Problem Chosen	2020	Team Control Number
C	**MCM/ICM**	**2017918**
	Summary Sheet	

2020 ICM Summary Sheet

Abstract:

Firstly, we exploited the word2phrase algorithm proposed by the previous reseach is used for text information to find phrases, and then pre-processing such as segmentation, deletion of stopwords, and stemming are performed. Then we chose TF-IDF and GloVe word embeddings as the text representation. Then, based on text representation and comment length, comment time, helpfulness ratio and other characteristics, KMeans clustering was performed to obtain two clusters, helpful and unhelpful, and the Euclidean distance between the cluster centers and sample feature vector is used to model helpfulness rating. Finally, we mark reviews as six categories of Hair Dryer (negative), Hair Dryer (positive), Microwave (negative), Microwave (positive), Pacifier (negative), Pacifier (positive), and use a supervised topic model LLDA to get topic words for those topics

Key words: Clustering ;machine learning

Dear sir or madam,

I am very honored to be hired by your company as your company's product market research data analysis consultant. After four days of conscientious work, we can confidently feedback to you the results of our work and propose wise advice on product sales monitoring, sales strategies, and product feature designs for the three products that your company will launch on Amazon: microwave ovens, baby pacifiers, and hairdryers.

First, we obtained about 3w reviews of the three categories of products on the Amazon e-commerce platform: microwave ovens, pacifiers, and hairdryers, including the review date, helpfulness ratio, review text, and star rating. Then the data is cleaned, and a large number of comments that are judged to be unreliable are deleted based on the number of valid votes. Then we perform preprocessing such as word segmentation and part-of-speech tagging on the review text, and represent the text as the classic TF-IDF value and GloVe word embedding in NLP for following text feature analysis.

After reading a lot of papers, and using autoregressive models to model the impact of previous reviews on subsequent reviews, and using a random forest algorithm to solve them, we conclude that the usefulness ratio of review validity measures on the current Amazon platform is unreasonable And there is the Matthew effect, that is, five-star evaluation will inspire more five-stars. So we can't directly trust comments with high helpfulness ratio.

In order to find out the practical helpfulness rating to help extract the suggestions in reliable reviews, we refer to the prediction methods of helpfulness ratings in other papers, and select the fields such as whether the review is officially certified, the review time, and the maximum TF-IDF value of the review as the feature pairs to analyze the samples using two-cluster analysis. Finally, we successfully divided the sample into two categories, helpful and unhelpful, and established a more reasonable prediction method for helpfulness rating. After ANAVO analysis, we obtained that the measure that has the most impact on helpfulness rating is whether the review has been officially certified and the review time (accounting for the most Important top 15%), followed by the length of the review (most important top 30%), which means that officially certified, up-to-date, longer reviews are most likely to be useful. In addition, we modeled a more valuable metric based on helpfulness rating and star rating, that is, reputation value.

We recommend that in the future analysis of product reviews, preferentially select the officially certified, latest, and longer reviews as the analysis sample. It is recommended to use our helpfulness rating prediction method and comprehensive star measurement to more accurately monitor the performance of the product in the market.

Moreover, we analyzed the three major categories of products: microwave oven, pacifier, and hairdryer. We first establish an autoregressive model and an autoregressive moving average model to analyze the changes in the reputation of various products in the online market over time. It is concluded that the reputation of hair dryers and microwave ovens in the online market has steadily changed over time, and the reputation of pacifiers in the online market has increased over time.

The online market for hair dryers and microwave ovens is relatively mature, and baby pacifier products are still in the development stage. Therefore, we recommend Sunshine to focus on baby pacifiers for sales and promotion.

Finally, we divided the three categories of products into six categories according to positive and negative evaluations, and then used the LLDA model to extract the topic words of these six topics. Based on the extracted keywords, we propose the following sales strategies and design features:

1. In the microwave oven market, Samsung and Whirlpool products have a poor reputation, and Sharp products have a mixed reputation.

2. For microwave oven products, customers are most concerned about the durability of the product and after-sales maintenance services, and whether the microwave oven can make full use of the cabinet space. The firepower and humanized button design of the microwave oven are also important requirements for customers.

3. In the hair dryer market, the main competitor is Conair, and Conair's products also have a mixed reputation.

4. For hair dryers, customers' biggest concern is safety, there have been sparks and even dangerous situations of fire. Lightweight, suitable for curlers and straight hair with a variety of hair dryers, with a diffuser, is more popular with customers.

5. In the baby pacifier market, the current reputation of mattress and monitor is relatively

poor which indicates that there is greater room for improvement, while crib and swing already have mature products.

6. For pacifier products, customers often want them to be easy to clean, soft and imitating nipple.

These are our recent work and achievements. Sincerely hope that our suggestions can be helpful to your company, and our team has benefited a lot from this task. Thanks again for your invitation!

Yours faifully,

Team 2017918

Contents

1 Introduction

1.1 Restatement of the Problem

Data show that there are approximately 12 million to 24 million e-commerce sites worldwide (e-commerce is a business transaction that conducts electronic transactions online). It is estimated that 95% of all purchases will be made through e-commerce, and 93.5% of global Internet users have purchased products online.

Amazon is the largest online e-commerce company in the United States. In 2017, Amazon accounted for 44% of total US e-commerce sales, and 59% of millennials would go to Amazon first when shopping online. In the United States, two-fifths of consumers spend (41%) receive 1-2 packages a week from Amazon. For consumers aged 18-25, this number jumps to 50%, and for consumers aged 26-35, this number jumps to 57%.

In Amazon's online marketplace, Amazon offers customers the opportunity to rate (rating) and evaluate purchases. Amazon's consumer rating system consists of two parts: 1. Personal rating-star rating; 2.

Rating of reviews-helpful rating. The evaluation system consists of two major blocks: 1. Order evaluation; 2. Product evaluation (As long as any buyer account has been purchased, you can write product reviews for almost any product on the platform, without necessarily buying this product). Order evaluation is for an order, and the evaluation content can include customer service, logistics, the product itself, and so on. Product evaluation can only be directed to the product itself, and has nothing to do with factors other than other products such as customer service and logistics. At the same time, for order evaluations that do not meet Amazon's requirements, product evaluation is not an evaluation of the product itself, but involves aspects that are not related to the product itself, and the seller can apply to Amazon for removal. If buyers and sellers do not intervene, Amazon will not actively remove order reviews, and the system itself will evaluate product reviews. If the system violates the rules, Amazon will delete the product reviews by itself. Order evaluation will affect the ODR indicator of the seller's account, and product evaluation will not affect the ODR indicator.

Amazon review is a very important task in the seller's operation. At the same time, the company uses this data to understand its market, time, and selection of product design features. Sunshine plans to launch and sell three new products in the online market: microwave ovens, baby pacifiers and hair dryers. Sunshine's data center now provides us with review data for these three products. We will use this data to determine key metrics related to other competing products, using special combinations and data types, especially the time-based metrics in the data above. And models to capture online sales strategies and identify important design features to increase product appeal.

1.2 Data Analysis

In order to analyze the relationship between star ratings, reviews, and help levels easily, we first define HelpfulnessRaito = HelpfulVotes / TotalVotes in reason, and temporarily use HelpnessnessRaito to characterize the help level. In addition, we temporarily try to use the text length of reviews and the number of reviews for the same product to quantitatively describe reviews.

