ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

Data Measures of Customer Reviews

He Chuning, Liu Ya Qi Lydia, Peggy Pan

Shenzhen Middle School, China

Abstract

After the transaction is completed, the review data of Amazon mall can be used to guide the subsequent sales of the analyzed products. Using the data of hair dryer, nipple and microwave provided by the Sunshine Company, a series of models are built for analysis. Using the LDA theme model, the comment content can be digitized. Using the k-means clustering model, the comment data could be classified. Using the spearman correlation coefficient, the correlation of internal parameters of the comment could be analyzed, and finally the trend of product's change could be studied by the seasonal time series model. Through the LDA theme model, it was found that the majority of comments were in line with the star rating after the text comments were digitized, because the star rating and the comment data result are similar. According to the correlation coefficient analysis of the data, it can be known that whether people think the comments are helpful or not has little relation with the level of the star rating. However, the higher the star rating, people tend not to give the help rating, and those who give the help rating, think it is more helpful. The relationship between help comments and the content of comments is similar to the relationship between star ratings and help comments. According to the seasonal time series model, the reputation of microwave ovens tends to decline over time, the reputation of pacifiers tends to rise over time, and the reputation of hair dryers tends to stay at a high level, although the reputation of the middle declines, but soon rises back up. According to the data of LDA model, there is little difference between the content of comments and the star rating, which also means that some words with emotional color are closely related to the rating level. It turned out that the microwave oven was the worst product and the hair dryer the most successful one. At last, sensitivity analysis is made and our models perform well. Then a letter summarizing our findings and advice is written to the Marketing Director of Sunshine Company.

Keywords

LDA theme model; K-means clustering model; Seasonal time series model.

1. Introduction

1.1. Background

Sunshine Company provided data on microwave ovens, baby pacifiers and hair dryers sold on amazon.com over time. The data includes customer ratings and evaluations of the product, as well as helpful ratings from other users for those reviews. Sunshine had not previously used the data for analysis before. They were interested in time-based patterns in the data to see if they would help sell products. Based on this, Sunshine needs to build appropriate mathematical models to analyze the data and guide the sales.

1.2. Our works

Task 1 The commodity review is digitized and processed quantitatively and qualitatively. Giving Sunshine Company a rough idea of the potential implications of the data set.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

Task 2 Mathematical models were used to get correlations between the evaluations. Investigate whether a particular star rating affects the reputation of a product and leads to more specific reviews

Task 3 By establishing a mathematical model to study the changing rule of product data set over time, it is shown that the changing rule of product reputation in the market over time.

Task 4 Identify potential most successful and most unsuccessful products based on reviews or ratings.

2. The Description of the Problem

2.1. Problem Statement

Amazon's online store has the function of rating and scoring. Now we need to build a mathematical model based on the rating and evaluation data of microwave oven, baby pacifier and hair dryer provided by the Sunshine Company. Based on this mathematical model, we need to help sunshine company analyze the change rule of product reputation over time, the correlation between score and rating and other factors. Through this research, it is helpful to the Sunshine company's three new products in the success of Amazon.

2.2. Analysis of Specific Issues

2.2.1. Analysis of Task 1

For task 1, we need to convert comment data to standard data, and make descriptive statistics on the data. By means of clustering, the corresponding evaluation of each product is clustered. According to the clustering results, the evaluation is divided into categories such as good, average and poor…

2.2.2. Analysis of Task 2

Task 2 needs to establish a mathematical model to study the relationship among star rating, comments and help rating of Sunshine Company. Therefore, a product can be taken as an example to obtain its internal relationship through the correlation coefficient among the star rating, rating and help rating of the product.

2.2.3. Analysis of Task 3

For task 3, we need to understand how the reputation in each dataset changes over time. According to task 2, star rating and review rating are strongly and positively correlated. Since the review rating can better reflect the real evaluation of buyers, the change of review rating can be used to reflect the change of product reputation.

2.2.4. Analysis of Task 4

For task 4, we need to determine the success and failure of the three products according to comments and star rating. Therefore, we can get inspiration from the sequence diagram in task 3 and the clustering analysis in task 1, and judge from multiple perspectives.

3. Basic Assumption

Vine users have more credibility in their comments

The comments of users who have not completed the transaction are less credible.

The interrelationships within a product can represent the characteristics of a sunshine product

When importing data, very few data were lost, which was ignored when using SPSS because the lost data was very small.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

4. Glossary & Symbols

4.1. Glossary

The LDA model [1-2]: LDA (Latent Dirichlet Allocation) is a document topic generation model, also known as a three-tier bayesian probability model, which contains three-tier structures of words, topics and documents

Clustering algorithm [3-4]: Cluster analysis, also known as group analysis, is a statistical analysis method to study classification problems (samples or indicators), and also an important algorithm for data mining.

