

Symbol	Definition
Hr	helpfulness ratings
helpful	the amount of helpful_votes
helpless	the amount of helpless_votes
total	the amount of total_votes
RI	reputation index

3 Data processing and overview

3.1 Dislocation value processing

We find that the data in some records is misplaced, for example, in "pacific.tsv" :

Table 2: Example of data misplaced

marketplace	customer_id	review_id	...	review_body	review_date
US	38939521	R32UHBXCITE4LC	...	8/18/2015	null

In this case, the review_body records date information, while the value of review_data is null. After further observation, we find that this is the result of the missing of the product_category column, which cause the data dislocation. In this case, we complete the product_category information as "baby" and repaire the data that has been misplaced.

3.2 Missing value processing

We find that some records had missing content, for example, in "pacific.tsv" :

Table 3: Examples of data exception

marketplace	customer_id	review_id	...	star_rating	...
US	34761565	r1coxfmkp3cqvq	...	null	...

There are a large number of missing items in this type of data that cannot be fixed, so we simply delete them.

3.3 Star rating overview

Star Ratings are integer type data that represent 1-5 Star Ratings for reviews and are direct numerical Ratings from customers. Through previous processing, we can get the data distribution after cleaning as follows:

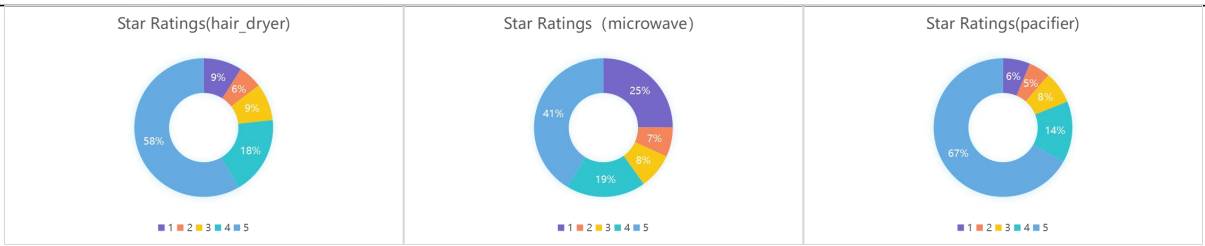

Fig 2: Star Ratings distribution for three data sets

In the data set of hair_dryer and pacifier, more than half of the samples received 5 star praise and 1 star negative evaluation was less than 10%.Less than half of the 5 - star positive comments in the microwave data set and 1 - star negative comments as high as a quarter. It can be seen that the distribution rules of the three samples are different, all of which have their corresponding research value, and cannot be combined for research. On the whole, the five star ratings are unevenly distributed, and the influence of the differences in the number of samples on the model's scientific nature should be eliminated in the process of establishing the model.

3.4 Review overview

Review is the data of string type, which is composed of review_headline and review_body. It is rather subjective and hard to distinguish. For the text with strong subjectivity, text mining is needed. The recognition of data determines the tuning effect of text mining algorithm. It is not difficult to find that most comments corresponding to the 2-4 star rating are "raise before raise" or "raise after raise". The text contains both derogatory and commendatory meanings, which has certain interference to text mining. In view of the above situation, it is necessary to merge the fields of review_headline and review_body from the comment tuple corresponding to the 1-star and 5-star scores, and then perform operations such as word frequency statistics and illegal word cleaning. Taking the microwave data set as an example, the frequency statistics of some high-frequency meaningful words are as follows:

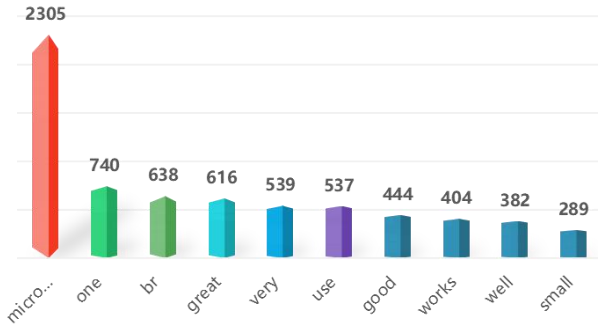

Fig3: Number of words in reviews

3.5 Helpfulness rating overview

Helpfulness ratings consist of helpful_votes and total_votes. Combined they can also be derived helpless_votes data. Total_votes helpful_votes and helpless_votes effect

is relative, if just the three simple combination will be difficult to handle the total_votes 0 case, so we need to find other ways to deal with it.

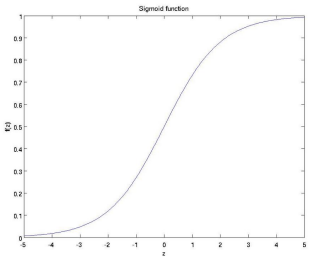

Fig 4: Sigmoid function

As shown in figure 4, the Sigmoid function maps variables into intervals, referring to the idea of data processing in machine learning. Because of its single increase and inverse function single increase properties, it can directly reflect the positive and negative effects of helpful_votes and helpless_votes, and deal with the originality of total_votes 0.

$$Hr = Sigmoid(helpful - helpless) = Sigmoid(2helpful - total) \quad (3\text{-}1)$$

Take the value of helpful_votes minus helpless_votes as an independent variable to get: when the total number of helpful votes is 0, the value is 0.5; when the number of helpful votes is large, the value is greater than 0.5; when the number of helpful votes is large, the value is less than 0.5. Moreover, the model has high discrimination and sensitivity.The distribution of final data processing results is shown in figure 5:

