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

Tracking reviews for better market occupancy

Summary

With the online market's large review data, customers become easier to find the products they want. Analyzing these reviews cloud helps companies design products and capture the market in the same way.

For question 1, this article preprocesses the data, deletes redundant data, and stores the data hierarchically in the form of a tree diagram.

For the A question, this article uses the three characteristics of review time, review length, and review star rating to measure the review value. The review value is digitized with useful votes. A multiple linear regression method is used to fit a function of the review value about the three features. Characterize the product attributes with the number of reviews, and then get the final relationship.

For the B question, the solution to detect one product's reputation is stars multiplies helpful votes on time-line because it is closely related to market sharing. It is supposed to aware that there remains a retardation time which is 30 days long. Based on the result of question B, we can predict products' market sharing by fame line and competitors' reviews for the C question.

For the D question, we use the DBCAN model to explore the review time-line. In positive review time zone, the average of review stars is 4.38, compared with 3.01 in negative time zone and 3.94 in the whole time, the possibility of giving higher star-rating increased 45% and 11.6%, which means high and low star ratings all incite more reviews of the same kind.

For the E question, the result is that: quality descriptors of text-based reviews strongly associated with rating levels. With kurtosis and skewness tests, these data do not obey Gaussian distribution, so our group chooses Spearman-test to calculate Correlation-Coefficient. By semantic analysis, we get scores of sentences based on word embedding and the mood of words. The Correlation-Coefficient of microwave, hairdryer, and pacifier are 0.519, 0.494 and 0.379 between star rating and sentence score, all significant at the 0.01 level(2-tailed). Based on word frequency we find some specific quality descriptors of different products.

Finally, we explore the time-line-based product's fame to identify potentially important design features and market strategy. First, microwave and hairdryer are supposed to have good performance in at least one year. Second, for pacifier, microwave and hairdryer, some of the important design features are water-tight, safe and small-size. Last but equally important, the market strategies based on data exploration are providing better delivery and keep the advertisement reliable.

Keywords: Spearman; DBSCAN; Time-line; Word Embedding;

Contents

MEMORANDUM

To: Marketing Director of Sunshine Company
From: Team #2001067
Date: Mar 9th, 2020
Subject: Online sales strategy recommend and important design features

Dear directors, we are honored to inform you about our achievement for tracking comments, select an online sales strategy, and design products after performing data analysis and modeling.

We established a review tracking model to classify the reliability of reviews for you to avoid fake reviews and focus on valuable comments. It is

$$U = \sum_{i=1}^{3} w_i x_i$$

where U represents the reliability of comments, x_1, $x_2$and $x_3$represents text length, star rating and how long it has been from the latest update. When U≥0.64, this comment is reliable.

To explore the market strategy, we build a Nature Language Processing (*NLP*) model, track the reviews based on the time-line. Based on the feature extraction, we set the online sales strategy as follows:

- When products' market share decrease, there's a high potential that there remain authentic negative reviews(long text, using radical words). So pay attention to the inflection point in your market occupation line, there may appear several low-star reviews one month ago. Please track the latest negative review to improve products.
- The company is supposed to provide better delivery because bad delivery will make customers fell disappointed(47 in 499 reviews cite "deliver" for microwave, 29.1% under two-stars).
- It is recommended to keep advertisements' reality. Customers are unwilling to give positive reviews when they feel cheated by advertisement: reviews cite "Advertise" have 11.3% lower potential to reach four-star.

Then we analyze the products potential design features by word frequency and semantic analysis, the results we get are as follows:

- Improve the quality of products to maintain company's market fame, because customers may update reviews after several months due to quality problems.
- Pacifier should avoid leaking problems due to reviews cites leak have 34% higher potential to be under 2-stars.
- Customers like small-sized hair-dryer because it is easy to put into luggage whi-le traveling. (87% reviews are positive when it comes to small size or traveling)
- Microwave is supposed to have anti-scald design due to 54.4% reviewers who have children at home give negative comments about heat and protection.

We sincerely hope that your company would occupy the market!

I. Introduction

Background

The development of e-commerce and the improvement of the level of logistics services have made more and more people choose to shop online. Due to the large variety and quantity of online products, in order to help customers better understand the products they want to buy, and improve online transaction customers Satisfaction, e-commerce websites have launched an online customer review system. By allowing consumers to evaluate the products they have purchased, other consumers can obtain more information than product descriptions, which greatly reduces the information between purchases and sales. Asymmetry, and this active and effective feedback mechanism can increase consumer trust, thereby reducing consumer perception of risk.

Due to the increasing number of reviews, e-commerce websites should help consumers identify valuable remarks to save time and increase consumer satisfaction, so mining the impact of reviews and ratings on user shopping has become a top priority.

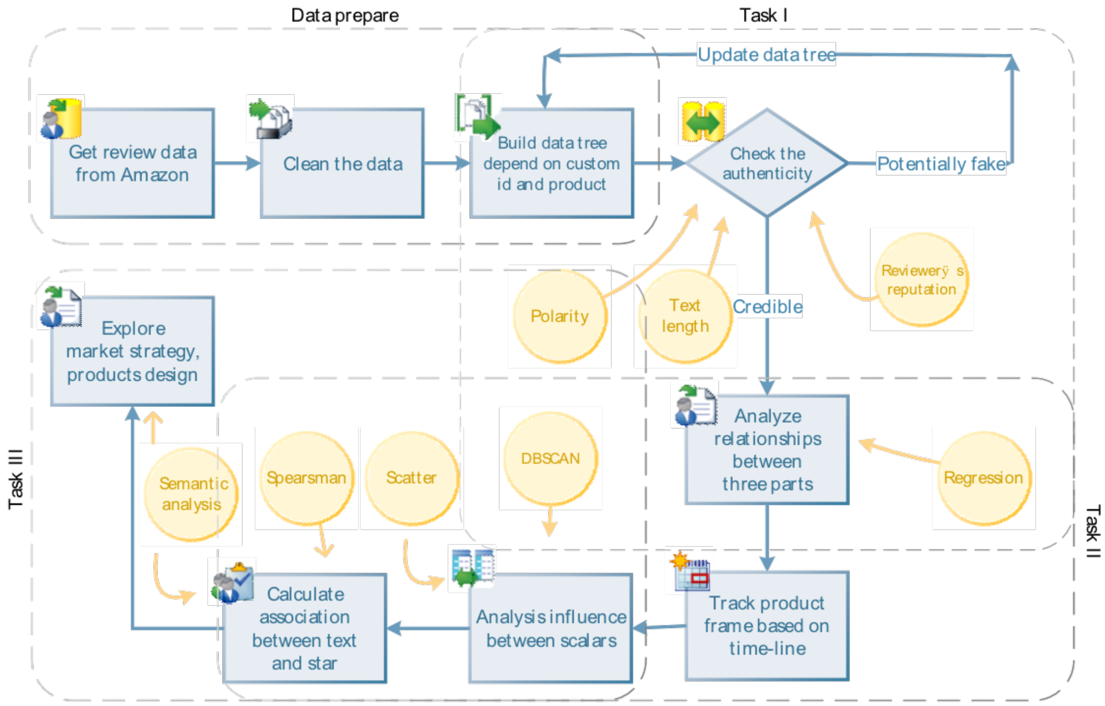
Our works

- Task 1

  We preprocesses the data, deletes redundant data, and stores the data hierarchically in the form of a tree diagram.
- Task 2

