| Problem Chosen | 2020 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2016002 |

# Research on consumer decision making and e-commerce marketing strategy under open online comment system


**Summary**

This paper analyzes the data sets of three commodities on Amazon e-commerce platform, and infers the consumer decision rules in Amazon e-commerce environment. It also explores the relationship among online reviews, stars, help reviews and other indicators, as well as the actual impact of these indicators on consumer purchase decisions.

At the beginning, we cleaned the three original data sets, and removed the missing items, single comments and less useful product data. Secondly, we use word2vec model to quantify the main content of the comment, and set an independent "loss" value to verify the accuracy of the model. Then, we set up a text similarity index to measure the similarity between the current comment and all previous comments. Then, through the establishment of negative binomial regression model, we analyze that vine membership, comment text similarity and comment subject information content will have a greater impact on the credibility of the comment. Finally, we establish a reputation evaluation system based on AHP, and discuss the changes of product reputation in time measurement using Markov chain.

Through the data processing and the analysis of the verified purchase index, we know the sales volume of the products. The negative binomial regression

equation is used to predict the text similarity, star level and sales volume. At the same time, the potential most successful and failed products of the three databases are B002KPM718, B00KPM, B00JH23DE, B003837V8A, B00O2KV5E4 andB00352M1RA.

In addition, we also use SentiWordNet in WordNet library to analyze the emotional characteristics of comment body, and objectively evaluate the emotional score of consumers for this kind of goods under the condition of retaining the original information to the maximum extent. And through a single product time node diagram, a single product time sequence diagram, analysis of the interaction between commodity stars and comment emotion. Then we get the conclusion that most of the stars have a positive correlation with the emotional scores of the reviews. That is to say, the promotion of star level will lead to the promotion of comment emotion; the decline of star level will lead to the decline of emotion score.

Finally, we make an overall evaluation of the model and put forward valuable suggestions for the online marketing activities and product design of sunshine company.

**Keywords**: Negative binomial regression; sentiment analysis; markoff chain;

# Contents

**1 Introduction** **1**

1.1 Problem Background . . . . . 1

1.2 Our Efforts . . . . . 2

1.3 Finally, we solved the following tasks . . . . . 3

**2 Statement of Our Model** **4**

**3 Data analysis** **4**

**4 Problem solving** **7**

4.1 Analysis of the question a . . . . . 7

4.1.1 Doc2Vec model obtains comment Text vector . . . . . 7

4.1.2 Text analysis method . . . . . 7

4.1.3 Model evaluation . . . . . 8

4.1.4 Model parameter adjustment . . . . . 9

4.1.5 Negative Binomial Regression Model . . . . . 10

4.2 Analysis of the problem b . . . . . 14

4.2.1 Determine reputation evaluation system . . . . . 15

4.2.2 Product reputation estimation of Markov chain . . . . . 16

4.3 Analysis of the question c . . . . . 17

4.3.1 Descriptive statistics of variables . . . . . 17

4.3.2 Correlation analysis of variables . . . . . 18

4.3.3 Analysis of regression model . . . . . 18

4.4 Analysis of the problem d . . . . . 19

4.5 Analysis of the problem e . . . . . 20

**5 Strengths and weaknesses** **22**

5.1 Strengths . . . . . 22

5.2 Weaknesses . . . . . 22

**6 The letter** **22**

**Appendices** **25**

**Appendix A First appendix** **25**

# 1 Introduction

## 1.1 Problem Background

"What are you going to do?"

"Pick up the express."

This kind of dialogue content appears more and more in people's life, which makes online shopping become a kind of life scene that people enjoy. Nowadays, the rapid development of Internet technology makes online trading platform occupy more and more broad market space. However, the online transaction behavior shows an exponential growth trend, but also because of the information asymmetry of both sides of the online transaction, it brings greater uncertainty and high risk to both sides of the transaction. As a result, the mechanism of consumer online comment came into being. The comment mechanism of online trading platform allows consumers to browse freely on the platform and comment on goods online. To some extent, the implementation of online reviews provides an effective way for buyers and sellers to obtain real commodity information, and also enables consumers to obtain more consumer surplus.

According to the 2019 consumer shopping behavior report released by Channel Advisor, 90% of respondents worldwide read online reviews, and 83% of respondents said that online reviews affected their shopping decisions. Consumers' preference for information sources in shopping obviously inclines to other users' online evaluation.

Take the online shopping mall created by Amazon in the title as an example. Amazon's online shopping mall sets a five-star product star index for customers, with five-star as the highest satisfaction. At the same time, Amazon's online mall has set up an open online comment system for customers. The open online comment system has subverted the traditional role of ordinary consumers in com-

modity trading and information dissemination. The massive online comment information released by a large number of consumer groups enables both businesses and consumers to obtain the feedback of the product's customer groups on the product through star rating, comments and other customers' help rating of the comment, and provides a reference decision for the potential consumer groups of the product.

Through the study of online reviews:

- It can provide effective decision-making basis for consumers' purchasing behavior;
- It can help the merchants of e-commerce platform to focus on effective reviews, so as to make a reasonable estimate of the expected sales of goods, which is conducive to enterprises to develop a more targeted marketing strategy.
- It is helpful to adjust and optimize the online comment mechanism itself, so as to help the buyer and the seller get higher utility.