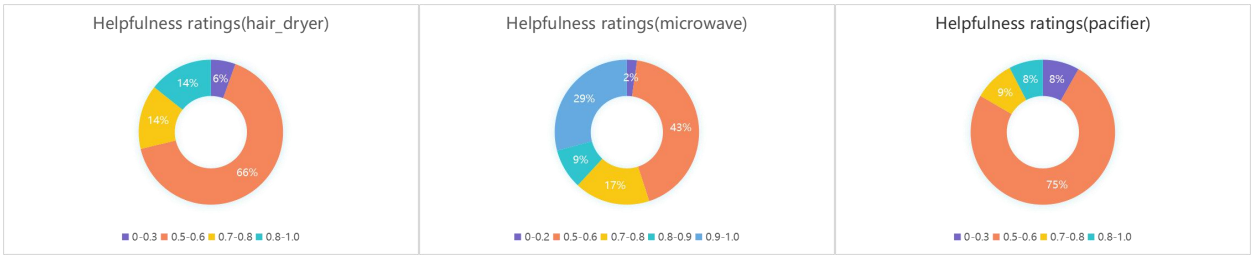

Fig5: Helpfulness ratings distribution for three datasets(take the lower limit)

4 Part1: Data pattern analysis

4.1 LDA theme model

LDA(Latent Dirichlet Allocation) is a document topic generation model, also known as a three-tier bayesian probability model, containing three-tier structures of words, topics and documents[4]. Documents to topics obey polynomial distribution, and topics to words obey polynomial distribution.Its purpose is to identify the theme, that is, the document-vocabulary matrix into the document-theme matrix and theme-vocabulary matrix.

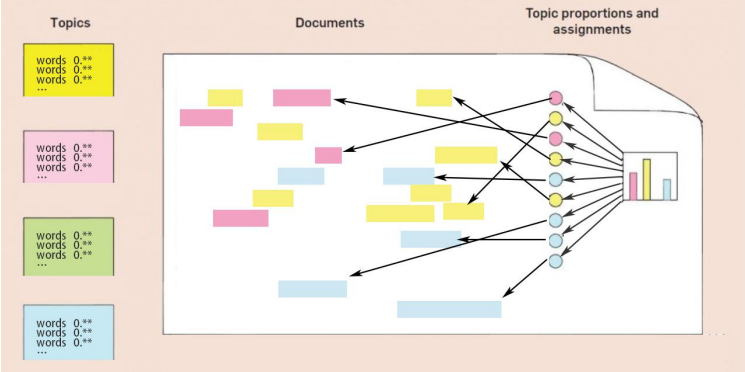

Fig6: LDA model schematic

A document has multiple topics, and each topic has a different word. The construction of a document, first with a certain probability to choose a topic, and then under the topic with a certain probability to choose a certain word, so as to generate the first word of the document. Repeat this process over and over again, and the whole article is generated. The use of LDA is the reverse process of the above document generation process, that is, according to a known document, to find out the topic of this document, and the words corresponding to these topics. Figure 7 uses the probability graph model of LDA to describe LDA accurately and formally:

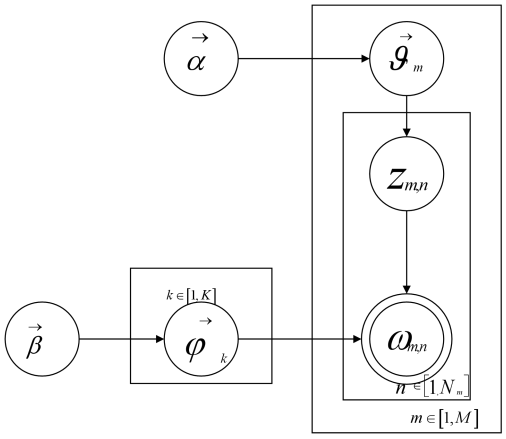

Fig7: LDA probability graph model

According to the above process, the sum under the current situation can be obtained by iterating until a certain degree of convergence. Further, the joint probability of the database can be calculated by the following topic distribution and word distribution:

$$p(\vec{z} \mid \vec{\alpha}) = \prod_{m=1}^{M} \frac{\Delta(\vec{\vartheta}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}, p(\vec{\omega}, \vec{z} \mid \vec{\alpha}, \vec{\beta}) = \prod_{k=1}^{K} \frac{\Delta(\vec{\varphi}_k + \vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{\vartheta}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \tag{4-1}$$

In terms of parameter selection, empirically, β is 0.01 and α is $50/k$, where k is the number of selected topics. From the data overview, we can see that the data set of the model is the set generated by merging the review_headline and review_body fields

from the comment tuples corresponding to the 1 star and 5 star scores. Therefore, the number of topics in the model is 2, that is, k =2.

After the above process, the subject word and its weight of each topic can be solved. In the microwave dataset, for example, the 1-star and 5-star records that were randomly typed were split into two topics. The top 10 topic words of each topic are shown in figure 8:

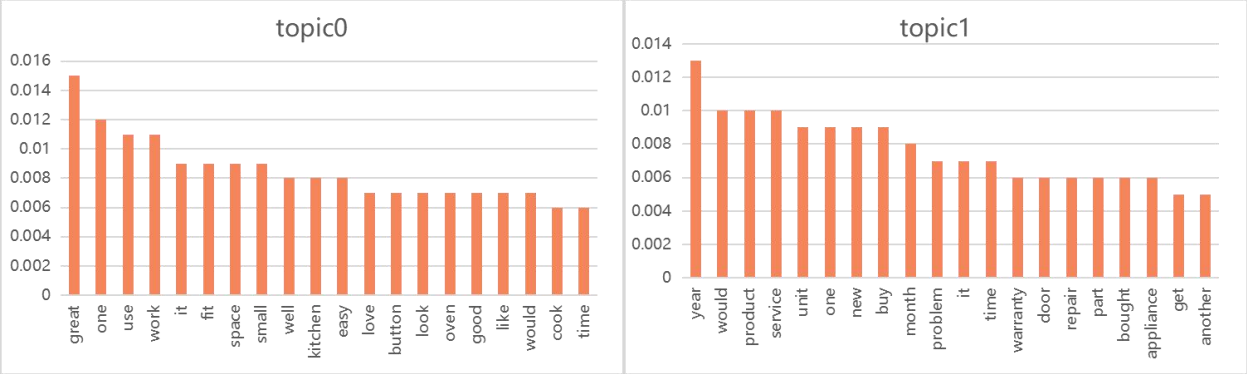

Fig 8: Microwave data set LDA topic model solution results (part)

More subject words means more precise division of topics. For the sake of both efficiency and accuracy, the number of theme words selected in this paper is 100, and the weight sequence of each theme word of the fourth theme is $\vec{v_i}$.