  We track the time-line and use multiple linear regression method, semantic analysis, and word frequency statistics to analysis relationships between review time, review text, and review star rating. And how they indicates the product's fame and success in market.
- Task 3

  Using word frequency statistics and nature language processing model, we mining the data and figure out some potential significant product design features and the online sales strategy.


Data prepare
Task I
Update data tree
Get review data from Amazon
Clean the data
Build data tree depend on custom id and product
Check the authenticity
Potentially fake
Polarity
Text length
Credible
Reviewerÿ s reputation
Task III
Explore market strategy, products design
Semantic analysis
Spearsman
Scatter
DBSCAN
Analyze relationships between three parts
Regression
Task II
Calculate association between text and star
Analysis influence between scalars
Track product frame based on time-line

II. The Description of the Problem

Problem statement

Amazon offers customers the opportunity to rate and review purchases. Inform customers of their online sales strategies by identifying key patterns, relationships, metrics, and parameters related to other competing products provided by customers in the past;Identify potentially important design features to increase product appeal. Sunshine Company has used data to inform sales strategies in the past, but they have never used this particular combination and data type before. Of particular interest to Sunshine Company is the time-based patterns in these data and whether they interact in a way that helps the company make successful products.

Analysis of Specific Issues

2.2.1 Analysis of Problem 1

This question can be analyzed based on the above analysis statistics to analyze this statistic combined with references, etc. to express some concerns. What will happen to each threshold range that is worth each statistic? This question is actually an analysis of the results of Model 1.

2.2.2 Analysis of Problem 2(a)

First, based on the nature of the data, the data is screened for completeness, redundancy, etc .;
Then merge similar data, standardize the data for easy processing, and convert text into mathematical symbol representation;
Finally, a function model is constructed to achieve the purpose of establishing the input-output relationship and reflecting the evaluation criteria.

2.2.3 Analysis of Problem 2(b)

Relevance processing is performed on ratings and star ratings, and data with consistent star ratings and reviews are filtered out. At this time, the roles of star ratings and reviews are consistent. Two columns of data are redundant and one column can be deleted

2.2.4 Analysis of Problem 2(c)

This method is similar to b. It is to analyze in different logical modes to find the key factors that can affect the inflection point of product ratings. Here you can still combine text reviews and ratings, and then look for an "event point or cause that can predict future product word of mouth.

2.2.5 Analysis of Problem 2(d)

n this question, it is only necessary to analyze whether there will be more concentrated positive or negative reviews over a period of time based on the time pattern of the previous questions.

2.2.6 Analysis of Problem 2(e)

This is a point I mentioned in a, how to associate the content of the review with the rating. The e-question does require a semantic analysis of the text and does involve knowledge of "NLP".

III. Basic assumption

1. Comment length has a positive impact on comment usefulness.
2. The word customers use always represent his or her attitude towards this product

IV. Symbols

Symbols	*Definition*	*Units*
x_1	Comment duration	Days
x_2	Star rating	Stars
x_3	Comment length	Words
U	Useful votes	dimensionless
Q	Social index of development	dimensionless
ω_i	Corresponding weight of x_i	dimensionless

V. Models

Analysis and Solving of Question One

Redundant data processing

Because marketplaces are all US, they can be removed as redundant data. The review id only plays a role of labeling the comments. We are more concerned about the specific content of the review than its id, so it is also removed as redundant data. , Product parent, product title, product category are also excluded as redundant data.

For the customer id, considering the case where the same id appears twice and the case of purchasing different products with the same id, further mining analysis is performed to find out why the customer is a repeat customer and whether there is a relationship between the products; but In actual data processing, the same ID only appears at most 2 times, and the number of repeated IDs is very small, which is not representative, so it can be deleted as redundant data.

Preliminary analysis and processing of remaining data

For the product id, use it as a basis for classification.

For stars, helpful votes and total votes, build a three-level evaluation system, use the stars to measure the quality of the product, use the total votes and useful votes to measure the credibility of the rating, and construct a function to fit its weight. Note that The point is that if the sentiment and rating of the review do not match, the rating is ignored.

For the green label, considering the conditions for obtaining the green label, the green label can be considered to be more representative. Therefore, for the green label review, the weight should be appropriately increased.

Whether to confirm the purchase: As the only condition for judging whether the evaluation is based, if the purchase is not made, the review is considered to be worthless to prevent black powder and write a favorable review.

For comment titles and comment subjects, they are used to analyze their emotions and perform word frequency statistics on high-frequency words for further analysis.

For review dates, this is used for time series analysis.

Based on the above analysis, we process the data in the form of a tree, as shown in Figure 1