1.Relationship between star rating and help level

In the online marketplace, all forms of customer reviews of any product are considered essential [13]. Because in the online market, consumers often need to understand the true quality of the product based on the opinions of previous buyers and their own experience, and may vary from person to person, these reviews are often the main factor for consumers to buy the product [16]. Studies have shown that by providing effective review information, sales and revenue of online stores and e-commerce sites can be increased [17]. Therefore, we have drawn a box diagram (Figure 1) to try to describe the relationship between the product star rating and the help level of the corresponding review: the center of the low star (such as 1-2 stars) is more inclined to the lower The help level is more widely spread, and the center of a high star (such as 4-5 stars) is more inclined to a higher help level and the spread is smaller.

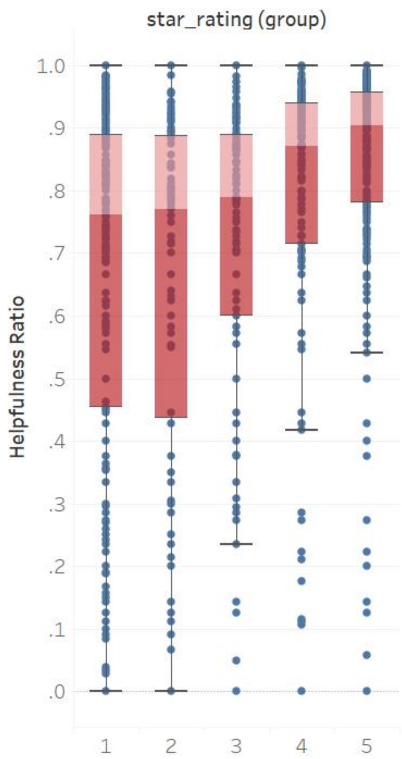


Figure 1

In addition, taking the product hair dryer as an example, we calculate that the correlation coefficient between the star rating and the help level of the product is 0.1127, and the covariance is 0.1098, showing a slight correlation.

1. The relationship between help levels and comments

By analyzing Amazon's online sales product review mechanism, we can speculate that the help level is directly related to the text content of the review, the text length of the review, and the time of the review. Here we only use the length of the comment as the character of the comment, and analyze the relationship between the help level and the comment. So we draw a line graph of help level and comment length (Figure 2), and calculate that the correlation coefficient between help level and comment length is 0.1184, and the covariance is 67.316356806672740

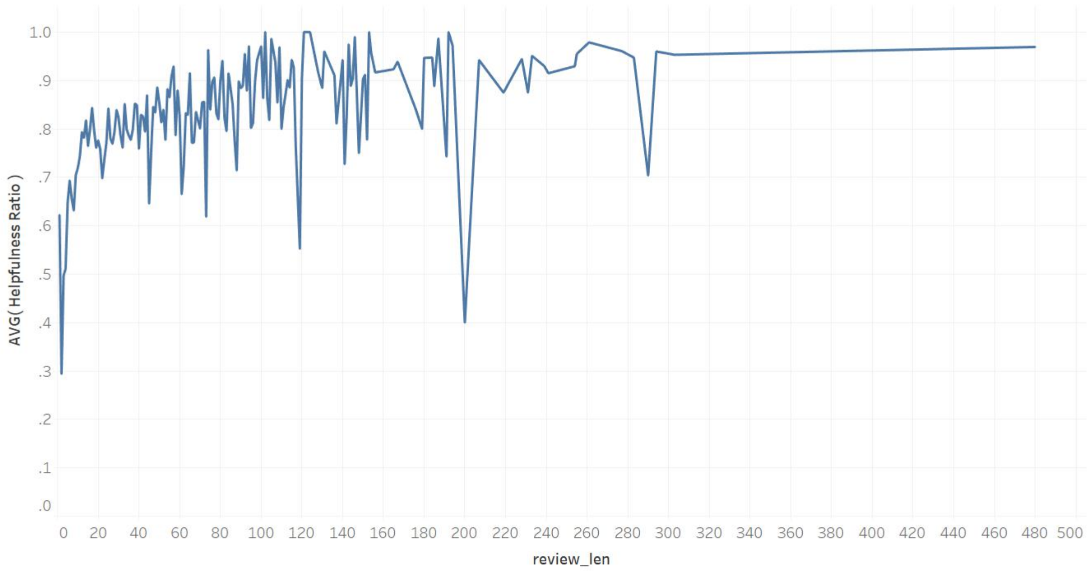


Figure 2

In order to further analyze the relationship between the help level and the comment length, we try to fit a three-dimensional polynomial on these sample points $f(x) = p_4x^3 + p_3x^2 + p_2x + p_1$, model parameters and simulation results are shown in Table 1 and Figure 3.

	Poly1	Poly2	Poly3
p_1	0.0007177	-4.516e-06	3.307e-08
p_2	0.7845	0.001967	-1.834e-05
p_3	0	0.7294	0.003461
p_4	0	0	0.6933
SSE	0.885	0.7553	0.7534
R-square	0.5735	0.636	0.6369
Adjusted R-square	0.5708	0.6315	0.6302
RMSE	0.07369	0.06828	0.06841

Table 1

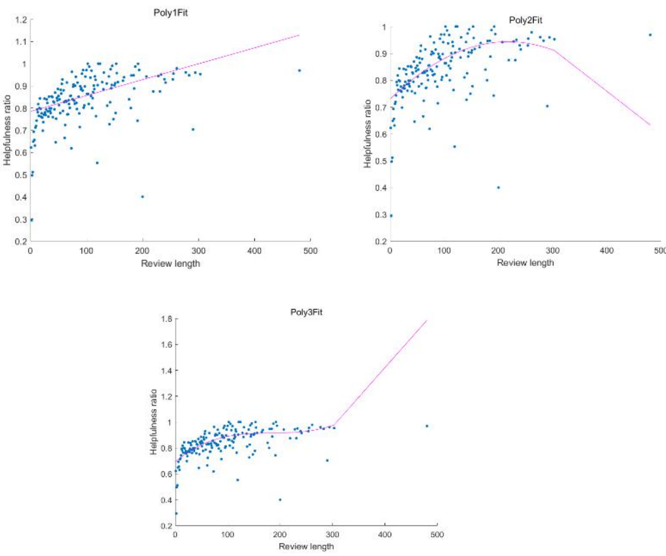


Figure 3

In addition, we also analyzed and interpreted the patterns between help levels and comments in Section 2a.

2 Assumptions and Notations

2.1 Notations

3 Analysis

3.1problem analysis

In order to analyze the relationship between star ratings, reviews, and help levels easily, we first define HelpfulnessRaito = HelpfulVotes / TotalVotes in reason, and temporarily use HelpnessnessRaito to characterize the help level. In addition, we temporarily try to use the text length of reviews and the number of reviews for the same product to quantitatively describe reviews.

1.Relationship between star rating and help level

In the online marketplace, all forms of customer reviews of any product are considered essential [13]. Because in the online market, consumers often need to understand the true quality of the product based on the opinions of previous buyers and their own experience, and may vary from person to person, these reviews are often the main factor for consumers to buy the product [16]. Studies have shown that by providing effective review information, sales and revenue of online stores and e-commerce sites can be increased [17]. Therefore, we have drawn a box diagram (Figure 1) to try to describe the relationship between the product star rating and the help level of the corresponding review: the center of the low star (such as 1-2 stars) is more inclined to the lower The help level is more widely spread, and the center of a high star (such as 4-5 stars) is more inclined to a higher help level and the spread is smaller.