4.2 Text-value conversion model

When a sample is selected, it is easy to obtain the existence or absence of each subject word of the i-th topic sequence $\vec{f_i}$. Simply multiply sequence $\vec{f_i}$ and sequence $\vec{v_i}$ to obtain the similarity of the sample for the i-th topic:

$$factor_i = \vec{f_i} \times \vec{v_i} \tag{4-2}$$

Subject words are only a part of the results of LDA subject model. In order to eliminate the interference of the above two cases, it is necessary to use the most classification model to carry out coefficient correction:

$$\begin{cases} \lambda_i = \dfrac{\sum_{k=1}^{K} \sum_{n=1}^{Maxn} v_{k,n}}{\sum_{n=1}^{Maxn} v_{i,n}} \\ factor_i = \lambda_i \times \vec{f}_i \times \vec{v}_i \end{cases} \tag{4-3}$$

By comparing the similarity of the classification model, the textual review data can be converted into the discrete value of positive and negative reviews to facilitate the verification of the model.

However, the discrete value of 0-1 distribution has a general effect in subsequent clustering analysis. Since the number of topics in this model is 2, the positive factor can

be generated in the following way. The positive factor is independent of the sum of the two similarity degrees, and it can accurately reflect the emotional tendency of the text information of review within the [0,1] interval.

$$Pos\_factor = \frac{1+(factor_0 - factor_1)}{2(factor_0 + factor_1)} \tag{4-4}$$

• **Model validation:**

In the process of writing a review, customers with 1 star and 5 star will clearly indicate their attitude towards the product, and their text review has a high correlation with the star rating. In the case of the microwave data set, from the solution results of the thematic model, topic0 is biased towards positive comments. The model can be verified by matching 5 star-level samples with topic0 samples and 1 star-level samples with topic1 samples. By calculation, the inaccuracy is 0.072, indicating that the model is more accurate.

4.3 Model of Star Rating, Review and Helpfulness Rating based on K-means

K-means clustering algorithm is an iterative clustering analysis algorithm. The steps are as follows: pre-divide the data into K groups, randomly select K objects as the initial clustering center, then calculate the distance between each object and each seed clustering center, and assign each object to the nearest clustering center. Cluster centers and the objects assigned to them represent a cluster. For each sample assigned, the cluster center is recalculated based on the existing objects in the cluster. This process will continue until some termination condition is met. The termination condition can be that no (or minimum number) objects are redistributed to different clusters, no (or minimum number) clustering center changes again, error squared and local minimum. The flow chart of K-means algorithm is shown in figure 9.

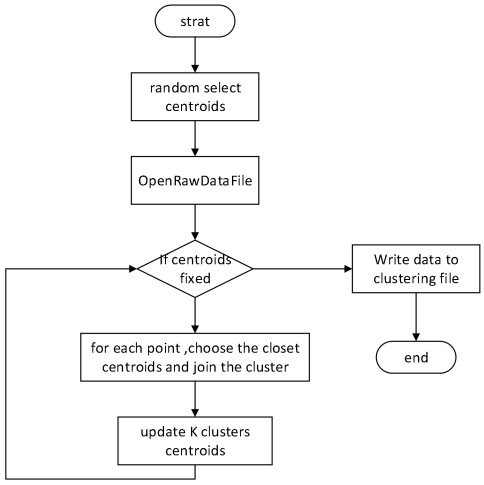

Fig9: Flow chart of K-means algorithm

The existing data set X is composed of $x^{(1)}, x^{(2)}, ..., x^{(m)}$. The model needs to classify the data cluster into k clusters : $\{C\} = [C_1, C_2, ..., C_k]$, and the minimum loss function is:

$$\begin{cases} E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \\ \mu_i = \partial g(x) 1 |C_i| \sum_{x \in C_i} x \end{cases} \tag{4-5}$$

Where μ_i is the central point of cluster C_i. Firstly, k sample points were randomly selected from the samples to act as the center point $\{\mu_1, \mu_2, ..., \mu_k\}$ of each cluster, and the following process was repeated: calculate the distance $dist(x^{(i)}, \mu_j)$ between all sample points and the center of each cluster, and then divide the sample points into the nearest cluster $x^{(i)} \in \mu_{nearest}$; According to the existing sample points in the cluster, the cluster center $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ is recalculated.

In terms of parameter selection, set the number of generated clusters to 4, the number of iterations to 300, and run the algorithm 10 times with different centroid initialization values, and the clustering effect is the best.

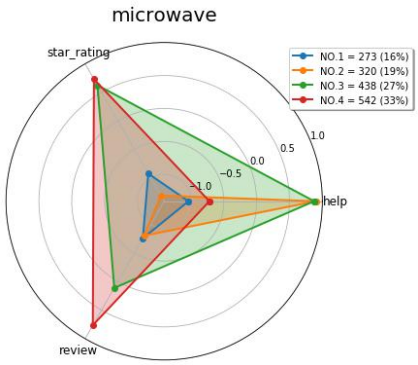

Fig10: Microwave dataset Star Rating, Review, Helpfulness Rating cluster radar chart

As shown in figure 10 to microwave data sets for example, Star Rating, Review and Helpfulness Rating can be divided into four types: the blue category is the fuzzy poor review with low score, low evaluation and small help; the yellow category is the clear poor review with low score, low evaluation and large help; the red category is the fuzzy good review with high score, high evaluation and small help; the green category is the clear good review with high score, high evaluation and large help.