## 1.2 Our Efforts

In order to explore the star rating of three kinds of commodities in Amazon online mall, help rating and the relationship between comments, we first analyzed the original data. After finding out the relationship among the three indicators of different kinds of goods, we set the screening rules for the original data, and cleaned the original data, excluding the impact of invalid data and malicious comments.

After sorting out the original data from Amazon online shopping mall into effective sample data, we transform the whole sample subject index into word vector. We try our best to make the transformed word vector contain the log-

ic information, vocabulary information, word order information and so on. In this way, not only can the transformed word vector fully reflect the original information structure, but also can reasonably explain the overall impact of the comments in the later regression.

At the same time, we consider the problem of information overload caused by massive information in the online comment system. Therefore, we use doc2vec text analysis technology to screen and analyze the usefulness of comments, and then study the characteristic factors that can reflect the usefulness of online comments.

## 1.3 Finally, we solved the following tasks

(1) The qualitative and quantitative variables are distinguished, and the explanatory variables and the interpreted variables suitable for negative binomial regression are selected from the data. Negative binomial regression was carried out to get star rating, score and help score the relationship among the three.

(2) According to the data after cleaning and processing, the sales of three kinds of commodities are visualized. The sales information based on rating and comment quantity is obtained, and the sales characteristics of three kinds of goods are summarized and analyzed.

(3) Use time series model to predict the future online reputation of three kinds of goods.

(4) Based on the user's comments and ratings on the products, the sales volume of the three products are predicted respectively, and the hot products in the future are selected.

(5) According to the emotional tendency of the comment text, we analyze the

specific influence of the specific star (such as: low star) on the consumer behavior (such as: the low star comment in the early stage of the consumer has a strong correlation with the low star comment in the later stage), and the correlation between the emotional color of the comment text and the star.

(6) Based on the above analysis, we summarize the online sales strategy of sunshine company and identify the design functions that can enhance the product performance.

# 2 Statement of Our Model

**Assumptions**

- There is a negative correlation between the usefulness of comments and the similarity of comments, that is, the higher the usefulness of comments, the lower the similarity of comments.
- We assume that the data in the dataset is arranged in time.
- We try to say Amazon's logistics takes three days and users try to write comments as soon as they receive the package.

# 3 Data analysis

Three datasets are provided in this paper, namely hairyer, microwave, and pacifier Among them, the data file of microwave is the smallest, with a total of 1615 comments. Secondly, there are 11470 pieces of hair guy data set, the largest of which is pacifier data set, with 18939 pieces of comment data. There are 15 indicators for each data.

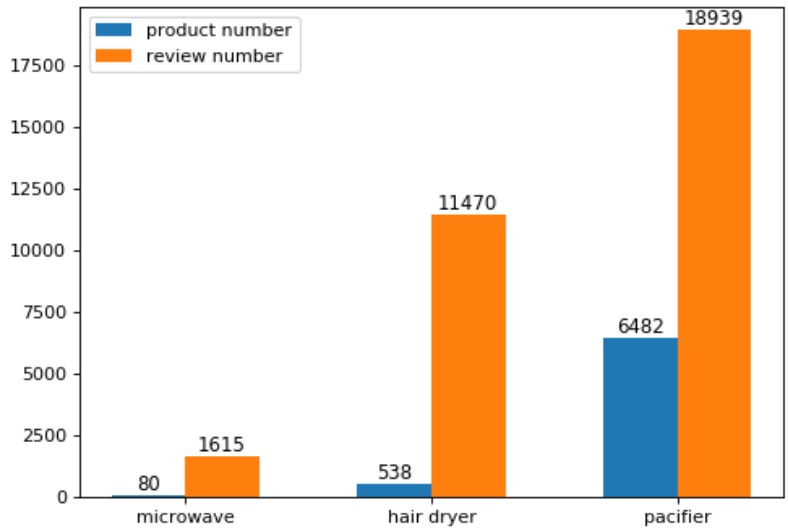


Figure 1: Number of products versus number of reviews

(1) First of all, we detect the missing value, and find that there are two missing value comments in the hair? Dryer and pacifier data sets. Because the impact is very small compared with the total number of samples, we directly take the elimination operation.

(2) First, the product review data with only one review is eliminated, because there is no comparability. There are 17 pieces of single product data in the microwave data set, 355 pieces of single product data in the hair player data set and 4935 pieces of single product data in the pacifier data set.

(3) To a certain extent, the helpful index can reflect the usability or value of the comment data, so it can further screen the comment data with low value. We will delete the data with helpful votes as zero, vine as zero, and verified "purchase" as zero, that is, there is no help to vote, and it is not a vine user, and there is no product review to buy this product. 129 pieces, 668 pieces and 1161 pieces of invalid data were deleted from the three data sets.

(4) In the further observation of the comment data, it is found that there are

some product comment data irrelevant to the subject product in the data. In order to solve this problem, we use "keyword filter" to further clean the data. For the data set of microwave, we search the key words of the product title index of the product. We use different combinations of upper and lower case to expand the search terms, "microwave", "microwave". The remaining two datasets also perform the above operations, and the final three datasets get 1470, 10428 and 9972 pieces of valid data respectively.