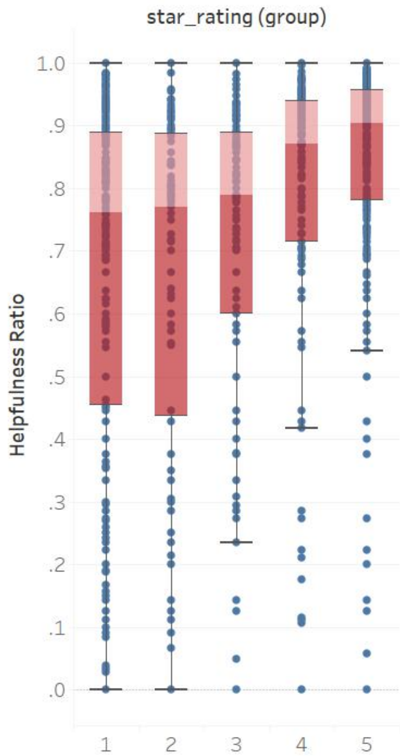


Figure 1

In addition, taking the product hair dryer as an example, we calculate that the correlation coefficient between the star rating and the help level of the product is 0.1127, and the covariance is 0.1098, showing a slight correlation.

1. The relationship between help levels and comments

By analyzing Amazon's online sales product review mechanism, we can speculate that the help level is directly related to the text content of the review, the text length of the review, and the time of the review. Here we only use the length of the comment as the character of the comment, and analyze the relationship between the help level and the comment. So we draw a line graph of help level and comment length (Figure 2), and calculate that the correlation coefficient between help level and comment length is 0.1184, and the covariance is 67.316356806672740

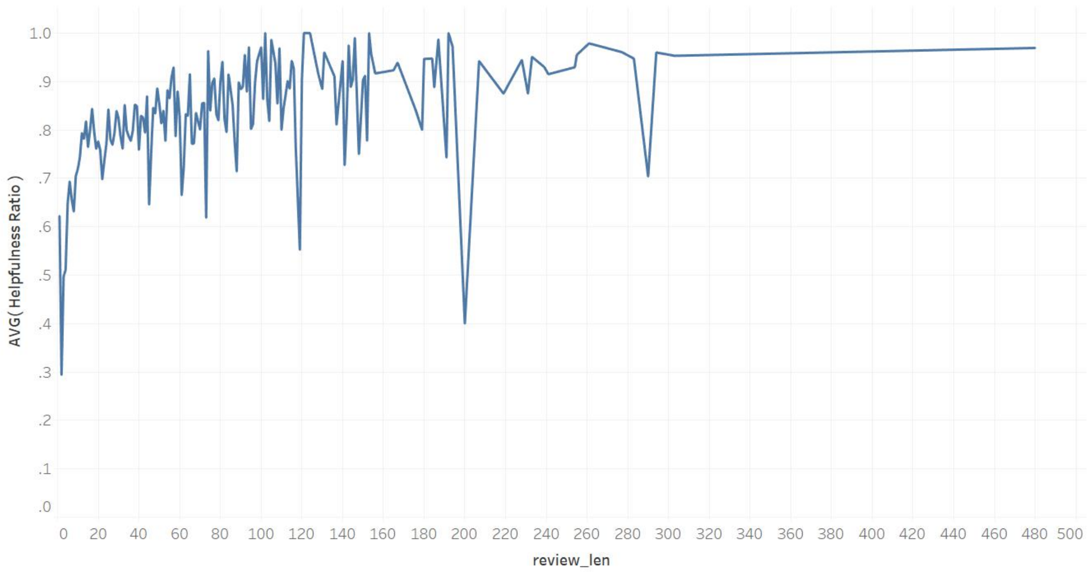


Figure 2

In order to further analyze the relationship between the help level and the comment length, we try to fit a three-dimensional polynomial on these sample points $f(x) = p_4x^3 + p_3x^2 + p_2x + p_1$, model parameters and simulation results are shown in Table 1 and Figure 3.

	Poly1	Poly2	Poly3
p_1	0.0007177	-4.516e-06	3.307e-08
p_2	0.7845	0.001967	-1.834e-05
p_3	0	0.7294	0.003461
p_4	0	0	0.6933
SSE	0.885	0.7553	0.7534
R-square	0.5735	0.636	0.6369
Adjusted R-square	0.5708	0.6315	0.6302
RMSE	0.07369	0.06828	0.06841

Table 1

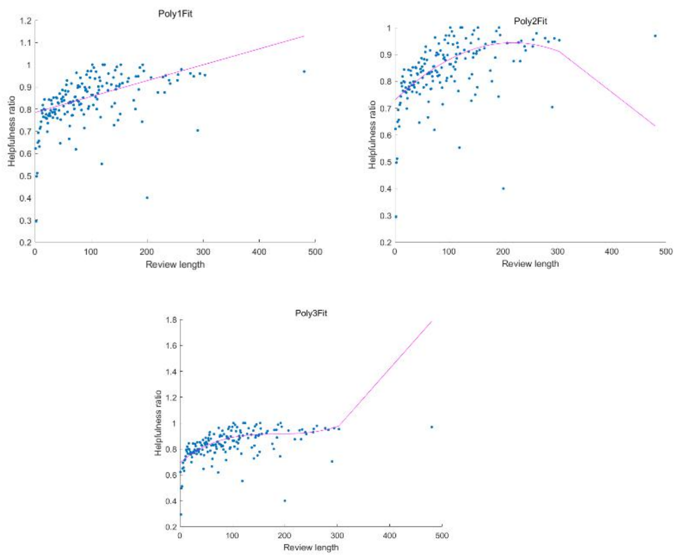


Figure 3

In addition, we also analyzed and interpreted the patterns between help levels and comments in Section 2a.

Labeled LDA for Extraction of Reviews' Topics

LDA [6] is a classic unsupervised Topic Model. It is a typical bag-of-words model, that is, a document is considered to be composed of a group of words, and there is no sequential relationship between words. At the same time, it assumes that each text is a mixture of multiple topics, and assumes that each word is generated by a topic. The probability that the word w appears in the text d is defined as:

\p(w|d)=p(w|\text{topic})\cdot p(\text{topic}|d)

After training, LDA can obtain the topic probability distribution of each document in the corpus, and then by analyzing some texts to extract their topics, we can perform topic clustering or text classification based on the topic.

LLDA [5] is improved on the basis of LDA to apply to labeled corpora(Figure x. shown the graphic model of LLDA). The difference between the two is that LDA performs Gibbs Sampling on all topics during training, but LLDA will only sample the topics to which the text belongs based on the topic labels. Other than that, the rest of LLDA's algorithms are the same as LDA.

The simple algorithm process of LLDA is as follows:

1. Initially assign random values to the topic distribution of each text and the topic of each word

2. Rescan the corpus. For each word, use Gibbs sampling formula to update its topic, and update the word number in the corpus. Gibbs sampling formula is defined as

$$P(z_i = j|\mathbf{z}_{-i}) \propto \frac{n_{-i,j}^{w_i} + \eta_{w_i}}{n_{-i,j}^{(\cdot)} + \boldsymbol{\eta}^T\mathbf{1}} \times \frac{n_{-i,j}^{(d)} + \alpha_j}{n_{-i,\cdot}^{(d)} + \boldsymbol{\alpha}^T\mathbf{1}}$$

Where n_{-I,j}^{w_i} is the count of word w_i in topic j, not including the current assignment z_i. Unlike LDA, j in LLDA can only be assigned as the topic from the topic label \lambda_d of text d .