5 Part2: Problem analysis based on "reputation index"

5.1 The construction of reputation index

In order to obtain the most valuable information from ratings and reviews for Sunshine Company to evaluate the products after they are launched, we define the "reputation index", which can reflect the level of a product's reputation in the market for Sunshine Company to track.

Therefore, EWM-TOPSIS evaluation model is constructed to solve the reputation index. Next we will elaborate on the corresponding process.

• Note: when we construct the evaluation model, we not only consider the two factors of "review" and "rating", because "helpful_votes", "total_votes", "vine" and "verified_purchase" also indirectly affect "review". Considering only "review" and "rating" will lead to a large deviation in the results.

5.1.1 Index selection

• **Review score** A_1 **:** in the previous chapter, we used LDA to process the text, so as to quantify the review and obtain its relative sentiment score. We recorded it as LDA_Score, which has a positive correlation with product reputation. In this paper, we directly use LDA_Score to reflect the comment score:

$$A_1 = LDA\_Score$$

• **Star rating** A_2 : the star rating of a product is positively correlated with the reputation, which directly and quantitatively reflects the customer's satisfaction with the product. Therefore, we need to calculate the average star rating of a product:

$$A_2 = \frac{\sum_{i=1}^{c} r_i}{c}$$

Where c is the total number of reviews of the product, and r_i is the star rating of the review i.

• **"Vine Voice" review** A_3 **:** Customers with the status of "Vine Voice" enjoy a high reputation in the community, so their reviews also have a greater impact on the overall reputation of the product. And because Vine Voice has long experience of commenting, we only consider their star ratings.

$$A_3 = \begin{cases} \dfrac{\sum_{i=1}^{c_v} r_{c_i}}{c_v} & if\ c_v \neq 0 \\ A_2 & if\ c_v = 0 \end{cases}$$

Where c_v is the total number of "Vine Voice" reviews of the product, and r_{c_i} is the star rating of the review i.

• **Helpful review** A_4 : a review that receives more "helpful votes" indicates that many customers also agree with the content of the review, indicating that the review is more reflective of the product's reputation. We use the processing results obtained in section 4 using the Sigmiod function to construct the index.

$$A_4 = Pos\_factor$$

5.1.2 EWM-TOPSIS evaluation model

Entropy was first introduced into information theory by shennong, and has been widely applied in engineering, technology, social economy and other fields[5].The smaller the information entropy of an index is, the greater the degree of variation of the

index value is, the more information it provides, the greater the role it can play in the comprehensive evaluation, and the greater its weight is.

TOPSIS method was first proposed by C.L.Hwang and K.Yoon in 1981. It is a method to rank according to the proximity between a limited number of evaluation objects and the idealized target, and to evaluate the relative merits of existing objects[6].

Next, we first use entropy weight method to define index weight, then use TOPSIS method to build reputation index.

Let $P=[P_1,P_2,\cdots,P_n]$ be the set of products, $A=[A_1,A_2,A_3,A_4]$ be the set of evaluation indexes, $X=(x_{ij})_{m\times n}$ be the decision matrix, and x_{ij} represent the value of the i-th product on the index j.

Step 1: data standardization

In order to eliminate the difference of data dimensions among different indicators, we should first conduct standardized data processing before data analysis:

$$x'_{ij}=\frac{x_{ij}-\min(x_{ij})}{\max(x_{ij})-\min(x_{ij})} \tag{5-1}$$

We call the dimensionless decision matrix $X'=(x'_{ij})_{m\times n}$.

Step 2: determine index weight based on entropy weight method

- Calculate the weight of product i on indicator j:

$$p_{ij}=\frac{1+x_{ij}}{\sum_{i=1}^{1+x_{ij}}(1+\mathrm{x}_{ij})} \tag{5-2}$$

- Calculate the coefficient of variation of entropy and index j:

$$\mathrm{e}_j=-(\ln m)^{-1}\sum_{i=1}^{m}p_{ij} \tag{5-3}$$

$$\mathrm{g}_i=1-e_j \tag{5-4}$$

- The final weight of indicator j can be calculated:

$$\mathrm{w}_\mathrm{j}=\frac{g_j}{\sum_{i=1}^{n}g_j} \tag{5-5}$$

- The weight set is as follows:

$$\mathrm{W}=[\mathrm{w}_1,\mathrm{w}_2,\mathrm{w}_3,\mathrm{w}_4] \tag{5-6}$$

Step 3: Evaluate based on the TOPSIS method

- First, calculate the weighted normalization matrix:

$$Z=(z_{ij})_{m\times\mathrm{n}} \tag{5-7}$$

Among them $z_{ij} = \mathrm{w}_j \times x_{ij}$.

• Then, the positive ideal solution Z^+ and the negative ideal solution Z^- are calculated:

$$\begin{aligned} Z^+ &= (z_1^+, z_2^+, z_3^+, z_4^+) \\ Z^- &= (z_1^-, z_2^-, z_3^-, z_4^-) \end{aligned} \tag{5-8}$$

Among them $Z_{\mathrm{j}}^+ = \max_i(x_{ij}'), Z_{\mathrm{j}}^- = \min_i(x_{ij}')$.