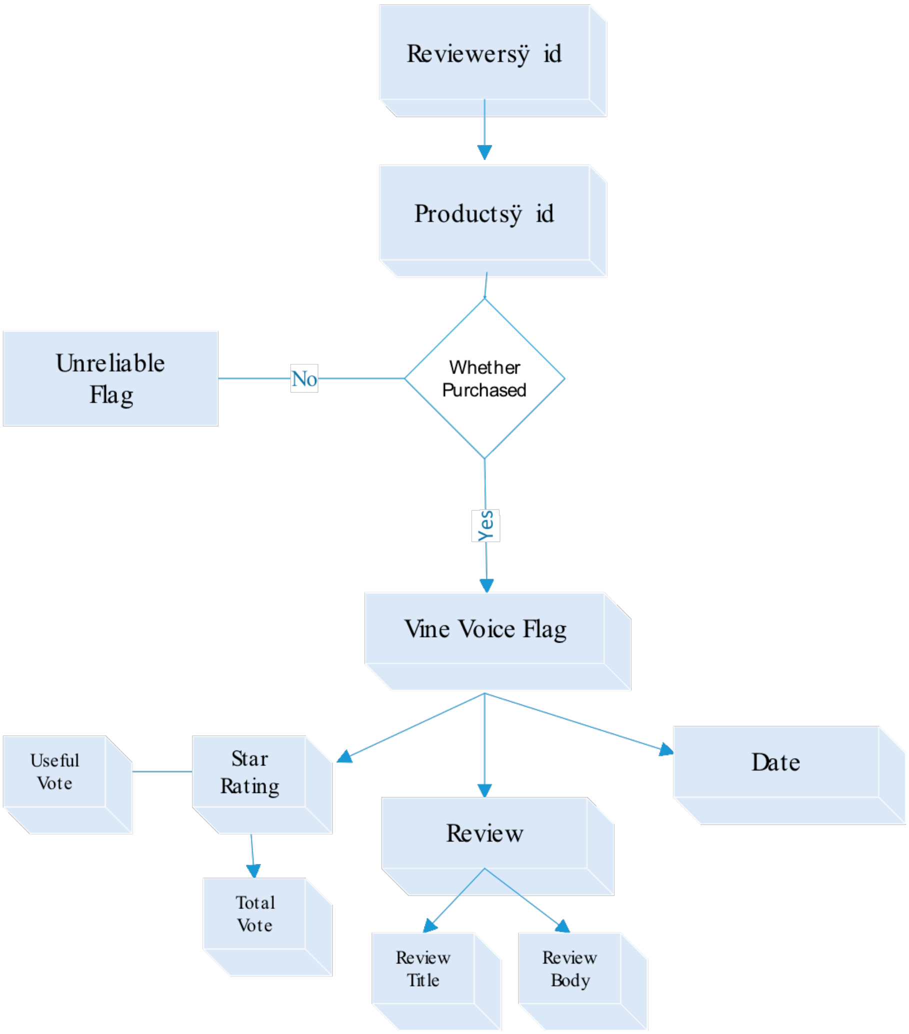


Fig.1

4.1.1 Word frequency statistics

In order to study the star rating given by the "topic" element in the specific content of the review, the degree of usefulness of the review readers and the effect on the reputation of the reviewers, we have adopted statistics on the frequency of certain keywords. method.

First of all, we know that a simple vocabulary has a series of derivative words, which may appear in the subject predicate attributive or other parts, such as the word 'problem' expressing possible defects of the product. The form exists and expresses the same meaning, so we use the word cloud method to perform a second-level word frequency statistics: first use a word cloud composed of a simple word and its derivative words as the same word to count the frequency, and then count the words How often each member of the cloud appears in the word cloud in the context of the review.

The "topic" element of our main research is obviously more appropriate to use the word cloud frequency to describe, so we define the word frequency f as follows:

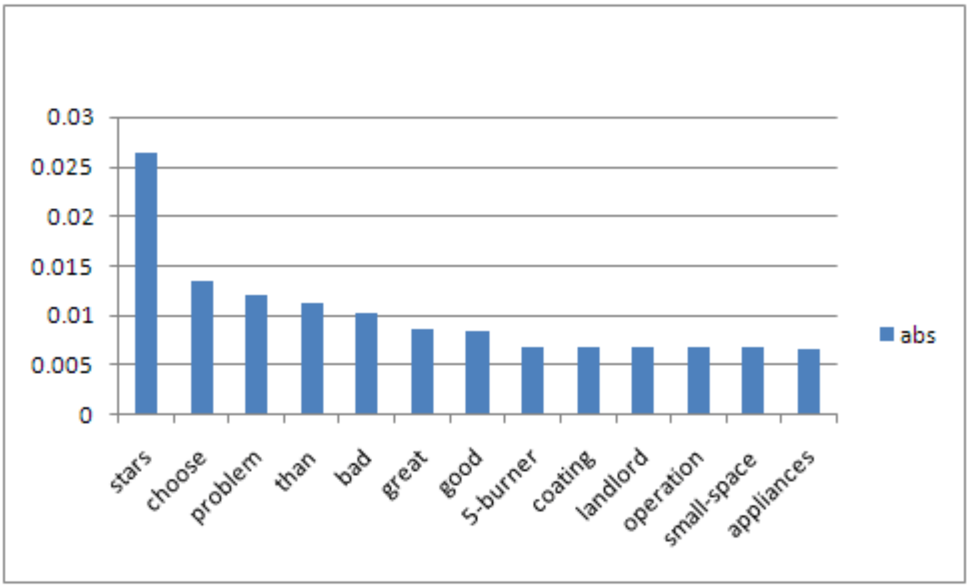


Fig.2

If a total of a type of words are used by the statistical review text objects, they are written as w1, w2, ..., wa, and there are n (w1), n (w2), ..., n (wa), respectively, forming b word clouds , C1, C2, ... Cb, b <= a, then

$$f(C_i) = \sum_{w_j \in C_i} n(w_j)/N$$

(N is the total number of words in the text objects being counted.

Then we counted the word frequency of the title and the content of the comment in turn, the word frequency of the comment title alone, the word frequency of the comment content, and the word frequency of each star rating comment. , The inflection point of the product's market share occurred some time ago for further research.

Analysis and Solving of Question Two（a）

Impact of comment length on comment usefulness

The length of a review can increase the recognizability of the information, especially when the information is available without additional search costs. The open review content of the website provides further explanation and explanation for the digital star rating, and affects other consumers' perception of the usefulness of the review. When buying a product, consumers often spend a lot of time comparing among several alternatives that have already been selected, but they lack the confidence to make purchasing decisions. The consumer may have a positive tendency for a product, but he has not analyzed the main reasons for choosing the product, or made a list of advantages and disadvantages; or the consumer has a negative tendency for a product, but lacks Motivation to process other optional product information. In these cases, an in-depth review from the consumer who did the work will help the consumer make a purchasing decision.

Longer reviews usually contain more product details and usage in different situations. Can reduce the uncertainty of product information and reduce consumers' perception of risk. Consumers'

online shopping is actually a decision-making process. In the process, there is a lot of uncertainty. Therefore, it is necessary to search for useful information to reduce this uncertainty. Before making a purchasing decision. Writing reviews is time-consuming and labor-intensive, and there are no incentives to write reviews on Amazon. Consumers write reviews solely because they are passionate about information sharing, not to get points for the word count. Longer reviews may contain more information and attract consumers to read them. At present, most websites filter out comments with less words as spam comments. Dangdang also limits the number of words that can be used for comments. Comments with too few words cannot be published successfully. So the longer this article is, the more useful information it contains. Having more information can increase consumer confidence in making purchasing decisions. Therefore, this article makes the following assumptions:
H1: Comment length has a positive impact on comment usefulness.