3. Repeat the Gibbs sampling based on the rotation of the axis in step 2 until the Gibbs sampling converges.

4. Summarize the topics of each word in each document in the corpus to obtain the document topic distribution θ_d, and summarize the distribution of each topic in the corpus to obtain the distribution of topics and words in LLDA \beta_k

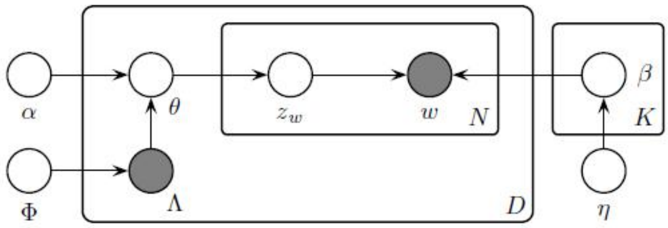


Figure . Graphical model of Label LDA(obtained from the original LLDA paper[5]), where \theta

is the topic distribution of the corpus, subject to the Dirichlet distribution affected by the parameter \alpha and the text labelset \Lambda, \theta determines the topic z of the text, and the word distribution of each topic \beta obeys the Dirichlet Distribution with parameter of \eta, topic z and topic distribution \beta together determine the words in the text.

We use the product category and the star rating of the review to construct the topic labels. After ignoring the 3-star review, the corpus is divided into six topics: Hair Dryer (negative), Hair Dryer (positive), Microwave (negative), Microwave (positive) , Pacifier (negative), Pacifier (positive). Then after filtering out reviews with helpfulness ratios that are below the threshold of 0.5 which have no reference value , we acquire the training dataset, and use LLDA Topic Model to find the topic words. Table x shows the 10 most frequent words(topic words) of the six topics, and the probability of the topic words appearing under the topic. After manual judgment, the topic words that can provide effective suggestions are marked in bold, which shows that LLDA can extract interpretable topic words.

microwave_negative	microwave_positive	hair_dryer_negative	hair_dryer_positive	pacifier_negative	pacifier_positive
samsung: 0.0304	oven: 0.0292	**conair: 0.0268**	blow: 0.0181	gate: 0.0495	baby: 0.0541
door: 0.0296	features: 0.0206	months: 0.0210	love: 0.0180	baby: 0.0419	love: 0.0306
whirlpool: 0.0270	cook: 0.0192	**fire: 0.0196**	time: 0.0126	**medicine: 0.0267**	easy: 0.0224
model: 0.0228	kitchen: 0.0192	**bonnet: 0.0196**	**heat: 0.0110**	**nipple: 0.0248**	paci: 0.0220
sharp: 0.0220	**easy: 0.0178**	**switch: 0.0189**	drying: 0.0092	**monitor: 0.0248**	bottle: 0.0179
appliance: 0.0220	door: 0.0171	blow: 0.0174	price: 0.0089	binky: 0.0248	bottles: 0.0168
repair: 0.0203	**cabinet: 0.0164**	**revlon: 0.0174**	dries: 0.0081	**lamb: 0.0209**	daughter: 0.0164
microwave_oven: 0.0203	**space: 0.0164**	**burned: 0.0167**	**cord: 0.0078**	arms: 0.0190	**nipple: 0.0149**
replaced: 0.0186	**sharp: 0.0157**	money: 0.0152	nice: 0.0075	these_pacifiers: 0.0171	**crib: 0.0131**
service: 0.0186	**button: 0.0157**	time: 0.0145	**conair: 0.0074**	hard: 0.0171	months: 0.0123
cost: 0.0177	cooking: 0.0149	**sparks: 0.0145**	**travel: 0.0060**	**hole: 0.0152**	**soothie: 0.0108**
replace: 0.0152	inside: 0.0135	**cord: 0.0145**	perfect: 0.0058	**wash: 0.0152**	**mattress: 0.0108**
call: 0.0152	model: 0.0128	**heat: 0.0138**	**power: 0.0058**	her_mouth: 0.0152	her_mouth: 0.0097
called: 0.0144	**watts: 0.0121**	**andis: 0.0138**	**diffuser: 0.0058**	**mattress: 0.0133**	**swing: 0.0097**
installed: 0.0135	**price: 0.0121**	started: 0.0123	**easy: 0.0057**	**chair: 0.0133**	**soft: 0.0090**
customer_service: 0.0135	**heat: 0.0114**	**power: 0.0123**	**skin: 0.0056**	**soap: 0.0133**	night: 0.0086
time: 0.0127	**popcorn: 0.0114**	**dangerous: 0.0123**	**straight: 0.0053**	babies: 0.0133	babies: 0.0086
warranty: 0.0118	**buttons: 0.0114**	**warranty: 0.0116**	**curly: 0.0052**	daughter: 0.0133	**wash: 0.0082**
replacement: 0.0118	nice: 0.0100	send: 0.0109	money: 0.0050	paci: 0.0114	hold: 0.0082
tech: 0.0118	range: 0.0100	disappointed: 0.0109	minutes: 0.0047	newborn: 0.0114	time: 0.0082

4 Model Implementation and Results

4.1 K-means Clustering Model for Helpfulness Rating

According to previous work and the analysis of the data mentioned earlier, we know that Amazon's helpfulness ratio, which is an evaluation of the helpfulness rating of review, is unreasonable. In addition, some papers point out that there exists sequential bias and preference bias in the helpfulness ratio [7]. At the same time, BC Wang el at. [8] proposed that the helpfulness rating of reviews will age over time, Nguy el al. [9]

suggested that the helpfulness rating of reviews is related to the length and the keywords of the reviews, and Tang, Jiliang, et al. believed that the helpfulness rating is also affected by the reputation of the evaluator.

Therefore, in this section, we propose a 2-clustering model, whose goal is to classify reviews into two categories: helpful and unhelpful. The model comprehensively considers the reputation of the reviewer (whether it is a member of amazon vine voices), whether the review is officially verified by Amazon, the date of the review, helpfulness ratio, review length, and the review keyword(represented as both TF-IDF and GloVe embedding) during the feature extraction of the review.

Before clustering, the feature vectors of the reviews need to be pre-processed. First, the vine and verified_purchase fields of the string type are converted into numeric types, specifically, if the value is N / n, it is mapped to -1, and if the value is Y / y, it is mapped to 1. Moreover, because clustering needs to compare samples, we standardize each dimension of the feature vector and map the values of each dimension to the interval [-1,1]. In doing so, we can prevent values of a certain dimension are much larger than other features, which results in the model can only learn the difference of features of this dimension between samples, and standardization can also speed up the convergence speed.

Clustering is an unsupervised classification model used to classify samples into several categories, and the requirement of clustering is to make the distance between samples of the same cluster as closes as possible and simultaneously the distance between samples of different samples is as far as possible. There are multiple implementation algorithms for clustering. The most commonly used KMeans clustring algorithm, exploiting the Euclidean distance, is selected in this paper, and the basic process of KMeans clustering is:

1. select initial cluster centers for k-mean clustering，

2. calculate the Euclidean distance between each sample and the current center of clusters, and classify the samples into the closest cluster，

3. after all the samples have been classified, each cluster center is updated according to all the samples in each cluster, and an iteration is completed here.