Table 1: Data set cleaning step table

| Remove | missing | single | low-credibility | irrelevant | Valid values |
|---|---|---|---|---|---|
| microwave | 0 | 17 | 129 | 0 | 1470 |
| hairdryer | 2 | 355 | 688 | 17 | 10428 |
| pacifier | 2 | 4935 | 1161 | 2869 | 9972 |

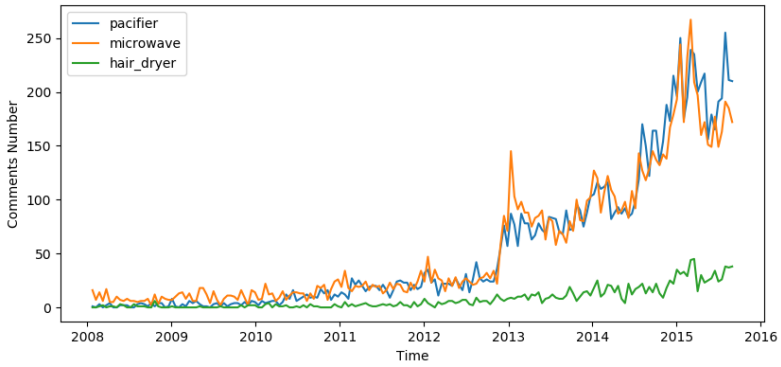


Figure 2: Comment time distribution

(5) We make a line chart of the data on a weekly basis. It is found that most of the data are concentrated around 15 years. The longer the data is, the less the reference value is. Therefore, in the following, we will not have a great impact on the overall relationship of the data if we intercept the data properly.

# 4 Problem solving

## 4.1 Analysis of the question a

### 4.1.1 Doc2Vec model obtains comment Text vector

Doc2vec is divided into two models, DBOW model and DM model, which have opposite structure. Both models can transform text into word vector, and retain a certain degree of content information in word vector.

### 4.1.2 Text analysis method

Because the comment content is text, we use cosine formula to calculate the text feature vector extracted from doc2vec model, and we can get the similarity of the two word vectors.

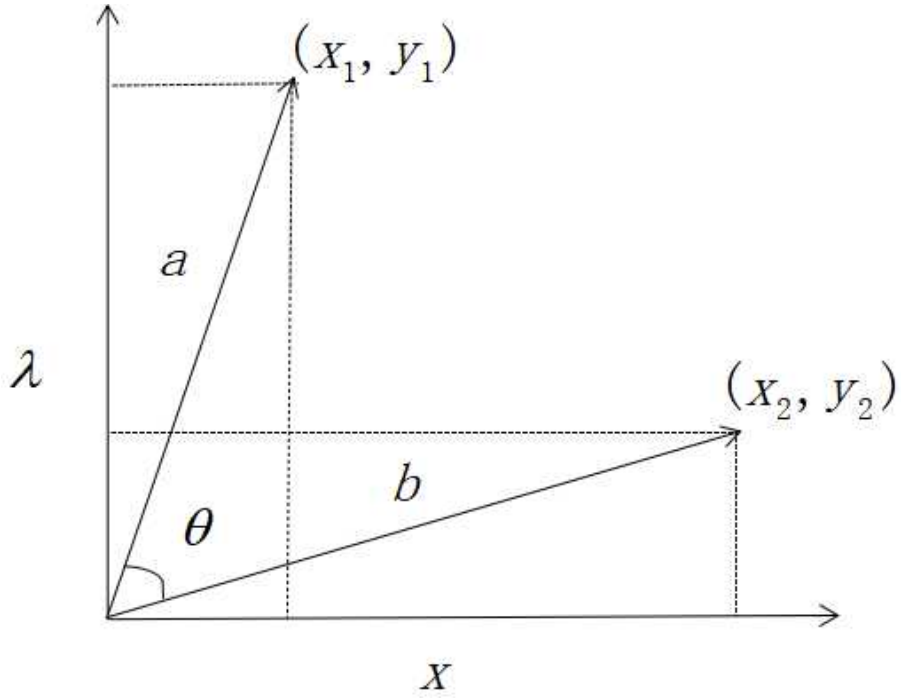


Figure 3: Cosine similarity

$$cos\theta = \frac{a*b}{\|a\|*\|b\|}$$
$$= \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}}$$

The closer the similarity is to 1, the closer the two vectors are to 0, the different they are. So cosine similarity can reflect the similarity of comment content to a certain extent.

#### 4.1.3 Model evaluation

We randomly divide the comment text of valid data into training set and test set according to the proportion of 7:3. If we are going to grade the model, the test set should have a label, so we use the variable "Star" as the label of the test set. And calculate the “loss” of the model through the following steps:

(1) We transform the test set text into word vector through the trained doc2vec model.

(2) We use doc2vec model to select the 20 word vectors closest to the current word vectors.

(3) Subtract the average star value corresponding to the 20 word vectors from the star value of the current word vector. This leads to our loss value. The smaller the loss value, the higher the accuracy of doc2vec model prediction.

$$starSum_i = \frac{k_i}{k_sum}$$

$k_i$ = The $i_th$ word vector is similar to the test word vector

$k_sum$ = The sum of all the similarities

#### 4.1.4 Model parameter adjustment

Because the three kinds of products are online reviews on Amazon e-commerce platform, the information contained in the review texts among the three kinds of products is quite similar. Therefore, the same model can be used for the three kinds of products, namely microwave oven, baby pacifier and hair dryer. We select the largest number of data set hairdryer to adjust the parameters of the model.