Time series diagram [5-6]: It is a statistical graph with time as the horizontal axis and observation variables as the vertical axis to reflect the relationship between time and quantity.

4.2. Symbols

Symbols	Definition
R	User provided comments
RH	Customer comment value on the product
SR	People's ratings of the product
V	VIP comment user
k	The LDA model determines the number of topics
N	Number of clustering centers

5. Models

5.1. Analysis and Solving of Task One

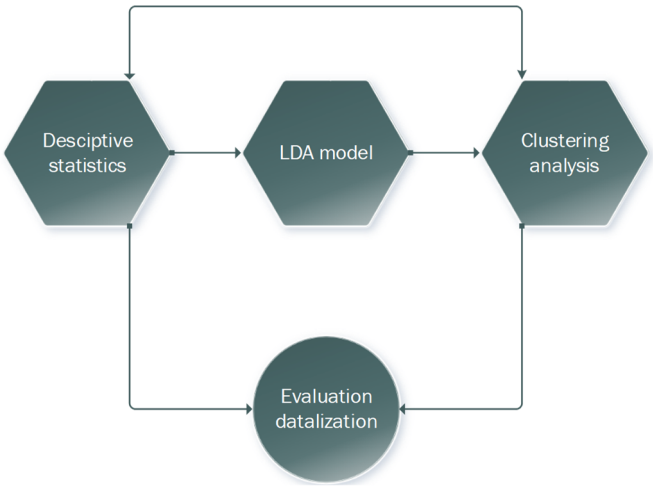

Fig 1. A mind map for task 1

5.1.1. Model Preparation

(1) Data Processing

Firstly, the evaluation data of hair dryer, baby pacifier and microwave oven should be fully digitized.

1. Convert non-vine comments to 0 and vine comments to 1.

2. Convert the verified purchase comments to 1 and the verified purchase comments to 0.

3. Put the title of the comment and the content of the comment in the same text, and digitize them through the LDA model.

(2) Assumptions

a. Vine users have more credibility in their comments, so when the LDA model is applied to digitize the comments, the judgment of the comments will be relatively loose. That is to say,

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

when we categorize, we consider as many terms as possible from vine users in order to characterize their comments.

b. The comments of users who have not completed the transaction are less credible, so it is more strict to regulate their words in the process of data transformation using LDA model.

c. In task 1, the comment data processing method of the baby pacifier and microwave oven is consistent with that of the hair dryer. Therefore, in task 1, the comment data of the hair dryer is temporarily taken as an example to construct quantitative and qualitative models. The other two product processing results will be involved in the following task

(3) The Foundation of Model

Before the quantitative and qualitative processing of comments, the text of comments should be digitized, and the LDA model is adopted. The schematic diagram is shown below.

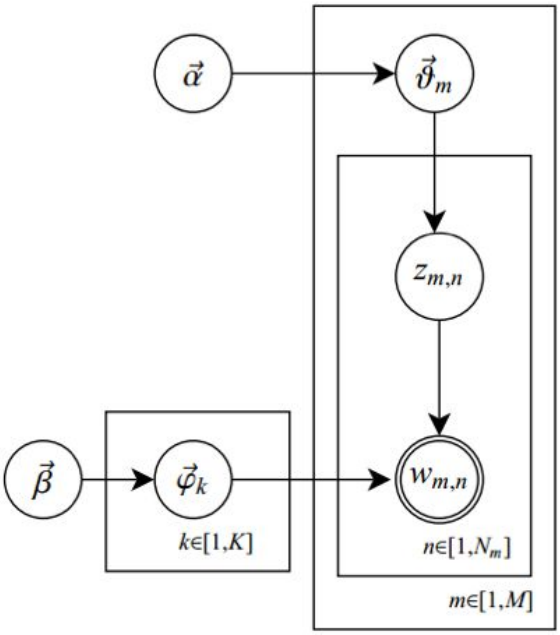

Fig 2. Schematic diagram of LDA clustering model

After digitizing the table of comments, the K-means clustering model is adopted in order to qualitative the comments, and its basic flow chart is shown in the following figure3:

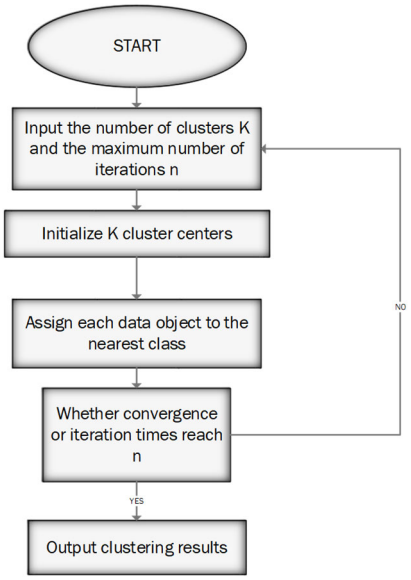

Fig 3. Flow chart of k-means clustering analysis

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

5.1.2. Model Establishment

The number of topics in the LDA model is denoted by k and its value is 5. So that it can be consistent with the star rating. The topics are called excellent, very good, average, poor, and very poor respectively. They are represented by 5, 4, 3, 2 and 1. The criteria for ordinary users are shown in the figure 4 below, and for vine users and those who have not completed the transaction are classified in the same way. But, as stated in the hypothesis, the word frequency criteria of the classification are modified accordingly.