4. Return to step 2 to reclassify all samples until the cluster to which all samples belong no longer changes, in another word, the model converges, or stop when the number of iterations reaches the set maximum threshold (this article is set to 300).

(补充距离函数)

The Euclidean distance between feature vectors of reviews in K-means Clustering Model for Helpfulness Rating is defined as:

$$\text{dist}_{i,j}=\sqrt{\sum\limits_k^8(\text{x}_{i,k}-\text{x}_{j,k})^2}$$

where dist_ {i, j} represents the Euclidean distance between the feature vectors X_i and X_j, and x_ {i, k} represents the k-th feature of the feature vector X_i. The feature vector X has 8-dimensional features, which are: vine, verified purchase, review date, helpfulness ratio, TF-IDF of keyword, review length, 1-st dimension of embedding of keyword, and 2-nd dimension of embedding of keyword.

Table x illustrates the two cluster centers obtained by training. Obviously, the two clusters have significant differences in the characteristics of verified purchase. Making use of this property, we can mark the cluster with a value of -1 in the verified purchase dimension as ‘unhelpful’ and the cluster with a value of 1 as ‘helpful’.

Table. The center of the two clusters after training

label	vine	verified_purchase	review_date	Helpfulness Ratio	tf-idf_max	review_len	embedding11	embedding2
unhelpful	-0.90877193	-1.	0.28096328	0.62035318	-0.09855958	-0.72670296	-0.23040571	-0.03358746
helpful	-0.99774266 4	1.	0.5479719	0.58448887	0.0403084	-0.8083521	-0.16317656	-0.01255133

unhelpful[[-0.90877193 -1. 0.28096328 0.62035318

-0.09855958 -0.72670296

-0.23040571 -0.03358746]

helpful[-0.99774266 1. 0.5479719 0.58448887 0.0403084 -0.80835214

-0.16317656 -0.01255133]]

Establishing a formula for calculating Helpfulness Rating based on the distance between the sample feature vector and the center of two clusters:

$$\text{helpfulness rating}_i = \frac{\frac{1}{dist_{pos,i}}}{\frac{1}{dist_{pos,i}}+\frac{1}{dist_{neg,i}}}$$

where $dist_{pos, i}$ represents the Euclidean distance between the feature vector of the review d_i and the center of the helpful cluster, and $dist_{neg, i}$ represents the Euclidean distance between the feature vector of the review d_i and the center of the unhelpful cluster.

(画一个 ratio 和 rating 对比图)

表 1 review closest to ‘helpful’ cluster center and 1 review closest to ‘unhelpful’ cluster center in corpus

Category

pacifier

Row Num

382

marketplace

US

customer_id

10213051

review_id

RR9T0D3HLV2XX

product_id

B005G37X4M

product_parent

379901061

product_title jj cole pacifier pod, mixed leaf (discontinued...

product_category

Baby

star_rating

5

helpful_votes

8

total_votes

10

vine

-1

verified_purchase

1

review_headline Easy

access paci pouch

review_body This pacifier pouch does the job very well.

It...

review_date

0.677019

Helpfulness Ratio

-0.777778

review easy access paci pouch this pacifier

pouch do...

phrase_reviews easy access paci pouch this pacifier pouch

do...

keyword

easy

tf-idf_max

-0.219847

review_len

-0.758621

helpfulness_rating

0.916642

Category

hair_dryer

Row Num

396

marketplace

US

customer_id

45107362

review_id

RR5M45RA6CHPD

product_id

B001B0TJCI

product_parent

862140913

product_title blo and go by laurie coleman - portable hair d...

product_category

Beauty

star_rating

1

helpful_votes

8

total_votes

10

vine

-1

verified_purchase

-1

review_headline

Disappointed

review_body Very disappointed with this product. It is so ...

review_date

0.167702

Helpfulness Ratio

0.333334

review disappointed very disappointed with this prod...

phrase_reviews disappointed very disappointed with this prod...

keyword

disappointed

tf-idf_max

0.159696

review_len

-0.868966

helpfulness_rating

0.0822725

Finally, ANOVA (Analysis of variance) is used to rank the importance of each feature to helpfulness rating. Through the results of selecting features according to a percentile of the highest scores based on ANOVA shown in Table x, we can see that the two most important features are verified purchased, review date, followed by the review length, which is also the most important text feature. Surprisingly, the helpfulness ratio basically has no effect on the helpfulness rating, which further illustrates the deficiency of the helpfulness ratio.

Table . results of selecting features according to a percentile of the highest scores based on ANOVA

15% verified_purchased，date

30% verified_purchased，date, rlen

60% verified_purchased， date, rlen, vine, max TF-IDF

80% verified_purchased， date, rlen, vine, max TF-IDF, embedding 1d

100% verified_purchased， date, rlen, vine, max TF-IDF, embedding 1d, helpfulness ratio, embedding 2d

4.2 Time series modeling

In order to more accurately analyze the possible changes in the reputation of the product in the online market (hereinafter referred to as "product reputation"), combined with the data provided in the data set, we use the product star rating reasonably here to quantify the product reputation, and use its as a characteristic statistic, a second-order autoregressive model for hair dryer and microwave oven and an autoregressive moving average model for baby pacifier were established. In addition, in order to facilitate the discussion of seasonal and long-term changes in product reputation, we reasonably choose the month as the time series unit.

4.2.1Autoregressive model and autoregressive moving average model

Autoregressive model, referred to as AR model, it predicts the current value X_t by a linear combination of one or more lagging periods. Specifically, a model with the following structure becomes a p-order autoregressive model, which is abbreviated as AR(p)：

$$\begin{cases} X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \\ \varphi_p \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E x_s \varepsilon_t = 0, \forall s < t \end{cases}$$

In particular, when $\varphi_0 = 0$, it is called a centralized AR(p) model. Here we use the Akaike Information Criterion (AIC) to determine the parameter p.

Autoregressive moving average model, referred to as ARMA model, is a time series model based on autoregressive model (AR model) and moving average model (MA model). According to Hamilton Smith[11] theory, using ARMA model to make predictions, first of all, it is necessary to perform a statistical test of the stationarity, and to determine multiple sets of model parameters through time series autocorrelation and partial correlation functions. Here we use the Bayesian Informationization Criterion (BIC) to determine a set of optimal model parameters.

4.2.2 Stationary Analysis of Time Series

Assume that a time series is generated by a random process, that is, each value of the time series $\{X_t\}$, $t = 1,2,...$ is randomly obtained from a probability distribution, if the following conditions are met: a. Mean $E(X_t) = u$ is a constant independent of time t; b. Var $Var(X_t) = \sigma^2$ is a constant independent of time t; c. Covariance $Cov(X_t, X_{t+T}) =$

γ_T is a constant that is only related to the time interval T and has nothing to do with time t, then the random time series is said to be stationary.

Here we use MATLAB software to perform Augmented Dickey-Fuller test and Daniel test on three groups of time series. The test results show that the three time series are all stationary series, and the results are consistent with the image characteristics of the three time series (Figure 1).