The initial parameters of the model are as follows:

| parameter | description |
|---|---|
| dm | Model training algorithm.dm=1,is DM.dm=0,is DBOW |
| size | Dimensions of vectors |
| window | The distance between the words used to predict context and prediction |

Figure 4: Model superparameter explanation

Table 2: Comparison of PVDBOW and PV-DM based on "loss"

| | dm = 0 | dm = 1 |
|---|---|---|
| window = 5 | 0.13376 | 0.10686 |
| window = 10 | 0.12539 | 0.03572 |
| window = 15 | 0.09847 | 0.06503 |
| window = 20 | 0.11150 | 0.02052 |
| window = 25 | 0.08997 | 0.10635 |

According to the broken line graph, the stability of the model is better when DM = 0, so we choose the model of DM = 0.

The final model parameters are:

Size=400, DM=1, window=15

According to this method, three model scores for three commodity data are respectively

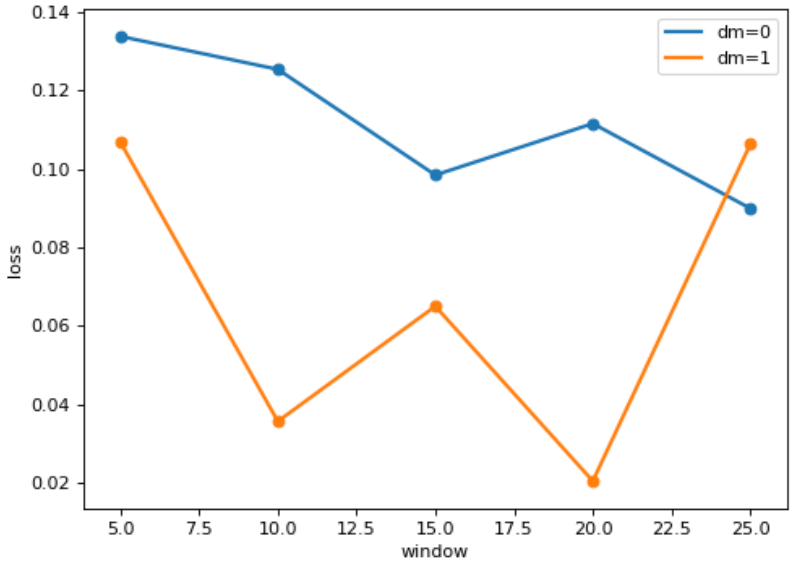


Figure 5: DBOW and DMin different window effect

Table 3: advantages and disadvantages of DM model in different dimensions

| | size=50 | size=100 | size=200 | size=400 |
|---|---|---|---|---|
| window=5 | 0.1076 | 0.1185 | 0.0587 | 0.0814 |
| window=10 | 0.030 | 0.060 | 0.0437 | 0.0351 |
| window=15 | 0.0627 | 0.0609 | 0.0322 | 0.0126 |

Mic=0.01264

Hair=0.019

Pac=0.009

The low value of loss indicates that doc2vec model is effective.

### 4.1.5 Negative Binomial Regression Model

**Model Selection**

Although we regard the comment as the explained variable, considering the

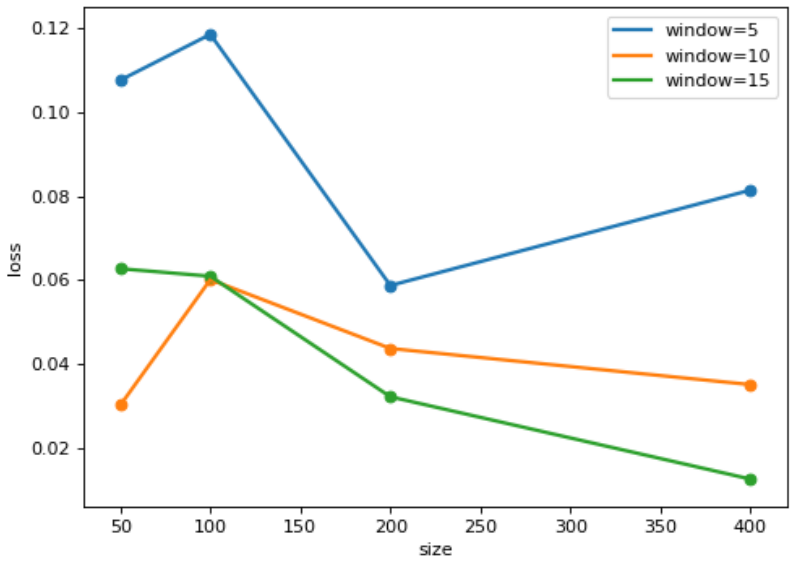


Figure 6: advantages and disadvantages of DM model in different dimensions

usefulness of the comment, the useful comment is a discrete variable with non normal distribution. That is to say, usability reviews do not meet the requirements of OLS model for data. Therefore, Poisson regression or negative binomial regression is more suitable for the regression of usefulness comments.

Moreover, the mean value and variance of the explained variables are 0.94 and 7.04 respectively. In this case, the negative binomial regression model is suitable for such scattered data. We choose to use negative binomial regression model to get better fitting results.

At the same time, negative binomial regression model can avoid Heteroscedasticity robust standard error method by using robust standard error method. It also avoids the endogenous problem caused by missing variables in the calculation.