• Calculate and evaluate the closeness between each sample and the optimal scheme and the worst scheme:

$$\begin{cases} D_i^+ = \sqrt{\sum_{j=1}^{m} w_j (Z_j^+ - z_{ij})^2} \\ D_i^- = \sqrt{\sum_{j=1}^{m} w_j (Z_j^- - z_{ij})^2} \end{cases} \tag{5-9}$$

Finally, we can evaluate the degree of closeness between each sample and the optimal scheme and define it as the reputation index RI:

$$RI_{\mathrm{i}} = \frac{D_i^-}{D_i^+ + D_i^-} \tag{5-10}$$

5.1.3 Reputation index solution

We take the four products in the data set hair_dryer with the highest number of reviews (product id is B003V264WW, B0009XH6TG, B00132ZG3U, B00005O0MZ) as examples to solve their respective reputation indexes. We record these four products as P_1, P_2, P_3, P_4 respectively.

Using the entropy weight method, we can get the weight $\mathrm{W} = [0.3748, 0.3125, 0.1792, 0.1335]$ of the four indexes. Then, using the TOPSIS method, we can calculate $RI = [0.7709, 0.3734, 0.6375, 0.2291]$. From this result, it can be seen that among the four products, P_1 has the highest reputation and P_4 has the lowest reputation.

The reputation index can be a powerful tool to help Sunshine Company track its products. By calculating the reputation index of the products launched, Sunshine Company can intuitively understand the reputation of its products in the market, and learn from products with higher reputation, so as to avoid making mistakes made by products with lower reputation.

5.2 Prediction of changes in reputation index

5.2.1 The establishment of ARIMA model

ARIMA model, fully known as autoregressive moving average model, is a famous time series prediction method proposed by Box and Jenkins in the early 1970s[7].

Generally speaking, ARIMA model is denoted as $ARMA(p,q,d)$, where p is the order of autoregression, q is the order of moving average, and d is the degree of difference.

For a product, let Y_t be the record of the t^{th} review from the shelf to the present, which is a sequence of non-stationary random variables. After the d^{th} difference, the stable sequence X_t can be obtained. Then, we use $ARMA(p,q)$ to fit X_t, that is:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t - (\theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}), t \in Z \quad (5\text{-}11)$$

Where, the first half represents the automatic regression process, and $\varphi_1, \cdots, \varphi_p$ is the regression coefficient; The latter part represents the moving average process, $\theta_1, \cdots, \theta_q$ is the moving average coefficient, and $\varepsilon_{t-q}, \cdots, \varepsilon_t$ is the unobserved white noise sequence, which follows the gaussian distribution.

5.2.2 Solution of ARIMA model

In order to analyze reputation index more intuitively and predict more accurately, we use ARIMA model to predict four secondary indexes respectively. In this paper, we take the product B000FS1W4U in hair_dryer data set as an example to explain in detail. First, by plotting the curve of the data sequence, we find that it is not a stationary time series, because it is growing. So we do a differential treatment of the sequence. After processing, because of $p = 0.1607 > 0.05$, we need to do the difference processing on the sequence again. And then we get $p = 0.035 < 0.05$, which means that the data passed the test. Because two differences are made here, so d is equal to 2.

Then, we draw the ACF and PACF diagrams of this sequence and select the appropriate p and q for the $ARMA(p,q,d)$ model. The autocorrelation value does not exceed the confidence interval after 4-order hysteresis, so we have $p = 4$ .Using the same method, we get $q = 4$.

Through the above process, we determine our model as $ARMA(4,4,2)$. Through Matlab solution, we can obtain the predicted results as shown in figure 12.

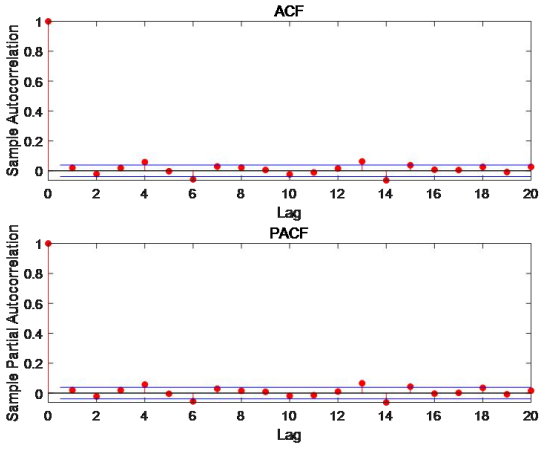

Fig 11: ACF and PACF

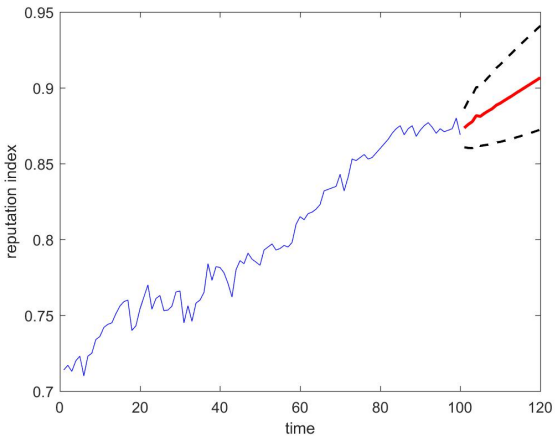

Fig 12: Predicted results

As can be seen from the figure, the reputation of product B000FS1W4U has a gradually rising trend.

5.3 Discovery of potentially successful or failing products

We consider the reputation value of products with potential success or failure to be significantly different from other products. For the convenience of the following description, we have named the abnormal product class and the mediocre product class respectively. Therefore, we use support vector machine discriminators to identify potential successful or failing products.