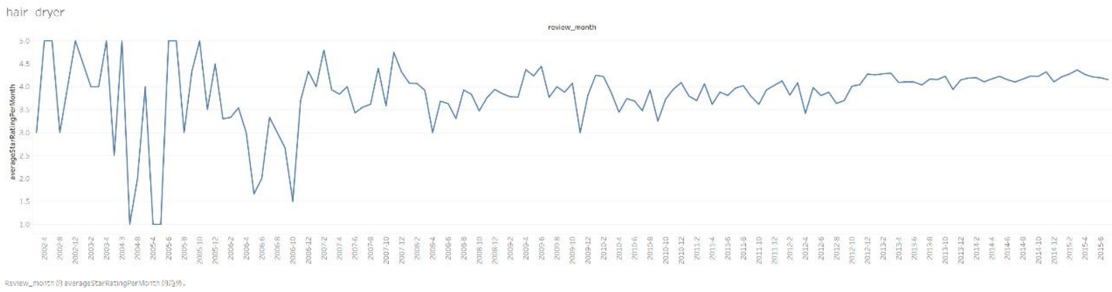


Figure 1(a)

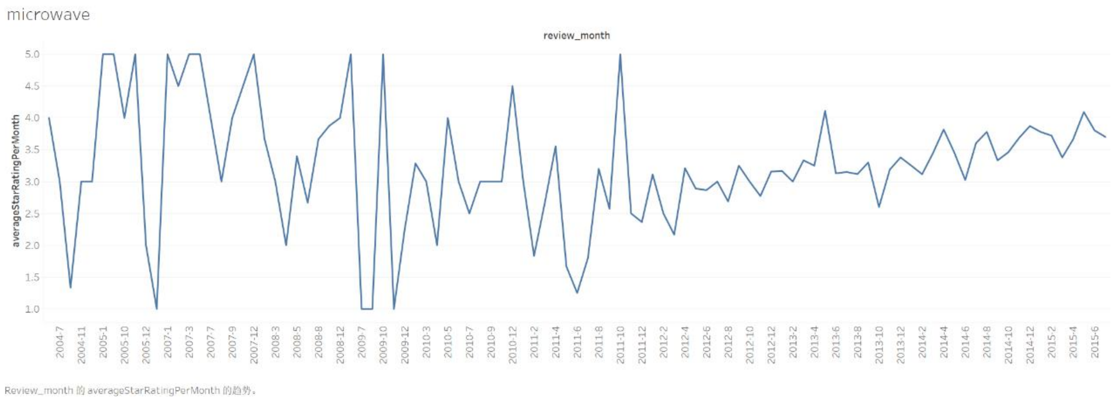


Figure 1(b)

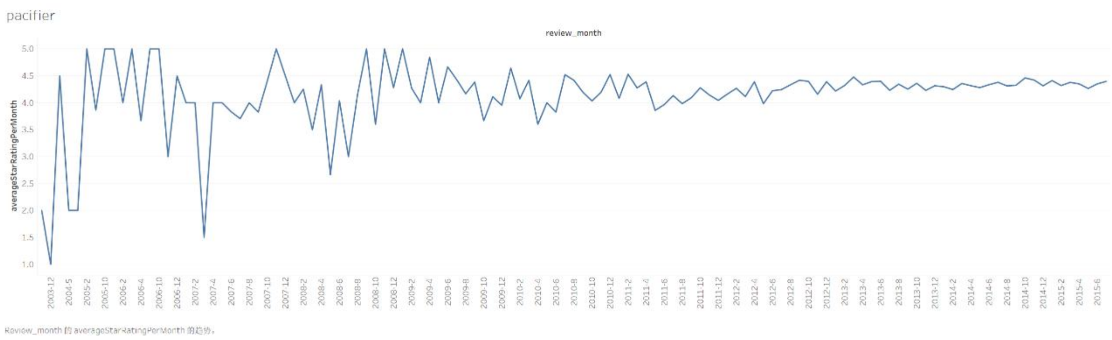


Figure 1(c)

For the Augmented Dickey-Fuller test, we used the adftest method provided by MATLAB software, which takes the existence of the unit root in the sequence as the null hypothesis, calculates the ADF statistics and the critical value of it at a given significance level, and compares the ADF statistics and the critical value. If the comparison result shows that the null hypothesis that the original sequence has a unit root can be rejected, then the original sequence is stable at the significance level.

Daniel test is based on the Spearman correlation coefficient. Unlike the ADF test, it's null hypothesis is that the sequence is stationary. Specifically, Daniel test calculates the Spearman correlation coefficient by using the rank statistic R of the time series $\{X_t\}$, that is, $q_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(t - R_t)^2$ and constructs the statistics $T = \frac{q_s\sqrt{n-2}}{\sqrt{1-q_s^2}}$ to decide: For a given significance level α, if the statistics T is greater than $\frac{t_\alpha}{2}(n-2)$, the null hypothesis is accepted, that is, the sequence is stable and can be accurately predicted, otherwise the null hypothesis is rejected, and the sequence is not stable. Taking $\alpha = 0.975$, We wrote MATLAB code to detect three sets of time series according to this principle. The results (Table 1) show that at the significance level of 97.5%, the three time series corresponding to the three products are stationary .

	T	$\frac{t_\alpha(n-2)}{2}$	Stationary

hair_dryer	0.617582335683350	1.977431212308178	True
microwave	0.524798680677441	1.984723186013982	True
pacifier	0.573161168709384	1.980447598683397	True

Table 1

4.2.3. Model Recognition And Order Determination

For a stationary random time series, the basic principles of model recognition are usaully as shown in Table 2[12].

ACF	PACF	Model
Trailing	p-degrees Truncation	AR(p)
q-degrees Truncation	Trailing	MA(q)
Trailing	Trailing	ARMA(p,q)