Impact of time on comment usefulness

Among a large number of comments, some comments can get a lot of votes, while others have only a small number of votes or even no one cares. This is not only related to the usefulness of the comments. Some comments are of high quality, but they can only get a small number of votes. Racherla and Friske believe that the number of votes a review receives is related to the length of time it appears on the site 2 . Since the comments were published early, since there are not many comments, each comment can be read by consumers, and the chances of getting a vote are great. Amazon.com sorts reviews by their usefulness by default, which means that some newly posted reviews are ranked behind because they have no votes. Comment
The permutation order has a serious impact on consumer behavior. With the increase in reviews, consumers often only read the reviews on the first few pages. In this way, the comments that rank first will get more and more votes, and the chances of getting votes later will be less and less, unless consumers choose to sort by the time of publication. Therefore, this article introduces the comment publication time as a control variable, and uses the number of days from the date when the comment is published to the date when the data is collected to measure.
This article uses useful votes to measure the usefulness of evaluation.
In summary, all variables and their metrics are summarized in Table 1.1.

Table 1.1

Variable name	Measure	Type
Comment length	Comment content word count	Continuous
Review extremeness	Star rating（1—5）	Discrete
Comment usefulness	Useful votes	Continuous
Comment time	Number of days from comment posting to collection date	Continuous

All sample description statistics

The description statistics of each variable in all samples in the research model are shown in Table 1.2.

Table 1.2.

Items	N	Minimum	Maximum	Mean	Std evaluation
Comment duration (days)	1845	6	1555	415.86	329.343
Ranking level	1845	1	5	4.96	.340
Useful ticket	1845	1	845	7.85	32.801
Star rating	1845	1	5	3.64	1.445
Comment length	1845	8	2718	119.19	152.820

Statistics of comment length frequency

According to the number of words, the comment length is divided into 0-10 words, 10-20 words, 20-50 words, 50-300 words, and more than 300 words for grouping statistics, as shown in Table 1.3.

Table 1.3

	Frequenc	Percentage	Effective Percentage	Cumulative Percentage
0-10	17	.9	.9	.9
10-20	212	11.5	11.5	12.4
20-50	476	25.8	25.8	38.2
50-300	978	53.0	53.0	91.2
Above	162	8.8	8.8	100.0
Total	1845	100.0	100.0	

regression analysis

Regression analysis is a widely used quantitative analysis method. It is mainly used to analyze the statistical relationship between things. It focuses on the quantitative change of variables. Finally, it reflects and describes this relationship in the form of regression equations to help people accurately grasp how variables are affected by one or more other variables. The degree of influence of the variables, which in turn provides a scientific basis for prediction. This paper uses multiple linear regression analysis to verify the impact of review length, review star rating, and reviewer ranking on review usefulness.

The multiple linear regression analysis method is to study the relationship with the dependent variable through the optimal linear combination of multiple independent variables, and jointly predict or estimate the dependent variable through the combination of multiple variables, which is more effective and consistent than predicting with only one variable. actual. In this paper, a stepwise screening strategy is used when performing multiple linear regression analysis. Stepwise Regression is a commonly used method to eliminate multicollinearity between variables and obtain the "optimal" regression equation. At each step in the calculation process, the partial regression sum of squares (that is, contribution) of the variables that have been introduced into the equation must be calculated, and then the significance test for the least contributing variable is given at a given level of significance. This variable remains in the regression equation, otherwise, if it is not significant, you need to delete the variable, and then follow this step to calculate other unintroduced variables until the variables in the regression equation cannot be eliminated and there is no new Until the variables can be introduced, the stepwise regression process ends.

First, for all sample data, a multivariate linear regression analysis is performed between the respective variables and the usefulness of the reviews. Due to the non-normal distribution of comment duration, comment length, and useful votes, the comment duration and useful votes are logarithmic, and the comment length is standardized, and outliers are removed. In addition, in order to study the impact of star ratings on the usefulness of reviews, two variables, star ratings and star squares, are introduced. If the coefficients of star terms are negative and the coefficients of star terms are positive, it indicates the shape relationship, and vice versa. For inverted relationship. The final analysis results are shown in Table 1.5.

Table 1.5(a)

Model	Sum of square	df	Mean square	F	Sig.
Regression	236.147	7	33.735	38.839	
Residual	1481.826	1706	.869		
Total	1717.973	1713			

From Table 1.5(a), it can be seen that the observed value of the F statistic is 38.839, and the corresponding probability P value is approximately 0, reaching a significant level, indicating that it is valid and statistically significant, that is, the explanatory power of the independent variables in the model has reached a significant Level.

Table 1.5(b)

Model	Nonnormalized C	Std Error	Normalized C	t	Sig.
Constant	3.147	1.372		2.293	0.022
Time	0.191	0.027	0.162	7.118	0
Length	0.179	0.054	0.118	3.333	0.001
Star	-0.621	0.125	-1.367	-4.965	0
Star^2	0.063	0.019	0.679	3.243	0.001

It can be seen from Table 1.5(b) that the coefficient of the comment length is positive and significant at a significance level of 0.01, indicating that the comment length is significantly positively related to the usefulness of the comment, assuming H1 is verified.

Analysis and Solving of Question Two（b）

4.1.2 Model establishment

The reputation of a product can often be measured by its market share. Good products occupy most of the market, and ordinary products can only survive. Therefore, for this question, this article uses time as a variable and the time span is 3 days. By examining the market share of a product at different times, you can get a rough idea of the current reputation of the product.

Then use the evaluation model obtained above to comprehensively measure the data of this period of time, you can get the image of the product's wind evaluation about time.