**model and variable setting**

Among them, text similarity and star level are the core variables, and the rest are the control variables.

| variable type | variable name | variable | description |
|---|---|---|---|
| independent variable | text similarity | similarity | Average cosine similarity of each comment to all previous comment text |
| | star rating | star_rating | Star rating for each comment |
| | The length of the te | length | Length of comment text |
| | Comment on exposure length | days | The distance from 2015.8.31 |
| | vine | vine1 | vine='N' |
| | | vine2 | vine='Y' |
| | purchase | verPur1 | verified_purchase='N' |
| | | verPur2 | verified_purchase='Y' |
| dependent variable | Help to vote | helpful votes | Help vote for each comment |

Figure 7: Variable design

**result analysis of the model**

We first use the hair dryer model for regression, because of the high similarity of the review data, so the regression results have a good universality, which is also applicable to MIC and PAC data sets.

**variable correlation analysis**

It can be seen that the correlation coefficient between simila and helpful is -0.0837, and the correlation coefficient between star level and helpful is negative.

The correlation coefficient between star rating and similarity was -0.3861, with a high correlation rate. The correlation coefficient between the independent variables is less than 0.4, which is not enough to produce the problem of collinearity.

**analysis of negative binomial regression results**

Because the variables are standardized, the regression coefficient can directly reflect the importance of each independent variable to the dependent variable.

So we can see. Whether vine is a member or not, similarity of text, and length of paper have great influence on the credibility of comments.

(1) Text similarity

Table 4: Variable correlation result

| | simila y | starr g | days | length |
|---|---|---|---|---|
| helpful s | | | | |
| similarity | 1 | | | |
| starrating | -0.011 | 1 | | |
| days | -0.3861 | 0.0978 | 1 | |
| length | 0.0511 | -0.1015 | -0.2497 | 1 |
| helpfulvo s | 0.0837 | -0.0483 | -0.2169 | 0.2974 |
| 1 | | | | |

According to the negative binomial regression model, the coefficient of similarity is negative and P value is significant, the coefficient is -. 1.129. The results show that similarity has a negative effect on helpful, and the larger the value of similarity is, the lower the helpful will be. That is, the higher the similarity between the current comment text and the previous comment text, the lower the usefulness of the comment and the less the information value. The smaller the text similarity is, the more new features will be mentioned in the comments, and the more useful votes will be.

(2) Star rating

The coefficient of stars is negative and P value is significant, the coefficient is - 0.190. Negative correlation with helpful. The higher the star rating, the lower the usefulness vote. The higher the star rating, the more product reference information will be ignored in the comments, and the more objective

Table 5: Negative binomial regression

| | helpfulvotes (***p<0.01) |
|---|---|
| similarity | -1.129*** |
| | (-47.647) |
| starrating | -0.190*** |
| | (-4.570) |
| length | 0.834*** |
| | (12.793) |
| days | -0.691*** |
| | (-8.335) |
| verPur1 | 0.496*** |
| | (3.534) |
| Vin1 | 1.607*** |
| | (6.204) |
| cons | -1.694*** |
| | (-6.235) |

comments will be more adopted by people.

## 4.2 Analysis of the problem b

For this problem, first of all, we need to establish the evaluation indicators of the product reputation. After analyzing the data set, the most intuitive indicators of the product reputation are star rating and review. The higher the star rating, the more likely the sentiment of the product review is to be satisfied, but the information contained in the review is far more than star rating. Therefore, based on the user's evaluation of the product, we can build the product reputation Comprehensive evaluation index system.

In this question, the semantic analysis of user comments is carried out, the user comments are automatically classified, the comment set is constructed, the comments are classified by AHP, and the quantitative results are obtained according to the user comments.

### 4.2.1 Determine reputation evaluation system

**user comment classification**

In the product review interface, some users' comments have nothing to do with the use of the product, and some comments are even malicious or false. A good comment should be of a certain length and full of logic. Simple "good" and "bad" can not accurately reflect the characteristics of Chu products. On the other hand, the content of the comment should be objective, fair and authentic. At the same time, comments should not be too far away from now, because they are time sensitive, and the quality of comments will decline with time.

Some common words in text, but due to the lack of information, it has little or no impact on text classification. For example, "a", "am", "of the" and "the" filter the words in the comments in English to improve the accuracy of classification.

Next, we calculate the similarity of the comments, and adopt naive Bayes algorithm. Naive Bayes algorithm follows the traditional classification method, which is based on Bayes theorem and is the modification of Bayes algorithm. It is based on the assumption that the occurrence of words is independent in the frequency of occurrence. By calculating the probability that a given text belongs to a certain category, we can determine the category of text.

**establish commodity evaluation system**

By dealing with commodity reviews, different indicators in reviews are separated, including: quality, price, description consistency, customer service attitude, product integrity, logistics speed, and after-sales situation. In the process

of selling goods, the appeal index is a hierarchical structure, which can be used to express the reputation of goods in the form of mathematical functions.

After establishing the hierarchical relationship, the importance of each layer of information elements is expressed by numerical value, the judgment matrix is constructed, the weight wi between elements is calculated, the single factor evaluation moment of commodity performance, delivery and after-sales service is bi respectively, and the comprehensive evaluation aggregation of each layer is determined, $Y_i = w_i j \times B_i$, and the reputation evaluation value of final commodity $Z = w_i \times (Y_i)^t$

#### 4.2.2 Product reputation estimation of Markov chain

Through the above process, we can get the reputation value of each product in each day. By using Markov process, we can analyze the tendency of increasing or decreasing the reputation of this product in the future.