Table 2

3.1 AR(P) Model Ordering

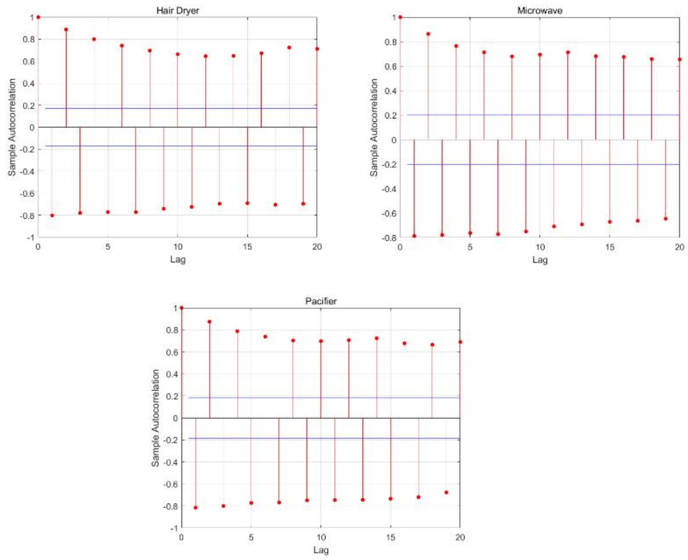


Figure 2

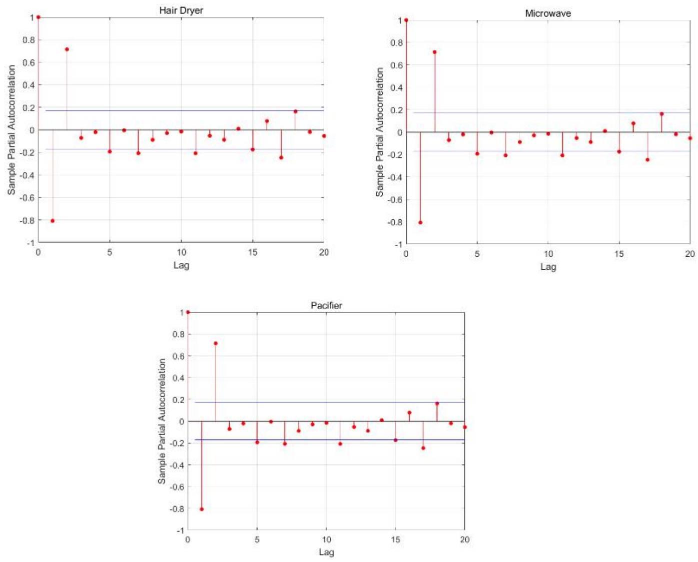


Figure 3

From the autocorrelation plots (Figure 2) and partial autocorrelation plots (Figure 3) of the three time series, we can observe that the autocorrelation functions (ACF) of the three time series are tailing, and the partial autocorrelation function (PACF) may be truncated. Therefore, we take the value of order p from 1 to 8, and try to establish the corresponding AR (p) model for the three time series, and use the aic method provided by MATLAB software to calculate the AIC value of the 8 models (Figure 4). Among them, AIC is a standard for measuring the goodness of fitting of statistical models. It is usually defined as[13]: $AIC = 2k - 2\ln(L)$, where k is the number of model parameters, and L is the likelihood function. When selecting the best model from the models, the one with the lowest AIC value is usually selected. In addition, we also use the Final Prediction Error(FPE) criterion to support our analysis (Figure 5). Its definition[13] is: $FPE(p) = \frac{N+p}{N-p}\widehat{\sigma_n}^2$, where N is the number of observation time series samples, p is the selected model order, and $\widehat{\sigma_n}^2$ is the model residual variance. Moreover, the model with the smallest FPE value is usually selected as the applicable model.

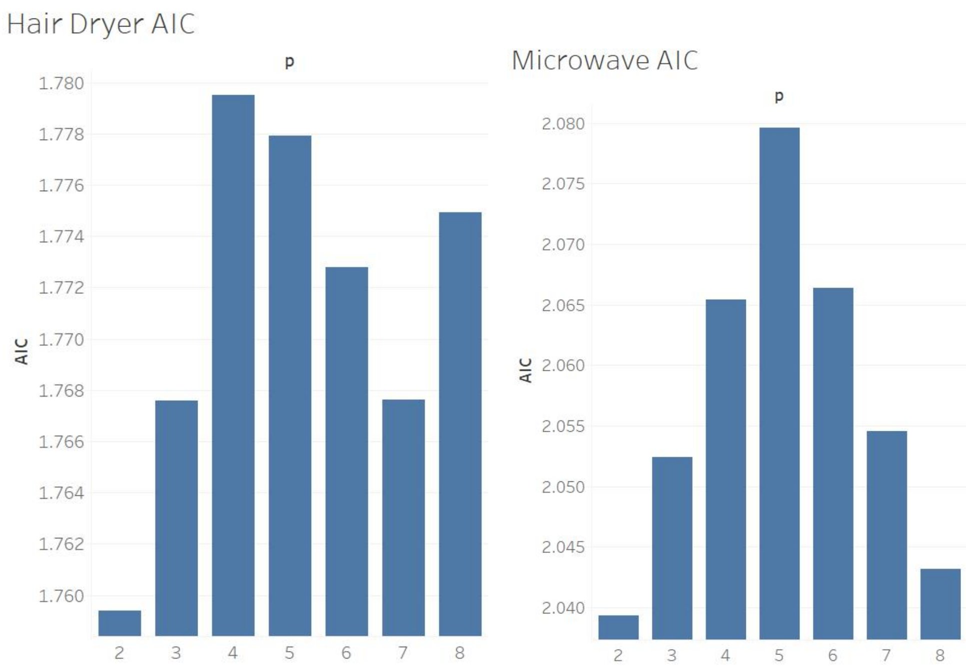


Figure 4

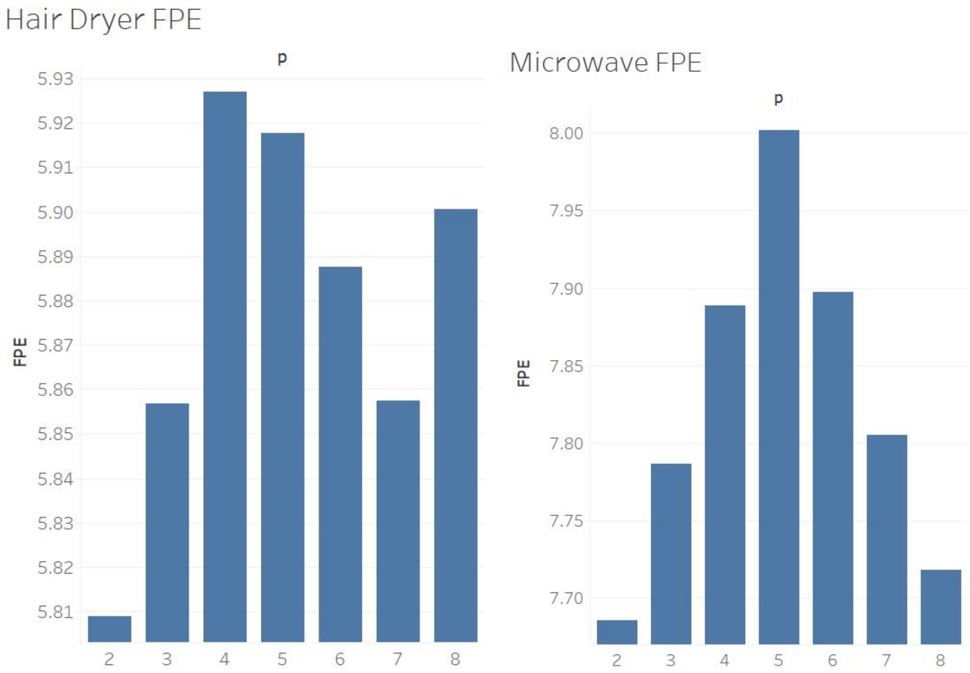


Figure 5

We have concluded through the experiments that hair dryers and microwave ovens can be analyzed and predicted using AR(2) models. However, when trying to establish an AR(p) model for a pacifier, the order p that minimizes the AIC value and the FPE value is 7, so we think that for baby pacifier, an ARMA (p, q) model should be tried instead.