5.3.1 The SVM recognizer

For 2D values, the SVM classifier trains a hyperplane line to separate two different classes, in this case the abnormal product class and the mediocre product class. The best hyperplane is defined by $\omega^T x+\gamma=0$, ω^T is the normal vector of the hyperplane, and x is the reputation value of the product. An effective hyperplane satisfies:

$$\begin{cases}\omega^T x_i+\gamma\geq 0 & for\ y_i=1\\ \omega^T x_i+\gamma\leq 0 & for\ y_i=-1\end{cases} \tag{5-12}$$

Among them, the training sample points closest to the optimal hyperplane make the equal sign in the above formula true, and they are called "support vectors".

5.3.2 One class SVM

In order to identify potential successful products and failing products, we need to classify the reputation index of different products, because the conditions for judging successful or failing products are not given in this paper. One class of support vector machines is introduced to separate successful products from failing products. The goal of one class SVM is to separate data points from the origin as much as possible in the feature map space. Firstly, the gaussian kernel is used to project the data points onto the feature space, and then the quadratic programming is solved to separate the projected data points from the origin. The expression for judging whether the new sample is an abnormal point is as follows:

$$(z-a)^T(\mathrm{z}-a)>R^2 \tag{5-13}$$

Where a is the weighted sum of all samples and is the radius of the smallest sphere. After introducing gaussian kernel, our discriminant function becomes:

$$k(z,z)-2\sum_i\alpha_i K(z,x^{(i)})+\sum\sum\alpha_i\alpha_j K(x^{(i)},x^{(j)})-R^2 \tag{5-14}$$

Where ω is the minimum radius, ζ_{i} is relaxation, $\mathrm{v}\in(0,1)$ determines the upper limit of the outliers. Finally, a line is trained in the feature map space to distinguish the origin from the data. The line and data are then converted back into the unprojected space, with outliers.

5.3.3 Identify potential successes and failures

Our SVM recognizer combines SVMS and LSVMS. We first use one class of support vector machines (SVM) to determine outliers for product reputation values. These

outliers indicate the existence of potentially successful or failing products. Then one class support vector machine is used to train the linear support vector machine as the threshold of potential success or failure of the product. Taking the hair_dryer data set as an example, we found the corresponding successful products and failing products. Some results are shown in table 4:

Table 4: Potentially successful or failing products

product_id	B003V264WW	B000BW5W50	B000BVB27S
result	successful	failing	failing
product_id	B0009XH6TG	B000C1Z1JM	B00132ZG3U
result	successful	failing	successful

5.3.4 Find the corresponding combination

After finding a potential success or failing product, we can find its corresponding index value. Taking the data in table 4 as an example, we can find that the reputation value of successful products, including the secondary indicators that constitute the reputation value, is higher than that of mediocre products. In contrast, the reputation value of failed products and their secondary indicators are lower than those of mediocre products.

5.4 The influence of star ratings on subsequent reviews

5.4.1 Matthew effect

The Matthew Effect, a phenomenon in which the strong get stronger and the weak get weaker, is widely used in psychology, education, finance and science. For example, prominent scientists tend to get more credit than unknown researchers; Even if their achievements are similar (Robert k. Merton,1968).

Similarly, we define the Matthew effect of customer reviews as: when customers see a higher or lower star review, they are more likely to write a higher or lower star review.

5.4.2 Correlation analysis

Correlation analysis using Pearson correlation coefficient is a common method to verify Matthew effect[8].Before we solve the correlation coefficient, we formulate the Matthew effect.

For a product P_i, we use its initial t comments to form the comment time series $X_{pi}(t), i = 1, \cdots, n$, and n is the total number of comments recorded for P_i. For another product P_j, if $X_{pi}(t_1) < X_{pj}(t_1)$, then:

$$X_{\mathrm{pi}}(t_2) - X_{pi}(t_1) < X_{pj}(t_2) - X_{pj}(t_1) \tag{5-15}$$

Where t_1 and t_2 are two moments and $t_2 > t_1$. This formula indicates that the increment of total stars will still be larger (smaller) in the sales of products with higher (lower) total review stars at a certain time. Here, the increment of total stars reflects the

level of rating of post-order customer reviews, which is consistent with the Matthew effect of customer comments as defined by us.

Now, we set $t_2 = t_1 + k$, that is, after the time point t_1, the product has added k review data. Now, we will analyze the correlation between the new total review star data at time t_2 and the old total review star data at time t_1, and observe whether the subsequent comments are related to the previous comments.

The mathematical description of Pearson's correlation coefficient is as follows:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}} \tag{5-16}$$

According to the size of r_{XY}, make the following judgment:

• If $r_{XY} \in (0.5,1]$, it is believed that there is a strong positive correlation between the total stars of the new review and the total stars of the old review, indicating that the Matthew effect of customer reviews is relatively obvious.

• If the $r_{XY} \in (0.3,0.5]$, it is considered that there is a weak positive correlation between the total stars of the new review and the total stars of the old review, indicating that the Matthew effect of the customer reviews is not obvious.

• If $r_{XY} \in [-1,0.3]$, there is no correlation or even a negative correlation between the new review total stars and the old review total stars, indicating that the Matthew effect of customer reviews does not exist.

We extracted two items each from the hair dryer, microwave and pacifier data sets, one of which had a higher overall star rating and the other a lower one. For $X_p(t_2)$, we get it by collecting all the historical review records; For $X_p(t_1)$, we get it by removing the most recent k records on the basis of $X_p(t_2)$. By taking $k = 20$ and using SPSS for correlation analysis, the results are shown in table 5:

Table 5: Results of correlation analysis

review_id	B0009XH6TG	B001UE7D2I	B0055UBB4O
PCC	0.820	0.755	0.964
review_id	B0052G51AQ	B0045I6IA4	B0028H3ACS
PCC	0.859	0.897	0.713

It turns out that our guess is correct and that there is indeed a Matthew effect in customer reviews. That is to say, a high star rating will arouse customers' desire for the same high star rating; Conversely, when customers see a series of low-star ratings, they are more likely to make similar ratings.