Figure 2 gives a product's market share and wind rating as a function of time:

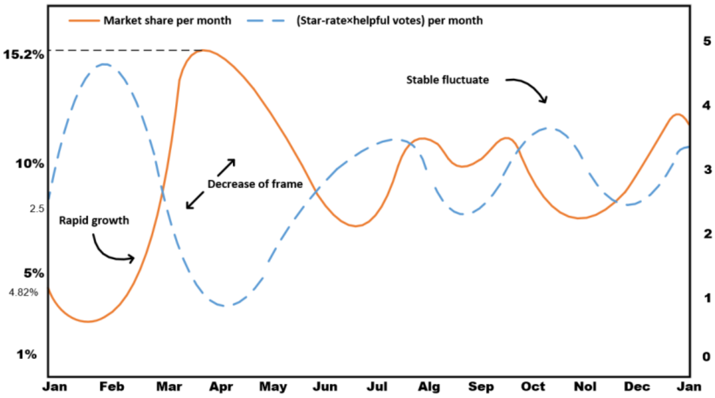


Fig. 3

4.1.3 Analysis and conclusion of the model

Because market regulation has a lag, that is, a product's reputation change always lags behind its wind rating change. From Figure 2, this article can know that the monotonicity of the reputation function is always monotonous with the wind rating function some time ago. Be consistent, with a time difference of about a month. Therefore, this article can give a conclusion that when the monotonicity of the wind assessment function changes, it indicates that the product's reputation is changing, which is specifically manifested as a twist in market share after a period of time.

Analysis and Solving of Question Two（c）

We use the market share to represent the products' succeed, in problem B we find that the products' fame will influence the market share, but there remains a retardation time t. To predict the products' succeed or not in the feature, we are supposed to know lately products' fame.

When it comes to the influence of the fame in the feature, we search the small time zone around the inflection point in the products' fame lines. Analogy to the bull news and bad news in stock market, we use star-rating multiply helpful votes rate to represent the news in products' market. Using microwave for example, we find that there is an increase of the frequency of "dangerous" around maximum point of fame line, which leads to more helpful votes on negative reviews and push customers away, making the market sharing decrease.

To improve the model's capability, we explore whether other companies' fame line influence objective function. The following picture shows an example in hair dryer:

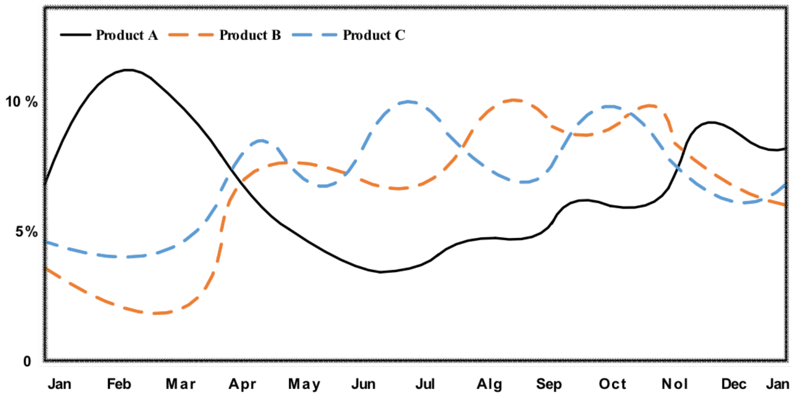


Fig. 4

It shows that when other companies' fame is increasing, there's high potential that objective function will decrease. Taking competitor's products fame into consider, we are able to predict the market share line for longer time steps.

In general, the combinations of measures that best indicate a potentially successful or failing product is the combination of star-rating multiplies helpful votes and competitors' product fame.

Analysis and Solving of Question Two（d）

4.1.4 Model preparation

After processing the timeline for obtaining the comments, we selected the data of the pacifier B00793CZAE to analyze whether there is a herd phenomenon in the comments. We first marked the revised star rating of each comment on the timeline to obtain an overall understanding of the data. as follows:

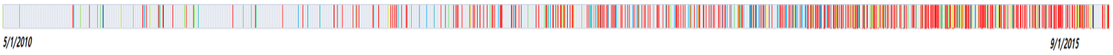


Fig.5

The closer the color is to red, the higher the star rating of the review, the closer to blue, the lower the star rating of the review, and the yellow is in the middle.

From the picture above, you can observe that the thin bars of the same color on the time axis do tend to connect into blocks, especially the five-star (red) comments. After that, we continue to use DBSCAN's clustering method to further quantitatively process the above data to prove this

observation.
The selection of the clustering method takes into account that the reviews change over time. If the stars are directly clustered, the clustering point will be a time point, and what this article hopes for is a time period. Therefore, this paper uses the DBSCAN algorithm. Process star rating data.

4.1.5 Model establishment:

Using DBSCAN's clustering method, several time periods with the highest density of different star reviews on the time axis are obtained. These time periods are marked on the time axis shown above, and the star ratings are evaluated according to the time period. Mode to determine the color of the border of the time period. The color standard is the same as that used to mark the thin line of the comment. The following processed time axis is obtained:

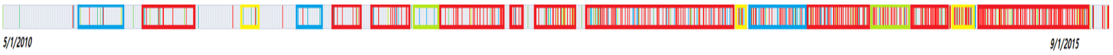


Fig.6

It can be seen that the reviews with the same star rating do show a bar-like distribution. Calculate the revised average number of reviews, the maximum number of reviews and the minimum number of reviews in the adjacent two-star review and five-star review clusters A and B. Number, and compared with the overall value of all reviews for this product. There are obvious differences.
In addition, we found that the wording or format of the review may also cause a certain degree of imitation. For example, it is the above product, and its title contains the word 'star' (that is, it uses the format of giving stars directly in the title). Comments), mark it on the timeline, and get:

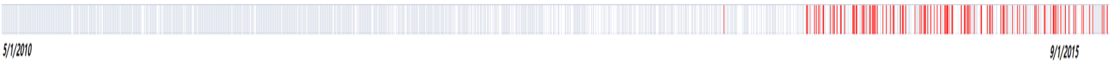


Fig.7

Using DBSCAN for clustering and labeling:

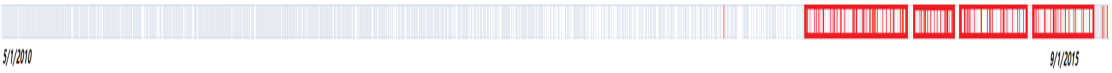


Fig.8

It is found that there is a process from scratch to this review format, and it also shows a block distribution after it appears. Therefore, even the wording or format of the commentary has a certain degree of imitation.

4.1.6 data analysis

As the density of comments increases, the preservation time of topics also becomes longer, that is, a common topic, or the phenomenon of follow-up imitation of comments, will become more obvious and lasting.

Analysis and Solving of Question Two（e）

Model preparation

To classify the positive or negative feeling when customers write down the review, we build convolutional neural networks for sentence classification based on word embedding and the convolutional neural networks model.

First, with the word embedding lexicon, we can express one word by a vector where the size and direction can represent its relationship with other words, just like the following picture shows. In this project, we use a public word embedding lexicon called Glove which calculates word vector by co-occurrence matrix and contains 400000-word vectors.