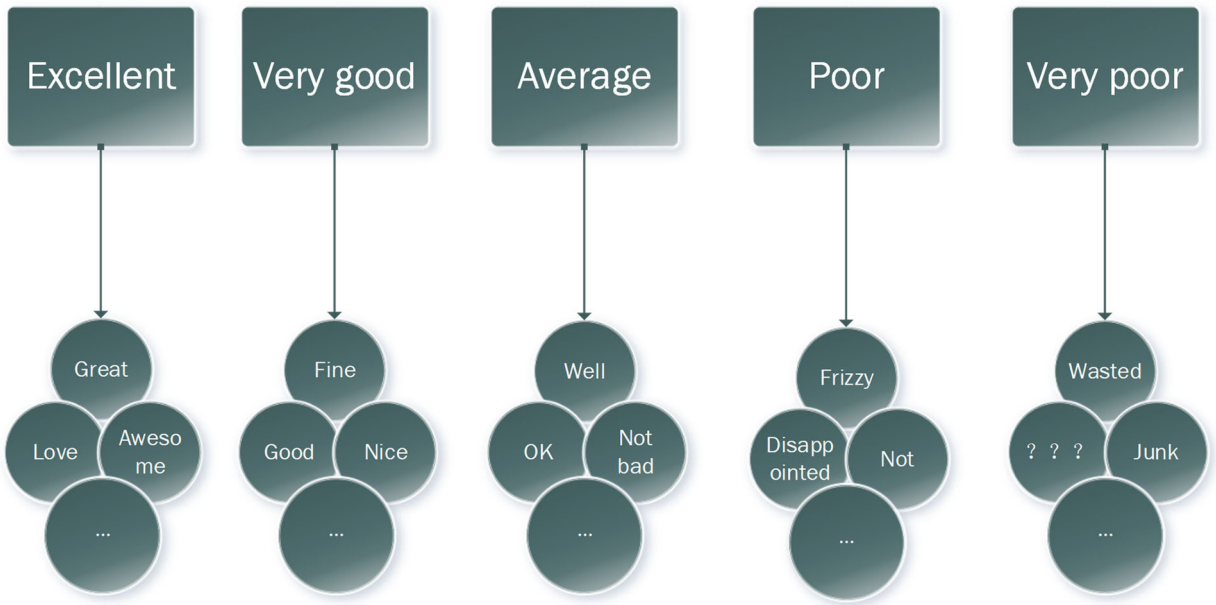

Fig 4. The correspondence between the word frequency of common users and the topic

For the data evaluation content, using the K-means clustering method in SPSS, The cluster number is set to 5, which is the same as the star rating. The initial clustering centers of the hair dryer are shown in the following Table1.

Table 1. The initial clustering centers of the hair dryer

	1	2	3	4	5
Serial number	1	3069	11468	774	5399
Star rating	5	5	5	5	4
Helpful votes	0	229	4	53	224
Useless votes	0	6	0	4	6
review	5	5	5	5	3
vine	0	0	0	0	0
Verified purchase	1	0	1	1	1

5.1.3. Results

Through the LDA model, the comments of the hair dryer are digitized and the data is filled into the following Table 2

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

Table 2. Hair dryer comments after digitized sample sheet

Objects / Serial number	Star rating	Helpful votes	Useless votes	Review	vine	Verified purchase
1	5	0	0	5	0	1
2	4	0	0	2	0	1
3	5	0	1	5	0	1

Descriptive statistics were conducted on all the data, and the results of the hair dryer are shown in the following Table 3.

Table 3. Descriptive statistics of hair dryer digitized comments

Objects	star_rating	helpful_votes	useless_votes	review	vine	verified_purchase
Average	4.116041848	2.17907585	0.383783784	4.0543156	0.01560593	0.855361813
Standard error	0.012141517	0.132974401	0.01720464	0.0129482	0.00115735	0.003284382
The median	5	0	0	5	0	1
Mode	5	0	0	5	0	1
The standard deviation	1.30033324	14.24130378	1.842584017	1.3867314	0.12395049	0.351751004
Variance	1.690866536	202.8147335	3.395115859	1.9230239	0.01536372	0.123728769
Kurtosis	0.531090093	406.1021131	1006.259976	-0.216908	59.1203573	2.084330957
Partial degrees	-1.350129055	17.76193968	24.24993621	-1.143215	7.81729164	-2.020882863
Area	4	499	98	4	1	1
The minimum	1	0	0	1	0	0
The maximum	5	499	98	5	1	1
Sum	47211	24994	4402	46503	179	9811
Number of observations	11470	11470	11470	11470	11470	11470