Markov chain is a kind of stochastic process with discrete time, discrete state and memory function. It is a mathematical model commonly used in prediction. If the state of the data itself at every moment only depends on the state of the random variables in front of it, and has nothing to do with the state in front of it, this is the "disorder" of Markov chain.

Understanding the reputation of this paper is an insensitive time series with no aftereffect. It can predict the reputation of each product according to the reputation data of product history.

First of all, it is a feature of frequency to find the transition probability of sequence fragments. When the number of fragments is large, the frequency can be regarded as changed equivalently, so it can be used to estimate the transition probability.

To test the "Markov chain" of discrete series of numbers:

In general, Markov chain of discrete sequence is used to test the "Markov" of the sequence with random variables. Statistics when large: After the data is read in and processed by python, the first-order Markov chain prediction is made for the data conforming to Markov chain property, and the result is output to excel table, then we can get whether the reputation of the product will rise or fall in the next time.

## 4.3 Analysis of the question c

First, we processed the data, added up the reviews of each product, and eliminated the unpurchased review data to get the approximate sales volume of the product.Meanwhile, the star average, the average of text similarity, the average of text length, the average of exposure time, and the average of voting are obtained.

### 4.3.1 Descriptive statistics of variables

Because hair has the largest number of samples and the highest data stability, and because the three data sets are all amazon product reviews with high similarity, the regression formula of hair sample is also applicable to the other two data sets

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| star_ratin~n | 174 | 4.151787 | .5788899 | 0 | 5 |
| sim_mean | 174 | .0010001 | .0018812 | -5.45e-06 | .0208191 |
| days_mean | 174 | 692.9445 | 513.7877 | 0 | 2976.25 |
| length_mean | 174 | 24.82263 | 11.40288 | 0 | 98.4 |
| helpful_mean | 174 | 1.907823 | 2.810338 | 0 | 21 |
| sale | 174 | 54.8908 | 83.16788 | 0 | 515 |

Figure 8: Negative binomial regression

### 4.3.2 Correlation analysis of variables

| | star_r~n | sim_mean | days_m~n | length~n | helpfu~n | sale | star_r~n |
|---|---|---|---|---|---|---|---|
| star_ratin~n | 1.0000 | | | | | | |
| sim_mean | -0.1164 | 1.0000 | | | | | |
| days_mean | -0.2334 | 0.5996 | 1.0000 | | | | |
| length_mean | 0.0061 | 0.2026 | 0.4941 | 1.0000 | | | |
| helpful_mean | -0.0735 | 0.2860 | 0.1285 | 0.3181 | 1.0000 | | |
| sale | 0.0407 | -0.0531 | -0.0838 | -0.0595 | -0.0807 | 1.0000 | |
| star_ratin~n | 1.0000 | -0.1164 | -0.2334 | 0.0061 | -0.0735 | 0.0407 | 1.0000 |

Figure 9: Variable correlation result

The correlation coefficient between Star and sale is 0.0407, while the correlation coefficient between sim-mean and sale is -0.1164, showing a negative correlation.Where the correlation coefficient between sim-mean and days$_meanisgreaterthan0.4, solidwi$

### 4.3.3 Analysis of regression model

| sale | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| star_ratin~n | .9462163 | .0706912 | 13.39 | 0.000 | .8076641 | 1.084768 |
| sim_mean | -19.3665 | 8.825071 | -2.19 | 0.028 | -36.66332 | -2.069676 |
| length_mean | .0171929 | .0109458 | 1.57 | 0.116 | -.0042603 | .0386462 |
| helpful_mean | .398949 | .29776 | 1.34 | 0.180 | -.1846498 | .9825479 |
| _cons | -1.606965 | .2951861 | -5.44 | 0.000 | -2.18552 | -1.028411 |
| sim_mean | 1 | (offset) | | | | |
| /lnalpha | .2693589 | .071983 | | | .1282747 | .410443 |
| alpha | 1.309125 | .0942348 | | | 1.136865 | 1.507485 |

Figure 10: Negative binomial regression

P<0.01 so reject the null hypothesis, significant.Among them, star and sim are significant and respectively equal to 0.9462163 and -19.3665. The constant term is -1.606965.The expression $sale = 0.946216 \times star + -19.3665 \times sim - 1.606965$.So if I plug in the data, I get.The potential most successful product in the Mic data set is B002KPM7L8.The product that failed the most was B002KPM7L8.The most successful in the Hair dataset is B00JH2C3DE. The biggest failure was B003837V8A. The most successful in the Pac data set is B00O2KV5E4.The biggest failure was B00352M1RA.

## 4.4 Analysis of the problem d

We tried to explore the effect of a given star on the emotion of the relevant comment.

We first used the sentiwordnet dictionary in the wordnet library to assign emotional values to the comments, in which positive words were positive, negative words were negative, and objective words were 0.Noun (goodness...)Verb (love...)Emotional assignment. And the final sum is the sentiment score of the single comment text.

Since the time span of our data is very large, and most of the data are concentrated in 2014 and 2015, we take January 2014 to August 2015 as the data timeline. In addition, we selected the data of hair dryer (productid B003V264WW) with the largest number of product comments in the three data sets for processing. Because the three data sets are commodity reviews on amazon platform, there is a high similarity. Meanwhile, hair dryer with productid of B003V264WW is the commodity with the most commodity reviews in the three data sets, so it has a strong representativeness and reliability.