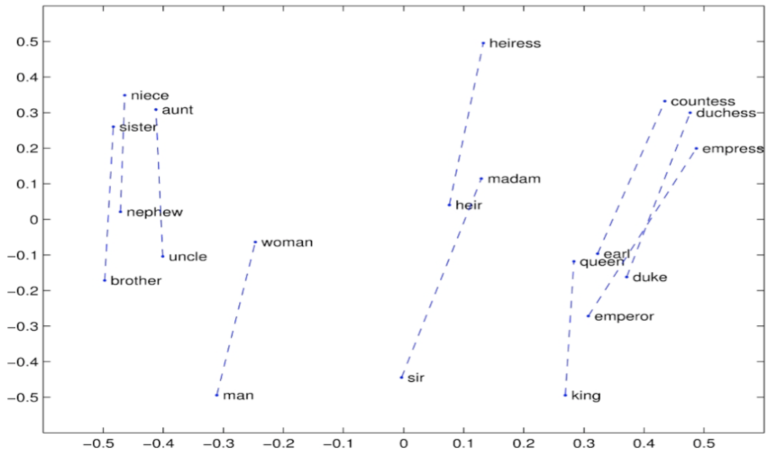


Figure: Word vector representation

Second, let $x_i \in \mathbb{R}^k$ be the $k-$dimensional word vector corresponding to the $i-th$ word in the sentence. A sentence of length n is represented as:

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n$$

where $\oplus$ is the concatenation operator. From here we get a large word matrix. Then with a filter $w \in \mathbb{R}^{h \times k}$ generate a window to monitor h words by:

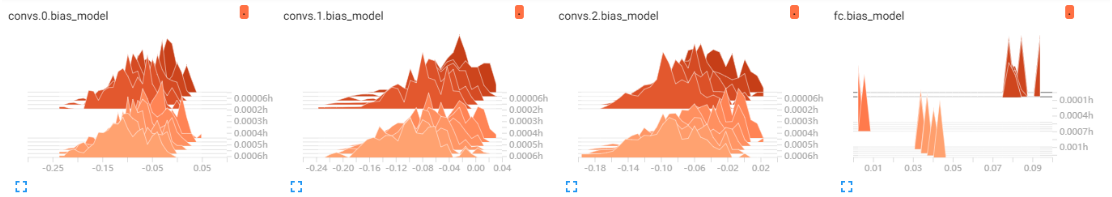
$$c_i = f(w \cdot x_{i:i+h-1} + b)$$

Every h -word-window will be converted by this filter, then create $c = [c_1, c_2, \ldots, c_{n-h+1}]$, after max-over-time-pooling to collect all these vectors we get $\hat{c} = \max\{\mathbf{c}\}$. With m filters, we get the penultimate layer $\mathbf{z} = [\hat{c}_1, \cdots, \hat{c}_m]$.

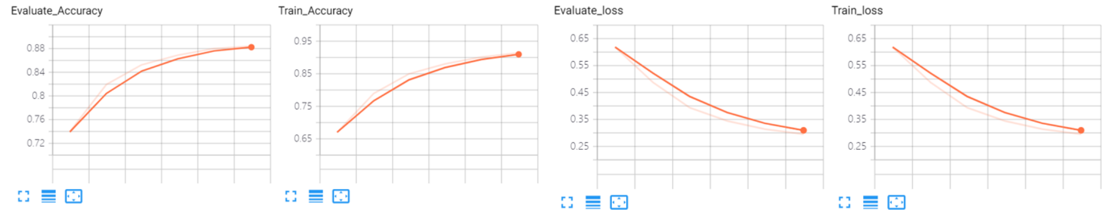
In the linear transformation layer, we use dropout rate r to make the model more robust.

$$y = \mathbf{w} \cdot (\mathbf{z} \circ \mathbf{r}) + b\,, \quad \mathbf{r} \in \mathbb{R}^m$$

According to the model architecture above, we build a sentence classification model. Based on a public dataset called IMDB we train the CNN model several times based on different learning rates seeking for the best model. In the following picture we show the changing of some network weights, it can be seen that during the training process, this model is trying to fit the data



The following picture is one of the training processes. From here you can see the loss is decreasing and training accuracy restrained around 93% and evaluate accuracy restrained around 88%.



Finally our model's evaluate loss restrained at 92%. Based on this model, we analysis the review titles bodies, each give them one classification score to represent it's positive or negative tone. The following chart shows some review scores of hair dryer, it shows that this model works good.

Review Headline	Review body	Body Score	Headline Score
Works great	Works great!	1	1
I love travel blow dryer ...	This dries my hair faster…	0.789	0.984
Five Stars	Love this dryer!	1	0.885
Gets extremely hot…	I have burned my hand ...	0.032	0.188

Model establishment

After trained a good nature language procession network, we establish the association model by the following steps.

First, we calculate the distribution of data, then based on the theory of test of significance for coefficient of correlation when it is not Gaussian distribution, which called Spearman test, to calculate the association of star rating and scores.