The final clustering center of k-means clustering analysis is shown in the following Table 4, The number of cases in each cluster of the hair dryer is shown in the following Figure 5:

Table 4. The final clustering center

	1	2	3	4	5
Serial number	860	4816	10043	2692	7291
Star rating	4	4	4	4	4
Helpful votes	1	1	6	1	1
Useless votes	0	0	1	0	0
review	4	4	4	4	4
vine	0	0	0	0	0
Verified purchase	1	1	1	1	1

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

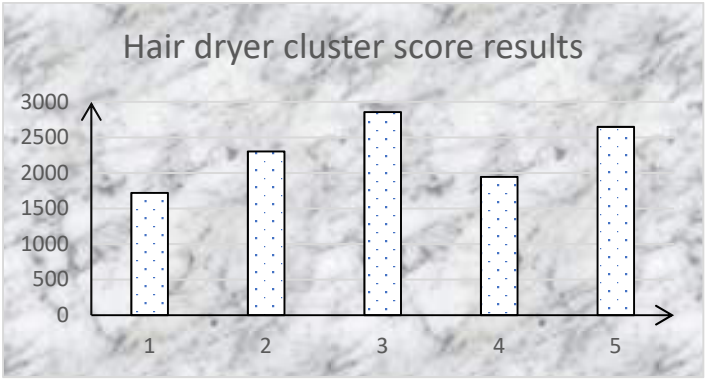

Fig 5. The number of units under each cluster center

5.1.4. Analysis of the Result

According to the statistical results of description after the comments were digitized in the LDA model, the star rating was roughly the same as the comment rating, although the star rating was slightly higher than the comment rating. Customers who bought hair dryers were the most likely to give five-star ratings and customers tend to think reviews are helpful for their shopping. According to the results of k-means clustering, both star rating and comment rating are grade 4, that is to say, the public thinks the hair dryer buying experience is a good choice. According to the data of LDA model, there is little difference between the content of comments and the star rating, which also means that some words with emotional color are closely related to the rating level.

5.2. Analysis and Solving of Task Two

Task 2 needs to establish a mathematical model to study the relationship among star rating, comments and help rating of Sunshine Company. Therefore, a product can be taken as an example to obtain its internal relationship through the correlation coefficient among the star rating, rating and help rating.

5.2.1. Model Preparation

(1) Data Processing

Task 2 follows the data of task 1 and takes the evaluation data of hair dryer as an example to build a correlation coefficient model to describe the internal relationship among star rating, comment rating and help rating of Sunshine Company.

(2) Assumptions

1. Since all the three products belong to the Sunshine company, it is assumed that the correlation between factors such as the infant pacifier, the microwave oven data set's star rating, the score, the help score and so on are consistent with the data set of the hair dryer.

2. The samples were sampled separately from each other.

(3) The Foundation of Model

The calculation formula of the spearman correlation coefficient is shown in the following formula1:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (1)$$

5.2.2. Model Establishment

After the normal distribution test, it is found that some comment data of the hair dryer do not conform to the normal distribution, so the spearman correlation coefficient model can be used

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

to study the correlation of various elements in the data set. Drawing a matrix scatter diagram initially describes the correlation between the elements as shown in the following Figure 6.

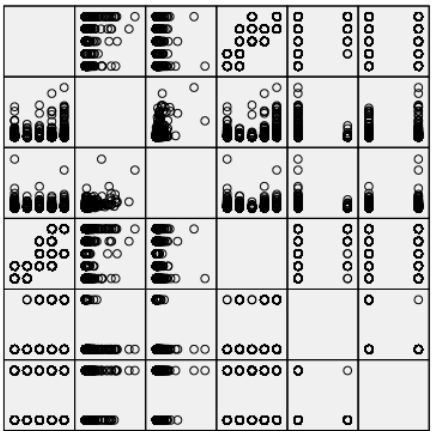

Fig 6. Internal matrix scatter diagram of hair dryer data

The spearman phase relation table is obtained by the spearman correlation coefficient function calculation method of matlab, which is shown in the results.

5.2.3. Results

Table 5 shows the spearman phase Correlation coefficient, All the results have passed the p-value test.