5.5 The association between descriptors and star levels

5.5.1 Association rule mining：Apriori

Association rule mining is one of the most active research methods in data mining. It was first proposed by Agrawal et al to solve the "shopping basket problem" in 1993 to find the association rules between different products in the mall transaction database. For example, if we find the following rules in the mall transaction database:

$$\{\text{bread, milk} \rightarrow \text{coffee}\}$$

This suggests that if customers buy bread and milk at the same time, they are more likely to buy coffee as well. In other words, there was a high correlation between bread, milk and coffee.

Based on this idea, we use Apriori algorithm to find the potential correlation between descriptors and star levels. At this point, each descriptor in the review can be considered an item in the shopping mall. We also see five type of stars as five goods. The question then turns out that when a customer buys a "star" item, he is more likely to buy a "descriptor" item together, and vice versa. Next, we will describe the establishment process of Apriori model in detail.

Step 1: data one hot coding

Since the input data of Apriori algorithm is in Boolean form, we need to convert the data in text form into the data in one hot coding form. The following example shows the process of one heat coding.

Suppose there are three review records, and their stars and reviews are shown in table 6:

Table 6: Data in text form

ID	star level	review
1	5	very useful
2	5	This is useful to me
3	1	so bad

Let $I = \{i_1, i_2, \cdots, i_n\}$ be the collection of "goods" in the shopping mall, then:

$$I = [5, very, useful, this, is, to, me, 1, so, bad]$$

Let $D = \{t_1, t_2, \cdots, t_m\}$ be the collection of records in the database, then:

$$D = \{1,2,3\}$$

Since "useful" appears in the first and second records, it can be expressed as $[1,1,0]$. The remaining codes are shown in table 7:

Table 7: Data in the form of one hot code

ID \ Factors	5	very	use-ful	this	is	to	me	1	so	bad
1	1	1	1	0	0	0	0	0	0	0
2	1	0	1	1	1	1	1	0	0	0
3	0	0	0	0	0	0	0	1	1	1

Step 2: formalized definition

• **Association rule:** represent the relationships between items, as follows

$$X => Y \quad where\ X, Y \subseteq I \tag{5-17}$$

Where X and Y are disjoint item sets, both of which are subsets of I, where X is called the antecedent and Y is called the antecedent.

• **Support:** X's support for T is defined as the ratio of the goods in transaction T to the goods in transaction T in both transaction T and item set X.

$$supp(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \tag{5-18}$$

• **Confidence:** an indicator of the probability of a rule being discovered. The confidence value of the rule (X→Y) relative to a set of transaction T is the ratio of transactions that contain X to transactions that also contain Y. Defined as:

$$\text{conf}(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{5-19}$$

• **Lift:** reflects the correlation between X and Y in association rules.

$$\text{lift}(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)} \tag{5-20}$$

Lift greater than 1 indicates that there is a strong correlation between X and Y, while lift less than or equal to 1 indicates that there is no effective strong correlation between X and Y.

Step 3: find the association rule between descriptors and star levels

The frequent item set is found by Apriori algorithm and used to generate rules that satisfy the lowest confidence. The specific steps of Apriori algorithm are shown in figure 13. Apriori has two important laws:

• **Law 1:** if a set is a frequent item set, all its subsets are frequent item sets.

• **Law 2:** if a set is not a frequent item set, then all of its supersets are not frequent item sets.

By using these two laws, the pruning effect can be achieved in solving the problem, so as to avoid the exponential growth of the number of item sets and control the time of calculating frequent item sets in a reasonable range.

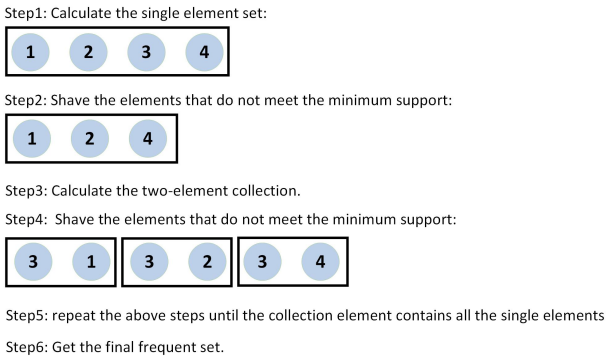

Fig 13: The steps of Apriori

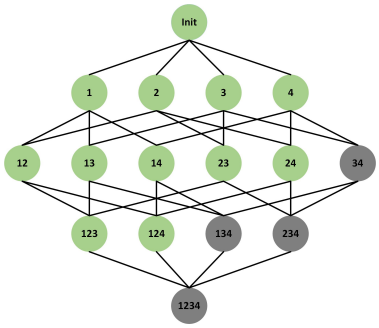

Fig 14: Apriori's law

An example is shown in 14. Gray circles represent infrequent item sets, such as {3,4} . From law two we can know that since the superset of {1,3,4} {2,3,4} {1,2,3,4} is {3,4} , they are not frequent item sets. That is, after the support of {3,4} has been calculated, there is no need to calculate the support of {1,3,4} {2,3,4} {1,2,3,4} .

Then we can calculate the association rules according to the frequent item sets obtained. Since a frequent itemset can generate multiple association rules, we need to try to construct pruning rules to prevent the association rules from growing exponentially. Through observation, we find that if a rule does not meet the minimum confidence requirement, then all subsets of the rule do not meet the minimum confidence requirement. Thus, we can effectively reduce the number of association rules.