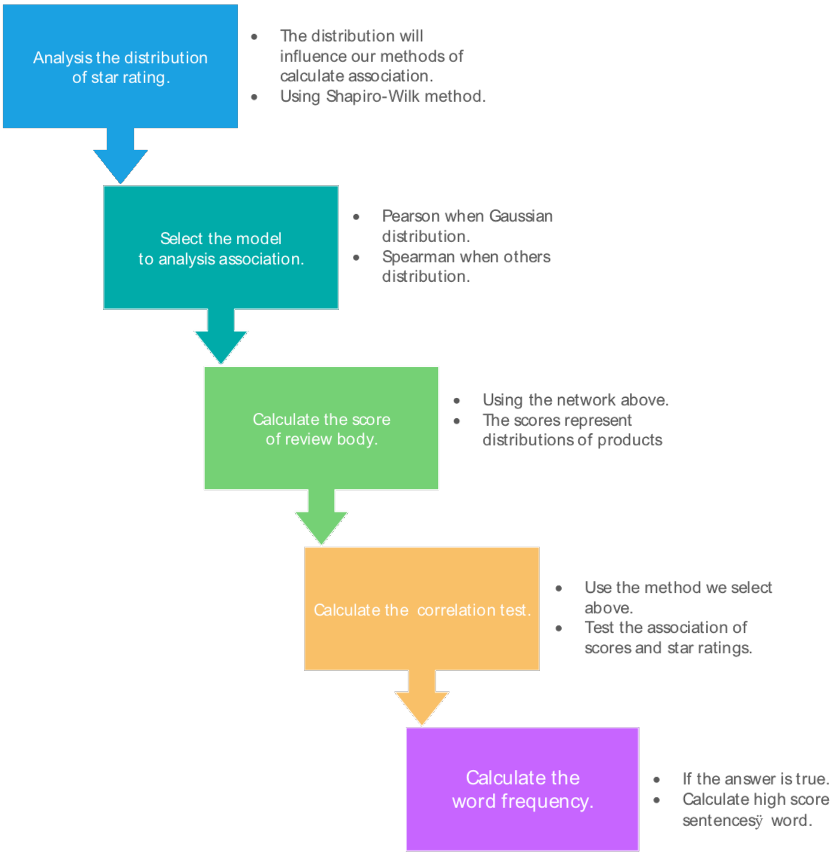

Analysis the distribution of star rating.
The distribution will influence our methods of calculate association.
Using Shapiro-Wilk method.
Select the model to analysis association.
Pearson when Gaussian distribution.
Spearman when others distribution.
Calculate the score of review body.
Using the network above.
The scores represent distributions of products
Calculate the correlation test.
Use the method we select above.
Test the association of scores and star ratings.
Calculate the word frequency.
If the answer is true.
Calculate high score sentencesÿ word.

Fig.9

First we explore the data of microwave(already washed), to select which method should we use to calculate distribution. In the following chart we explore the data's shape and using skewness and kurtosis to calculate whether the data obeys Gaussian distribution.

Table1.7

Descriptive				
			Statistic	**Std. Error**
Star Rating	Mean		3.74	0.067
	95% Confidence Interval for Mean	Lower Bound	3.6	0
		Upper Bound	3.87	0
	5% Trimmed Mean		3.82	0
	Median		4	0
	Variance		2.267	0
	Minimum		1	0
	Maximum		5	0
	Range		4	0
	Skewness		**-0.872**	**0.109**
	Kurtosis		**-0.759**	**0.218**

The skewness of the distribution is -0.872(standard error 0.109) and Kurtosis value is -0.159, using the zero-value formula:

$$Zero-score_{skewness}=\frac{Skewness}{Std-error}$$

$$Zero-score_{kurtosis}=\frac{Kurtosis}{Std-error}$$

When the zero-score is around ± 1.96, we can say the data obeys normal distribution. Nevertheless, in this case, the zero-score is -8 and -3.48, so these data does not obey normal distribution, so we are supposed to use Spearman test:

$$r_s=\rho_{\mathrm{rg}_X,\mathrm{rg}_Y}=\frac{\operatorname{cov}\left(\mathrm{rg}_X,\mathrm{rg}_Y\right)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}$$

Where:

•ρ denotestheusual Pearson correlation coefficient, but applied to the rankvariables,

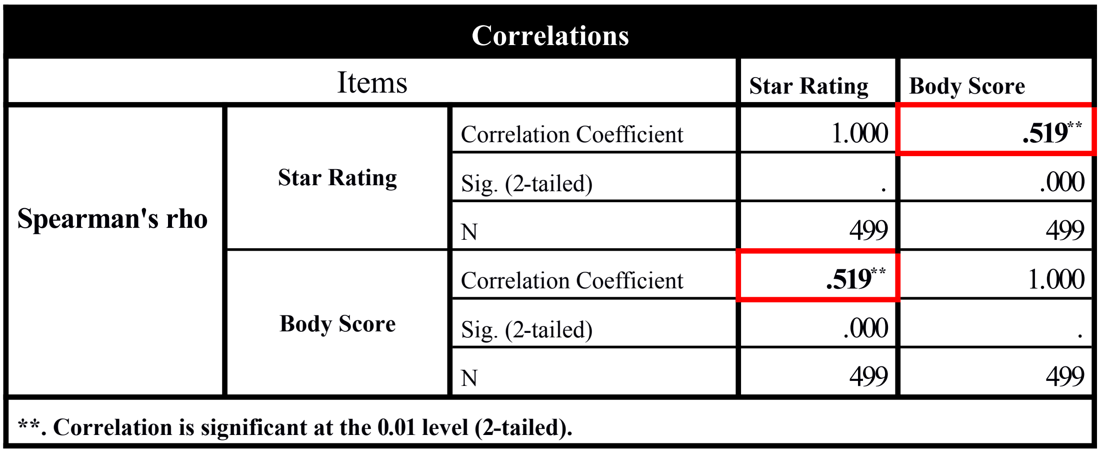
•$\operatorname{cov}(rg_X, rg_Y)$, $\operatorname{cov}(rg_X, rg_Y)$ is the covariance of the rank variables,

•$\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_X}$ and $\sigma_{\mathrm{rg}_Y}\sigma_{\mathrm{rg}_Y}$ are the standard deviations of the rank variables.

Result

After established the model above, we calculate the Spearman-Correlation-Coefficient in SPSS, and get the result below:

Table 1.8



Correlations

Items			Star Rating	Body Score
Spearman's rho	Star Rating	Correlation Coefficient	1.000	.519**
		Sig. (2-tailed)	.	.000
		N	499	499
	Body Score	Correlation Coefficient	.519**	1.000
		Sig. (2-tailed)	.000	.
		N	499	499

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

The chart shows that the Spearman-Correlation is significant at the 0.01 level (2-tailed), which means a strong association. So specific quality descriptors of text-based reviews strongly associated with rating levels.

After doing this example on microwaves' review, we calculate the scores of hair-dryer and pacifier, the Correlation-Coefficient of microwave, hair dryer and pacifier is 0.519, 0.494 and 0.379. Spearman-Correlations are all significant at the 0.01 level(2-tailed), So there is a strong association among these tree products' star rating and review text.

To find which specific word that influence the star rating, we do some word frequency statistics based on sentence score, and link them together with stars.