Table 5. Hair dryer data internal phase Correlation coefficient

	Star_rating	Helpful_votes	Useless_votes	Review	Vine	Verified_purchase
Star_rating	1	−0. 1738538	−0. 2027136	0. 91208961	0. 01501038	0. 106875071
Helpful_votes	−0. 1738538	1	0. 43864309	−0. 1620472	0. 03936175	−0. 195739912
Useless_votes	−0. 2027136	0. 43864309	1	−0. 1842716	0. 09292503	−0. 154810873
Review	0. 91208961	−0. 1620472	−0. 1842716	1	0. 020925	0. 097350745
Vine	0. 01501038	0. 03936175	0. 09292503	0. 020925	1	−0. 304192122
Verified_purchase	0. 10687507	−0. 1957399	−0. 1548109	0. 09735075	−0. 3041921	1

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

5.2.4. Analysis of the Result

From the results, it can be found that the correlation between the star rating and the comment is 0.92, showing a strong positive correlation, indicating that the star rating and the comment content of Sunshine company's product are basically the same. The correlation between the star rating and the help comment is weak and negative, which proves that whether people think the comments are helpful or not has little relation with the level of the star rating. However, the higher the star rating, people tend to not give the help rating, and those who give the help rating think it is more helpful. The relationship between help comments and the content of comments is similar to the relationship between star ratings and help comments.

5.3. Analysis and Solving of Task Three

For task 3, we need to understand how the reputation in each dataset changed over time. According to task 2, star rating and review rating are strongly and positively correlated. Since review rating can better reflect the real evaluation of buyers, the change of review rating can be used to reflect the change of product reputation.

5.3.1. Model Preparation

(1) Data Processing

Since it is necessary to know the changes of reputation in the data sets of the three products over time, not only the data-based results of the comments on the hair dryer in task 1, but also the data-based results of the comments on the other two products are needed. The following two figures show the trend of the contents of pacifier and microwave oven changing with time after digitization, respectively.[7-8]

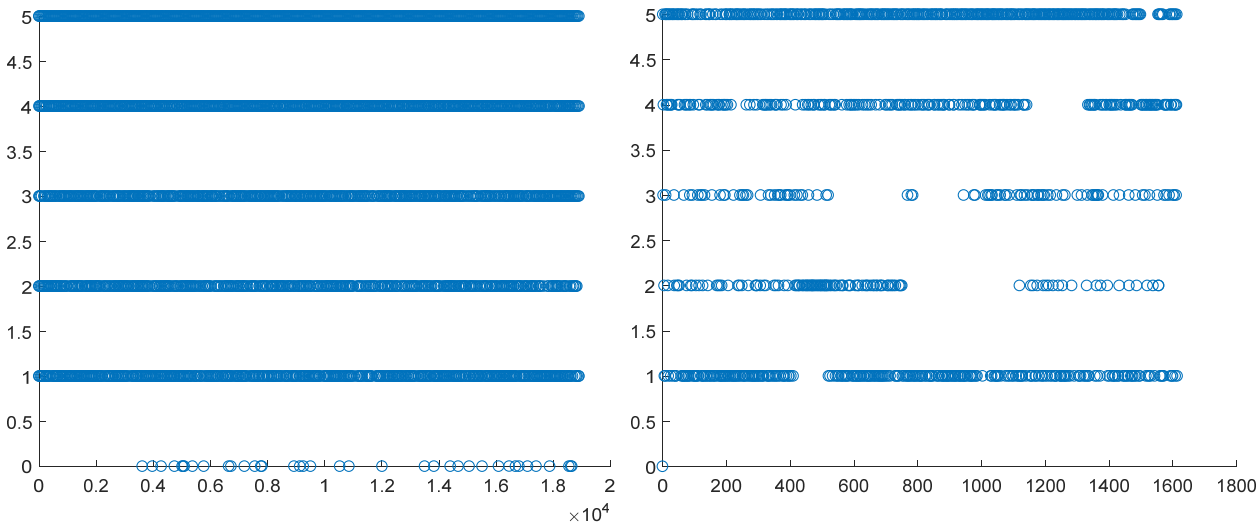

Fig 7. The trend of the contents of pacifier and microwave oven changing with time

(2) Assumptions

In the data set of the pacifier, individual star ratings exceeded the maximum value and were ignored

(3) The Foundation of Model

Time series, also known as dynamic sequences, refers to the numerical sequences that arranges the index values of a certain phenomenon in a chronological order. Its basic framework is shown in the following figure 8.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

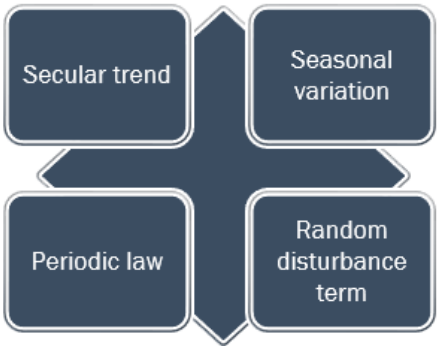

Fig 8. The basic framework of a time series diagram

5.3.2. Model Establishment

According to the rule of the data set, we firstly decomposed the comment scores according to the seasonality, and set the time and date of four quarters each year. According to the season, 4 data are selected each year from the data processed in task 1 to use SPSS to draw a seasonal time series diagram, if there are more data in the current quarter, the average is taken. The following Table 6 shows the model fitting degree of the hair dryer comment data when establishing the time series model.