We divide the reviewbody of B003V264WWW data into three emotional comments according to the emotional score: positive, objective and negative.Combine the data of every half month as a node.Make a line chart of the number of comments and a stacked bar chart of the star series. As can be seen from figure 1, more four-star stars can drive more sales and attention, and increase the number of positive reviews.

We output the data in chronological order, excluding the influence of time sales volume on the number of comments and stars.The data were grouped into groups of 10 and assigned to their average values, flattening the data curve.And selected 200 data to chart.It can be found that in the short term, most of the time, the emotion curve of stars and comments is the first to close, and the increase

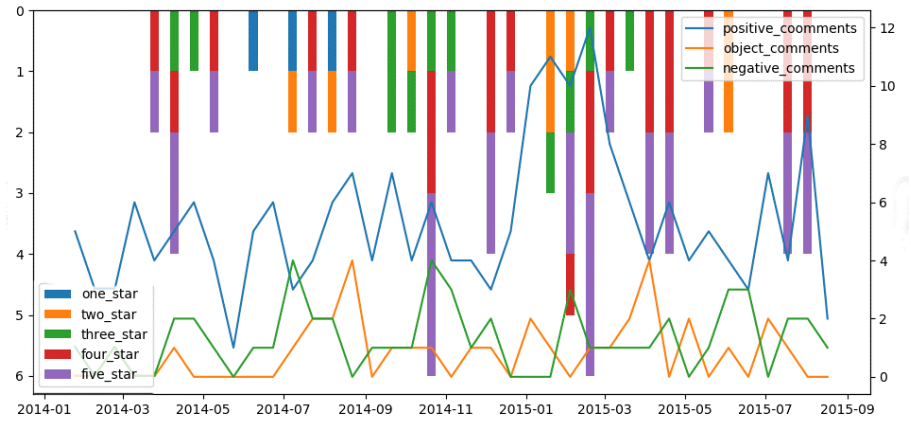


Figure 11: Contrast charts of stars and comments on emotions

of stars will lead to the improvement of comment emotion, that is, the number of positive comments will increase.The decline of stars will also drive down the enthusiasm of comments to a certain extent.

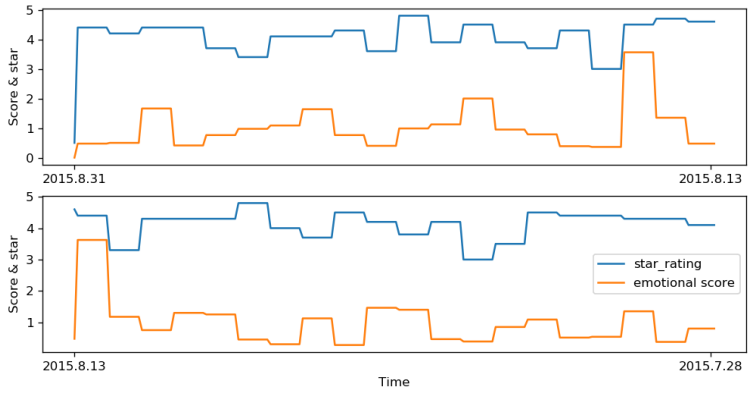


Figure 12: Go to time to affect star level and comment on emotion contrast chart

## 4.5 Analysis of the problem e

In our common sense, we think that the rating of a commodity review should be related to the content of the review.For example, a product with a high star rating will get a higher rating in the review.We want to further explore the relationship between commodity reviews and commodity ratings.

We visualized the first 50 data of hair$_dryerdataonthebasisofdataprocessingofd$.

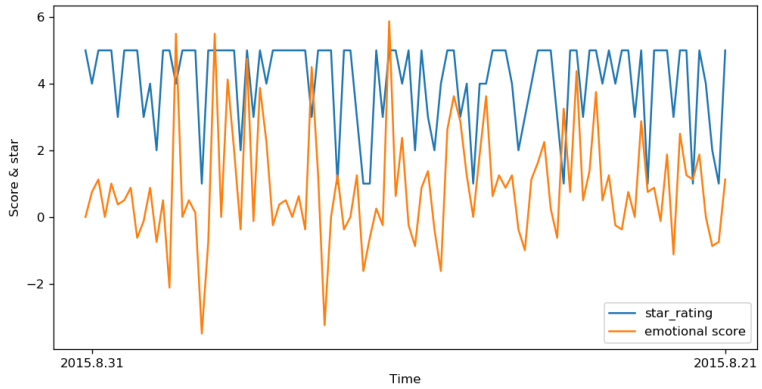


Figure 13: Comparison charts of small sample stars and comments on emotions

It can be seen that there are obvious trend differences in the middle part, and we need to further analyze the relationship between the two.

correlation test

(1) we first standardized the two data to remove the influence of dimension.

(2) because the sample size is large, we conducted JB test on the data. P <0.05 rejected the null hypothesis, and the data does not follow normal distribution.

(3) Spearman correlation coefficient was used

Number of obs = 10428
Spearman's rho = 0.2003
Test of Ho:emo and star are independent

$$Prob > |t| = 0.0000$$

P<0.05, the correlation coefficient is significant. The correlation coefficient was 0.2003.

It indicates that the correlation is weak and the two are not closely related.