Here is our result of links between specific word and star ratings, we only monitor one-star and five-stars to monitor polarity, and select the description of quality:

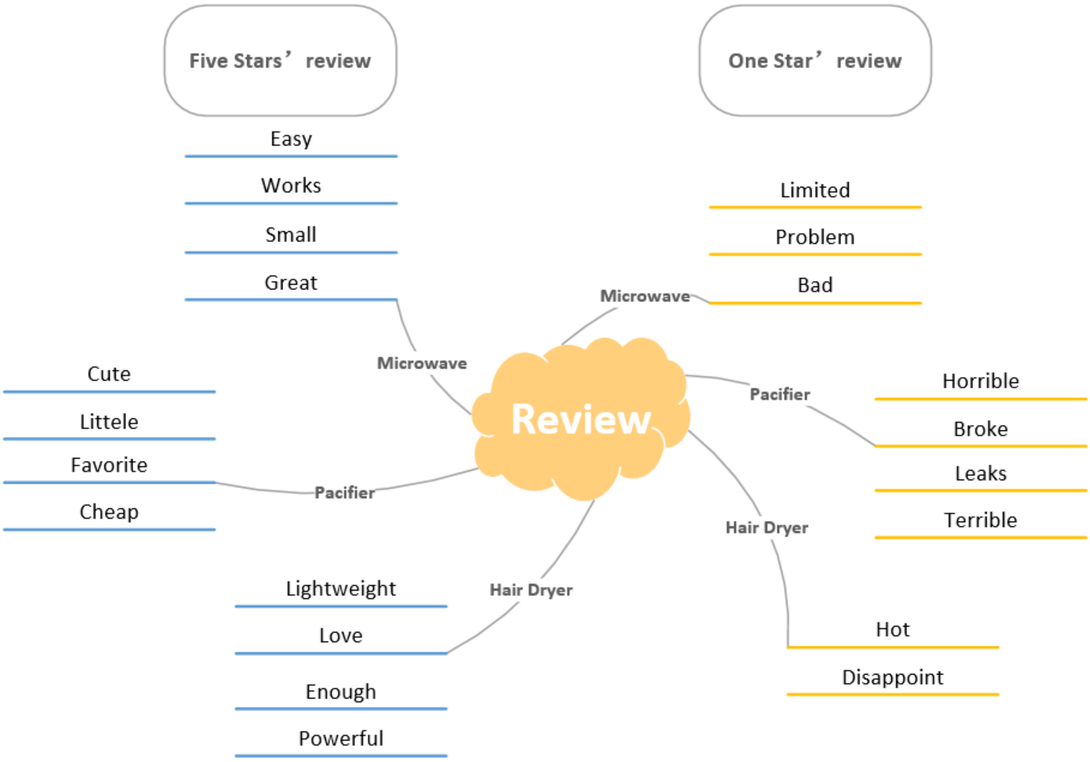


Fig.10

Analysis of the result

All the data pass the Spearman- correlation-test and the Correlation-Coefficient of microwave, hair dryer and pacifier is 0.519, 0.494 and 0.379. Correlations are all significant at the 0.01 level (2-tailed).

From the result above we are able to analysis the products' performance.

For microwave, it's easy to use and small size made customers love this product, but others fell disappointed at the limit of its function, probably due to the small size. For pacifier, five-stars reviews think the shape is small and cute, it becomes babies' favorite pacifier. Nevertheless, some reviews say it leaks and some times broke. When it comes to the hair dryer, customers love its light-weight and it is enough for them. But negative comments tend to say it is very hot and they are disappointed.

VI. Error Analysis and Sensitivity Analysis

5.1 Correlation analyze

The main purpose of the model in this paper is to study the correlation between the respective variables and the dependent variable, and use the correlation analysis to make a preliminary judgment. Correlation analysis is a commonly used statistical method to study the correlation between random variables. The correlation coefficient r is used to indicate the degree of correlation between the two variables, and the value of r is calculated using sample data, which ranges from -1 to 1. $|r| > 0$ indicates that there is a positive correlation between the two variables, otherwise it is a negative correlation; $r = 0$ indicates that there is no correlation between the variables; $|r| > 0.8$ indicates that there is a strong correlation between the variables, and $|r| < 0.3$ indicates that the variables are between The correlation is extremely weak and can be considered irrelevant.

Relevant analysis is made on the star rating, comment length, comment duration, and comment useful votes. The results are shown in Table 1.4.

Table 1.4

Items	Coefficient	Useful votes	Comment duration (days)	Star rating	Comment length
Useful votes	Spearman correlation	1	$.060^{**}$	$-.073^{**}$	$-.308^{**}$
	Two-sided test		.009	.002	.000
	N	1845	1845	1845	1845

**. Significant correlation at .01 level (both sides)

It can be seen from Table 1 that the correlation coefficients between the respective variables are very small, and the absolute values are all below 0.2. It can be considered that there is no correlation between the variables, and all of them can be included in the research model for analysis. In addition, it can be seen that the respective variables and the control variables have significant correlations with the dependent variables (both of which are significant at the two-tailed level). The correlation coefficient between comment duration and useful votes was 0.06, which was significantly positively correlated at the 0.01 level. Comment length was also significantly positively related to usefulness at the 0.01 level. This is a preliminary test of the previous hypothesis H1, and what is the specific impact relationship and its impact relationship under different brands, then continue to explore through the subsequent regression analysis, and then get support for each research hypothesis.

VII. Evaluation and Promotion of Model

7 Strength and Weakness

7.1Strength

While making full use of the data, redundant data is discarded, which simplifies the calculation of the model. Each quantity is digitized and a variety of appropriate algorithms are used to implement the model.

7.2 Weakness:

In practice, reviews often include pictures, and the influence of pictures on the value of reviews cannot be ignored, but this article is limited by the type of data provided, and ignores the influence of pictures on the value of reviews.

7.3 Promotion

The model uses the DESCAN algorithm to analyze the relationship between data and time. This algorithm is superior in time analysis and worthy of reference. At the same time, the model uses a convolutional neural network algorithm for natural language processing. This algorithm is useful for complex Problem processing has significant effects and can be applied to other natural language processing problems.

VIIIConclusions

1 Conclusions of the problem

- Track the useful reviews on online market website.
- Analyze relationship between text, star rating, and helpful votes.
- Explore the potential association between product fame and reviews.
- Identify better online sales strategy and significant product design feature based on reviews.

2 Methods used in our models

- DBCAN model
- Nature language processing model
- Word frequency statistics
- Spearman test
- Correlation coefficient matrix
- Time-line model
- Multiple regression

I X. References

[1]Kim, Yoon. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 10.3115/v1/D14-1181.

[2] Mikolov, Tomas & Sutskever, Ilya & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems. 26.

[3] Pennington, Jeffrey & Socher, Richard & Manning, Christoper. (2014). Glove: Global Vectors for Word Representation. EMNLP. 14. 1532-1543. 10.3115/v1/D14-1162.

[4]Jie lin,Pingchun Wang.Research on the factors influencing the usefulness of online comments in e-commerce trade[J].Commercial economic research,2017(10):73-75.

[5]Zhuwang Zu,Zhenqiu Yuan.The dimensions and research methods of online comment quality in e-commerce are discussed[J].University library and information journal, 2016,34(04):99-103+122.