Table 6. The model fitting degree of the hair dryer comment data

Model fitting degree

Fitting a statistical	Avera'ge	Standard error	MIN	MAX	percentile						
					5	10	25	50	75	90	95
R party smoothly	.260	.368	6.661E-16	.521	6.661E-16	6.661E-16	6.661E-16	.260	.521	.521	.521
R2	.438	.619	6.661E-16	.876	6.661E-16	6.661E-16	6.661E-16	.438	.876	.876	.876
RMSE	.585	.566	.185	.985	.185	.185	.185	.585	.985	.985	.985
MAPE	14.553	15.360	3.692	25.415	3.692	3.692	3.692	14.553	25.415	25.415	25.415
MaxAPE	162.736	209.272	14.759	310.714	14.759	14.759	14.759	162.736	310.714	310.714	310.714
MAE	.439	.417	.144	.733	.144	.144	.144	.439	.733	.733	.733
MaxAE	1.784	1.872	.460	3.107	.460	.460	.460	1.784	3.107	3.107	3.107
BIC	-1.593	2.312	-3.228	.042	-3.228	-3.228	-3.228	-1.593	.042	.042	.042

5.3.3. Results

The seasonal time series of microwave ovens, baby pacifiers, and hair dryers are shown in the figure below. The blue line is the trend change of the original comment, and the red line is the seasonal trend.