Step 4: calculate the descriptors that are most relevant to each star level

Taking the hair_dryer data set as an example, we use Apriori to find out the corresponding association rules, as shown in table 8:

Table 8: Apriori results

rating level	quality descriptor
5	hot,powerful,love
4	great,good
3	back
2	problem
1	longer time,waste money

We find that the high star rating in hair_dryer data set is more correlated with “hot”, which suggested that the hair dryer with relatively high temperature is more popular among customers. Also, low-star rating is more correlated with “longer time” and “waste money”, which indicates that customers do not like time-consuming hair dryer and think it is a waste of money.

6 Sensitivity analysis

In section 5.4, the value of parameter k is manually specified. Will the change of k have a big impact on the result of the solution or even change the conclusion? Therefore, on the basis of the original $k = 20$ we adjust the corresponding parameters and observed the changes of the results. The results are shown in table 9:

Table 9: Results of sensitivity analysis

k=15	**review_id**	B0009XH6TG	B001UE7D2I	B0055UBB4O
	PCC	0.788(↓4%)	0.770(↑2%)	0.945(↓2%)
	review_id	B0052G51AQ	B0045I6IA4	B0028H3ACS
	PCC	0.833(↓3%)	0.870(↓3%)	0.720(↑1%)
k=25	**review_id**	B0009XH6TG	B001UE7D2I	B0055UBB4O
	PCC	0.804(↓2%)	0.793(↑5%)	0.954(↓1%)
	review_id	B0052G51AQ	B0045I6IA4	B0028H3ACS
	PCC	0.885(↑3%)	0.924(↑3%)	0.727(↑2%)

As shown in the results, the change of k value has an impact on the solved Pearson correlation value, but all PCC values are still greater than 0.5, which indicates that the conclusion that there is Matthew effect in customer reviews has not changed, and our model is robust and reliable.

7 Strengths and Weaknesses

7.1 Strengths

• LDA topic model can well analyze the weight value of each word in the topic, so as to facilitate the conversion of text data into numerical data.

• Combining EWM and TOPSIS, we can fully consider all indicators to make a comprehensive evaluation on the reputation of the product for Sunshine Company to track and analyze.

• The introduction of the Matthew effect, an interdisciplinary concept, to analyze the impact of previous reviews on subsequent reviews is an innovation and a good result has been achieved.

• Apriori algorithm can help us to obtain the association between descriptors and star levels, and at the same time to effectively find the association between descriptor phrases and star levels.

7.2 Weaknesses

• Apriori algorithm runs for a long time, and the solution results need simple manual screening.

Letter

To: Marketing Director of Sunshine Company

From: Team 2004898

Date: March 9, 2020

Subject: The result of our team

Dear director, we are honored to inform you of our achievements after data analysis and modeling.

First, we use LDA model to convert qualitative Review data in text form into quantitative *Pos _ factor* data in numerical form, so that customers' likes and dislikes can be judged by intuitively observing the value. We also use the k-means algorithm to find the internal relationship among Star Rating, Review and Helpfulness Rating, and divide the review into four categories: fuzzy poor review, clear poor review, fuzzy favorable review and clear favorable review. In this way, when you select data for market research, you can directly choose clear negative reviews and clear positive reviews for research, so as to avoid the waste of time caused by the research on fuzzy reviews.

Next, we use the EWM-TOPSIS model to build the "reputation index", which allows you to track the competitiveness of products in the market by observing their reputation index. We also built a time series model with ARIMA, which can help you easily observe whether a product's reputation is rising or falling.

We also use the SVM algorithm to identify potential successful or failing products. We find that in order for a product to become a potential successful product, it should perform very well in all the secondary indicators of the reputation index.

In addition, through correlation analysis, we find that there is Matthew effect in customers' shopping reviews, and existing bad reviews will incite future customers to make the same bad reviews. Therefore, if the product you sell has some low star rating, it is necessary to take corresponding improvement measures for the problems reflected in the rating in time, so as to prevent the continuous decline of product reputation caused by the Matthew effect.

Finally, using Apriori algorithm, we find that, for hair dryer, a high star level often corresponds to "hot", "quickly", while a low star level often corresponds to "waste money", "longer time". Therefore, your hair dryer should be able to produce a high temperature to dry hair quickly, and the price should not be too high.

For microwave oven, a high star level often corresponds to "fit", "easy", and a low star level often corresponds to "repair", "warranty". Therefore, your microwave oven should be easy to use and provide excellent after-sales service.

For pacifier, a high star level often corresponds to "quality", "soft", "clean", while a low star level often corresponds to "old", "shape", "small", and "hard". Therefore, your baby pacifier should be soft, clean and designed to fit babies.

So that's the summary of our research. We sincerely hope that it can provide you with useful information and look forward to your reply. Thank you!

References

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

[2]Junkui Wang. Research on the usefulness of online reviews for e-commerce websites. Diss. Xidian University, 2014.

[3]Cao, Qing, Wenjing Duan, and Qiwei Gan. "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach." Decision Support Systems 50.2 (2011): 511-521.

[4]"LDA (LDA Document Topic Generation Model) _Baidu Encyclopedia". Baike.Baidu.Com, 2020, https://baike.baidu.com/item/LDA/13489644?fr=aladdin.

[5] Chen Lei, and Wang Yanzhang. "Research on Evaluation Method of Entropy Weight Coefficient and TOPSIS Integrated Evaluation." Control and Decision 18.4 (2003): 456-459.

[6]"TOPSIS Method_Baidu Encyclopedia". Baike.Baidu.Com, 2020, https://baike.baidu.com/item/TOPSIS%E6%B3%95/3094166?fr=aladdin.

[7]"ARIMA Model_Baidu Encyclopedia". Baike.Baidu.Com, 2020, https://baike.baidu.com/item/ARIMA%E6%A8%A1%E5%9E%8B/10611682?fr=aladdin.

[8]Tu Xianfeng. "Empirical Analysis of the Matthew Effect in Online Retail Transactions." E-Commerce 2 (2015): 7-9.