The possible reason is that people express positive words in a variety of ways, which makes it difficult to detect all of them.In the data, there is data noise with 5 stars bad evaluation and 1 star good evaluation.

# 5 Strengths and weaknesses

## 5.1 Strengths

- We use word2vec to vectonize comments to maximize the efficiency of the use of comment text information.
- We use the negative binomial regression model, which has better regression effect on scattered data, so our model regression result is better.
- The emotional analysis of nouns, verbs and adjectives in the comments ensures the reliability of the emotional score.

## 5.2 Weaknesses

- Our model of correlation between stars and comments is relatively simple and cannot well represent the deep relationship between the two
- We are shallow in mining the time attributes of the data.Not a good reflection of the temporal characteristics of the data.

# 6 The letter

Dear marketing director of sunshine company: Hello, we are a MCM team.

Our team conducted in-depth research and Analysis on the review data of hair dryer, microwave oven and baby pacifier in the specific sales period of A-

mazon online mall. We get the text similarity to measure the similarity between the current comment subject and the overall comment, and analyze the emotion of the comment to get the emotional scores of different types of comments. On this basis, we analyze the impact of consumer behavior and online comment system. Based on our research, we put forward the following suggestions for the sales strategy of sunshine company in selling the above three products online:

(1) In terms of comment usability: according to the analysis of negative binomial regression model, the comments written by vine members, the similarity with the text and the length of comment text have a great impact on comment usability. To improve the use value of comments for consumers, that is, to improve the usability of comments, can promote the number of comments of vine members on related products, encourage consumers to write relatively long comments, and encourage consumers to make personalized comments on products.

(2) Potential successful goods and unsuccessful goods: according to the negative binomial regression model $Sale = 0.946216 \times star - rating - 19.3665 \times sim - 1.606965$ This formula can predict the sales volume of goods according to the star trend and text similarity of products, and then the potential successful or failed goods can be obtained. Help your company to produce successful products, so as to retain and attract more customers.

(3) Impact of commodity review rating on sales volume: According to the analysis of data visualization results, four-star and five-star reviews can improve product sales relatively more. There is a certain correlation between product star rating and comment emotion score. Product star rating can be achieved by improving comment emotion score.

From:

MCM

# References

[1] Wu Weifang. (2018). Research on the influence of text similarity based on Doc2Vec on the usefulness of online comments(Doctoral dissertation).

[2] Wang Mohan. (2015). Research on the impact of open online reviews on consumers' purchasing decisions. (Doctoral dissertation)

[3] Product evaluation and prediction based on markov chain.https://www.cnblogs.com/TTyb/p/5692330.html

# Appendices

## Appendix A First appendix

Here are simulation programmes we used in our model as follow.

**Input matlab source:**

new roman.fd New Roman.fd

```
import re

def etl(content):
    content = re.sub("[\s+\.\!\/_,$%^*(+\"\']+|[+——！，。？、~@#￥%……&*（）]+",
                     " ", content)
    content = content.replace('-', ' ')
    content = content.replace("'", '')
    content = re.sub(r" {2,}", " ", content)
    return content



review_temp = []
i = 0
for i, rev in enumerate(pac['review_body']):

    if type(rev)==float:
        print(i, type(rev))
    rev = str(rev).lower()

    rev = etl(rev)
    review_temp.append(rev)

temp_list = []
review_temp2 = []
```

```
for i in review_temp:
    temp_list = i.split()
    review_temp2.append(temp_list)

from nltk.corpus import stopwords

review = []
sr = stopwords.words('english')
for i in review_temp2:
    temp = []
    for j in i:
        if j not in sr:
            temp.append(j)
    review.append(temp)
print(review[:2])

from sklearn.model_selection import train_test_split
train, test = train_test_split(review, test_size=0.3)
print("train: " + str(len(train)))

documents = [TaggedDocument(doc, [i]) for i, doc in enumerate(train)]
model = Doc2Vec(documents, dm=1, vector_size=400, window=15, min_count=1, workers=4)
train_model = model.train(documents, total_examples=model.corpus_count,epochs=10)

import numpy as np

def bit_product_sum(x, y):
    return sum([item[0] * item[1] for item in zip(x, y)])

def cosine_similarity(x, y, norm=False):
    assert len(x) == len(y), "len(x) != len(y)"
    zero_list = [0] * len(x)
    res = np.array([[x[i] * y[i], x[i] * x[i], y[i] * y[i]] for i in range(len(x))])
    cos = sum(res[:, 0]) / (np.sqrt(sum(res[:, 1])) * np.sqrt(sum(res[:, 2])))
```

```
    return 0.5 * cos + 0.5 if norm else cos
m = 0
n = 0
sim = []
for i in range(len(review)):
    inferred_vector = model.infer_vector(doc_words=review[i], alpha=0.025, steps=300)
    m = 0
    if i<20:
        for j in range(0, i):
            temp_vector = model.infer_vector(doc_words=review[j], alpha=0.025, steps=
            m += cosine_similarity(inferred_vector, temp_vector)
        if i!=0:
            m = m/i
        sim.append(m)
    if i>=20:
        for j in range(i-20, i):
            temp_vector = model.infer_vector(doc_words=review[j], alpha=0.025, steps=
            m += cosine_similarity(inferred_vector, temp_vector)
        if i!=0:
            m = m/i
        sim.append(m)

    n += 1
    if(n % 50 == 0):
        print(n)
```