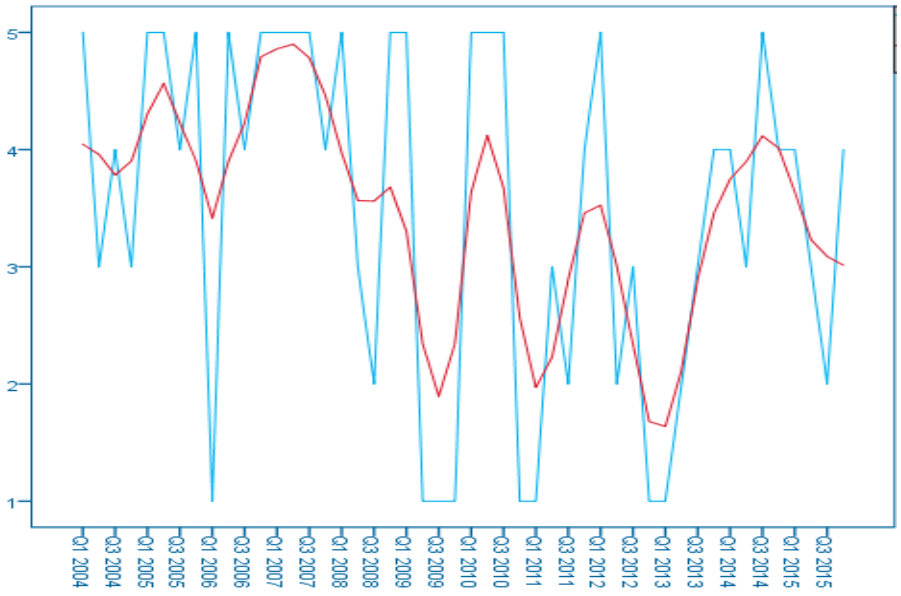

Fig 9. Trend chart of microwave reputation over time

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

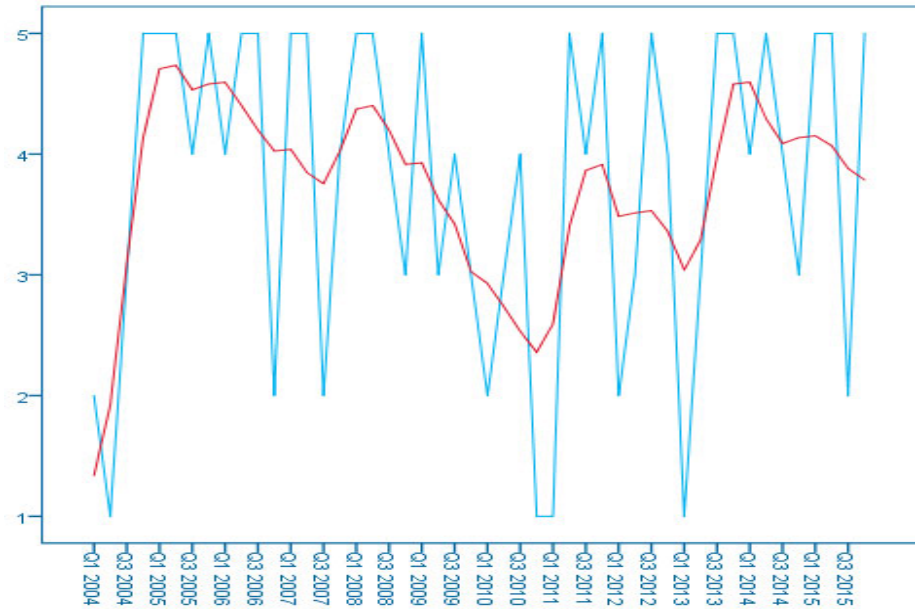

Fig 10. Chart of the trend of pacifier reputation over time

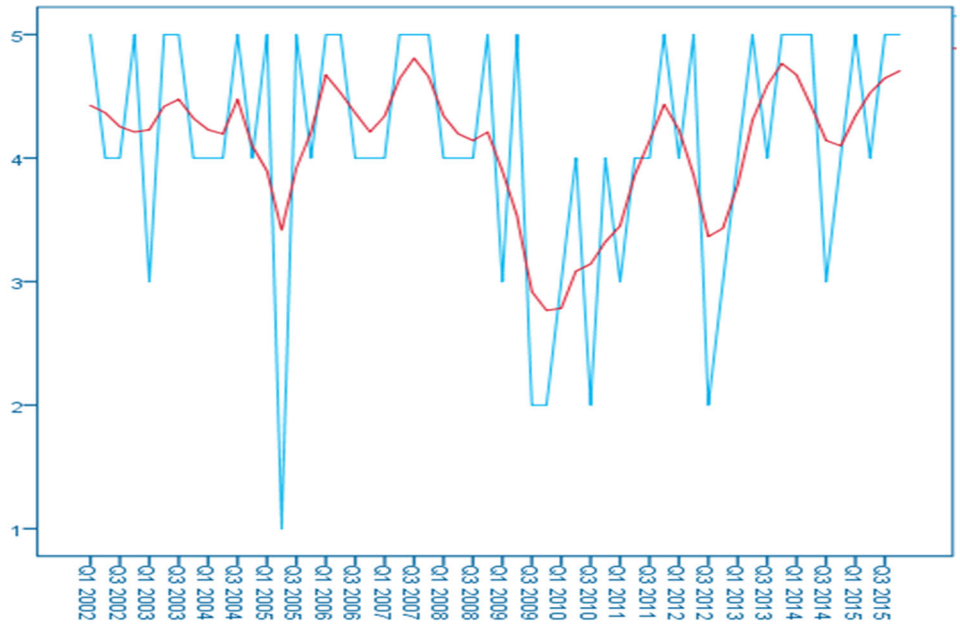

Fig 11. Trend chart of hair dryers' reputation with time

5.3.4. Analysis of the Result

The seasonal timing chart of the three product review data sets shows that the reputation of microwave ovens tends to decline over time, the reputation of pacifiers tends to rise over time, and the reputation of hair dryers tends to stay at a high level, although the reputation of the middle declines, but soon rises back up.

5.4. Analysis and Solving of Task Four

For task 4, we need to determine the success and failure of the three products according to comments and star rating. Therefore, we can get inspiration from the sequence diagram in task 3 and the clustering analysis in task 1, and judge from multiple perspectives.

5.4.1. Model Establishment and Results

Task 1 has shown the clustering results of the evaluation data of the hair dryer. Now let's look at the clustering results of the evaluation data of the microwave oven and the baby pacifier.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

Table 7. The final clustering center of microwave evaluation data

The final clustering center

	clustering				
	1	2	3	4	5
Serial number	237	1095	509	1450	690
star_rating	4	3	5	3	4
helpful_votes	1	6	814	11	4
Total_votes	2	7	848	13	5
review	4	3	5	3	4
vine	0	0	1	0	0
verified_purchase	1	1	0	0	1

Table 8. The final cluster center of infant pacifier evaluation data

The final clustering center

	clustering				
	1	2	3	4	5
Serial number	1033	16171	3419	10971	6658
star_rating	4	4	4	4	4
helpful_votes	0	2	0	1	0
total_votes	1	2	1	1	1
review	4	4	4	4	4
vine	0	0	0	0	0
verified_purchase	1	1	1	1	1

5.4.2. Analysis of the Result

From the clustering results, the score of the hair dryer was set at 4 points, the clustering results of the microwave oven were inconsistent, the average clustering result was 3.8 points, and the clustering result of the baby nipple score was also 4 points. Combined with the conclusion obtained from task 3, we can perceive that microwave oven is the most failed product, not only the overall score is low, but also the reputation declines year by year. The hair dryer is the most successful product, the score and reputation are stable at a high point.

6. Error Analysis and Sensitivity Analysis

6.1. Error Analysis

6.1.1. Error Analysis of Model One

The main error in task 1 is that the representation of words under each topic in the LDA model is not necessarily accurate. Even with modifications, the purchase comments of vine users and unconfirmed users are not necessarily representative. Although the k-means clustering method divides the comments into 5 categories, the results fail to converge.

6.1.2. Error Analysis of Model Two

The error of model 2 is that the calculation accuracy of the spearman correlation coefficient is limited. Although the P value of some results is greater than 0.05, the value greater than 0.05 is very limited, which may lead to errors in the calculation of correlation.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

6.1.3. Error Analysis of Model Three

The error of model 3 is that the seasonal time series adopts the method of averaging when selecting the representative values of each season, which may lead to the fact that the selected data cannot represent the score of the current season.

6.2. Sensitivity Analysis

The sensitivity of the model is general. For the first model, the LDA model is responsible for digitizing the comments. If the number of topics is increased, the results will be more accurate, and the results of each operation may produce different results. The k-means clustering model has no convergence and poor sensitivity, while the time series model is sensitive to the change of seasonal representative data.

7. Evaluation and Promotion of Model

7.1. Strength and Weakness

7.1.1. Strength

The LDA model can well translate textual comments into comment ratings.

The k-means clustering model can divide the data into the required categories.

The seasonal time series model can reflect the product over time.

7.1.2. Weakness

The representation of words under each topic in the LDA model is not necessarily accurate.

The calculation accuracy of the spearman correlation coefficient is limited.

The selected data may not represent the score of the current season.

7.2. Promotion

The LDA model should adopt more words as the classification basis, the k-means clustering model should improve the algorithm to make the results converge, and the seasonal time series model can adopt a more scientific sampling method when selecting the data representing each season.

8. Conclusions

8.1. Conclusions of the Problem

The star rating was roughly as same as the comment rating, although the star rating was slightly higher than the comment rating. Customers who bought hair dryers were the most likely to give five-star ratings. According to the results of k-means clustering, both star rating and comment rating are grade 4, that is to say, the public thinks the experience of buying hair dryer is a good choice. According to the data of LDA model, there is little difference between the content of comments and the star rating, which also means that some words with emotional color are closely related to the rating level.

The star rating and the comment content of the sunshine company's product are basically the same. Whether people think the comments are helpful or not has little relation with the level of the star rating. The higher the star rating, people tend to not give the help rating, and those who give the help rating think it is more helpful. The relationship between help comments and the content of comments is similar to the relationship between star ratings and help comments.

The reputation of microwave ovens tends to decline over time. The reputation of pacifiers tends to rise over time. The reputation of hair dryers tends to stay at a high level, although there is a decline, it soon rises back up.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202101_4(1).0002

Microwave oven is the most failed product, not only the overall score is low, but also the reputation declines year by year. The hair dryer is the most successful product, the score and reputation are stable at a high point.

8.2. Methods Used in Our Models

LDA theme model

K means clustering model

Seasonal time series model

Acknowledgements

These authors are contributed equally to this work.

References

[1] Hamed Jelodar,Yongli Wang,Chi Yuan,Xia Feng,Xiahui Jiang,Yanchao Li,Liang Zhao. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey[J]. Multimedia Tools and Applications,2019,78(11).

[2] Massimiliano Giacalone,Raffaele Mattera,Eugenia Nissi. Economic indicators forecasting in presence of seasonal patterns: time series revision and prediction accuracy[J]. Quality & Quantity,2020,54(1).

[3] Juliet Waterkeyn, Anthony Waterkeyn, Fausca Uwingabire, Julia Pantoglou, Amans Ntakarutimana, Marcie Mbirira, Joseph Katabarwa, Zachary Bigirimana, Sandy Cairncross, Richard Carter. The value of monitoring data in a process evaluation of hygiene behaviour change in Community Health Clubs to explain findings from a cluster-randomised controlled trial in Rwanda[J]. BMC Public Health,2020,20(1).

[4] Wang Xu Wen,Zhang Yu,Guo Zhen,Li Jiao. Identifying concepts from medical images via transfer learning and image retrieval.[J]. Mathematical biosciences and engineering : MBE,2019,16(4).

[5] Anna Newton-Levinson,Megan Higdon,Jessica Sales,Laurie Gaydos,Roger Rochat. Context matters: Using mixed methods timelines to provide an accessible and integrated visual for complex program evaluation data[J]. Evaluation and Program Planning,2020,80.

[6] Hyunjoong Kim,Han Kyul Kim,Sungzoon Cho. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling[J]. Expert Systems With Applications,2020,150.

[7] Li Li,Song Qifa,Yang Xi. Categorization of β-cell capacity in patients with obesity via OGTT using K-means clustering.[J]. Endocrine connections,2020,9(2).

[8] Kim Yoonhee,Ratnam J V,Doi Takeshi,Morioka Yushi,Behera Swadhin,Tsuzuki Ataru,Minakawa Noboru,Sweijd Neville,Kruger Philip,Maharaj Rajendra,Imai Chisato Chrissy,Ng Chris Fook Sheng,Chung Yeonseung,Hashizume Masahiro. Publisher Correction: Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model.[J]. Scientific reports,2020,10(1).