4.3 ARMA(p,q) Model Ordering

We take the values of the order p and q from 1 to 6 respectively, and use the arima method provided by MATLAB software to calculate 36 sets of model parameters (Table 3), then use the Bayesian information criterion (BIC) to determine a set Optimal model parameters. The BIC definition[13] is: $BIC(p) = \ln \widehat{\sigma_n}^2 + p \ln N/N$, where N is the number of time series samples observed; p is the selected model order; $\widehat{\sigma_n}^2$ is the residual variance of the model. For 36 combinations, the model order with minimum BIC value is taken as the applicable model order, that is, $p = 3, q = 4$.

	p = 1	p = 2	p = 3	p = 4	p = 5	p = 6
= 1	591.8191	594.9387	592.5738	585.4352	590.4749	571.6361
= 2	588.0565	591.8396	600.8759	589.5241	592.9526	586.7818
= 3	592.3286	579.9982	597.4429	565.6617	585.7225	588.9007
= 4	580.7916	576.7992	587.1968	578.2552	586.204	592.8485
= 5	603.062	580.0842	568.403	577.8342	586.3038	597.5115
	595	582	581	594	566	571

= 6	.829	.8092	.5509	.3235	.9013	.6503

Table 3

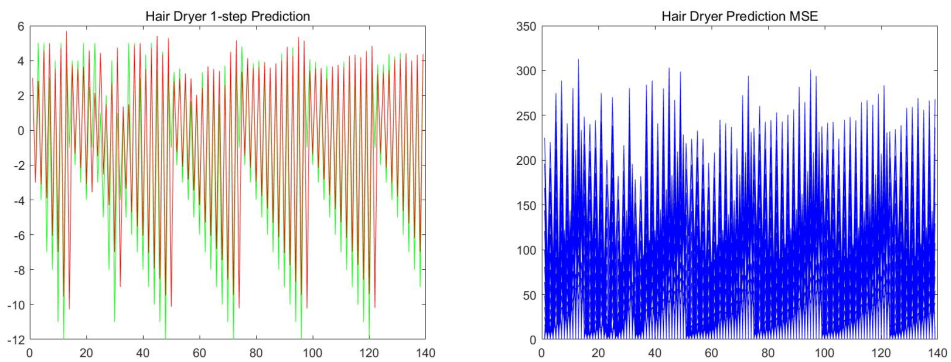
4.3.1 Model Solution

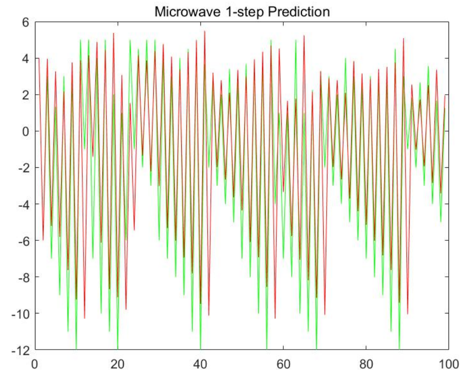
AR(p) Model Solution

Establish AR(2) model for product hair dryer and microwave oven:

$$A(z) = 1 + 0.1216z^{-1} - 0.8065z^{-2}$$

With the help of MATLAB software, we get the model's FPE value of 7.686, mean square error (MSE) of 7.381, and obtain the one-step prediction results for these two products (Figure 6). From the predicted images of the two products, we have concluded that the reputation of hair dryer and microwave oven in the online market has steadily changed over time. Moreover, this is consistent with the characteristics shown in the time series of the two products (Figure 7).




Microwave 1-step Prediction

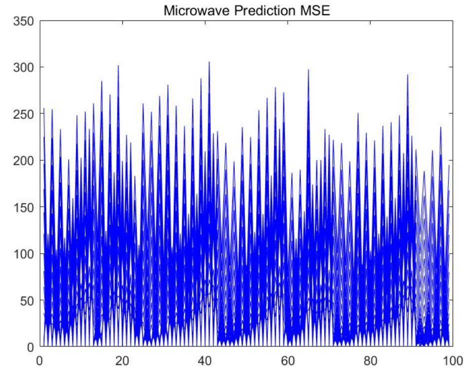
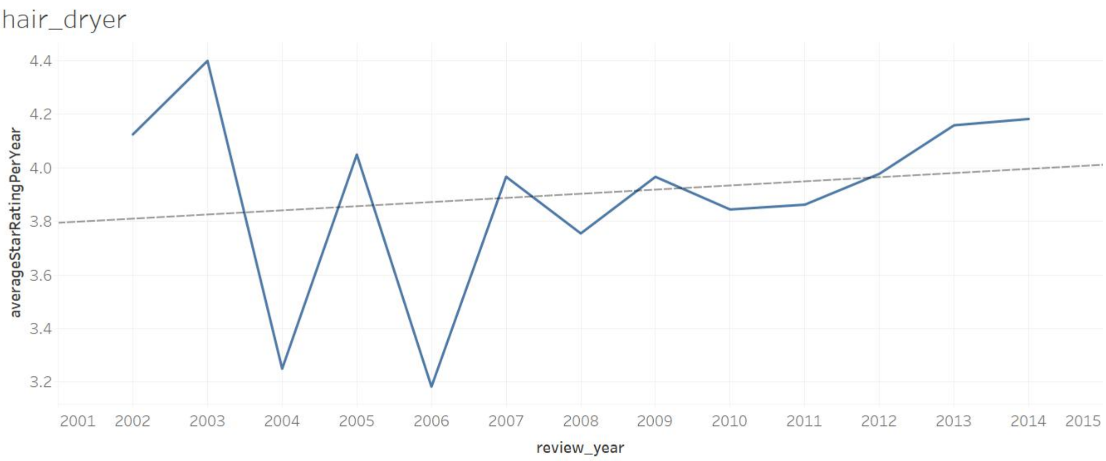

Microwave Prediction MSE

Figure 6


hair_dryer
averageStarRatingPerYear
review_year

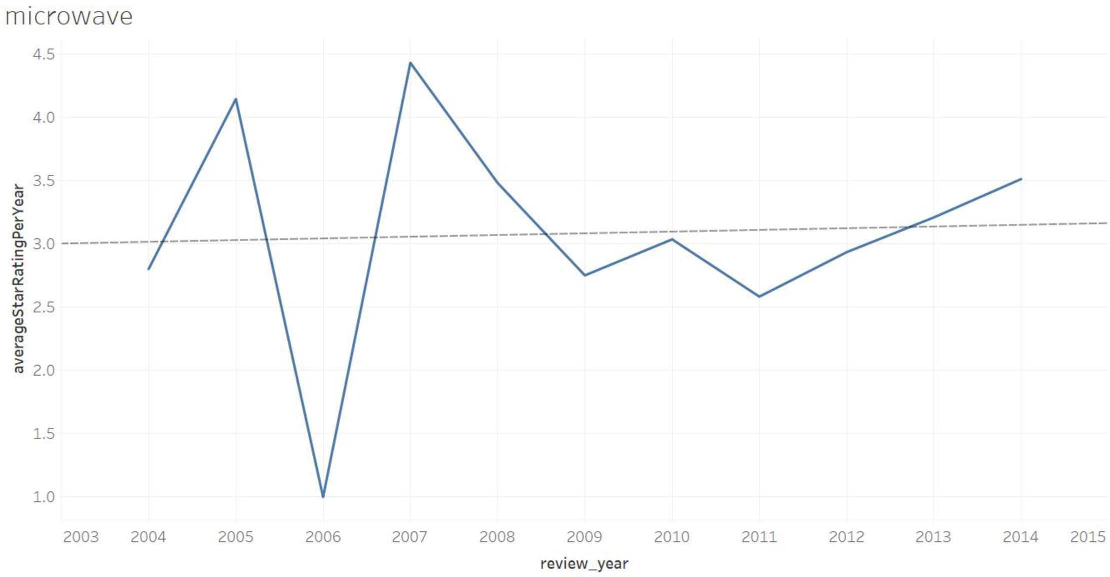


Figure 7

4.3.2 ARMA (p, q) model solution

Establish a discrete ARMA (3,4) model for the product pacifier. Because the estimation of the ARMA (3,4) parameters is non-linear, it is difficult to obtain accurate estimates of the ARMA model parameters. We estimate AR parameters and MA parameters separately legitimately (Table 4), thus greatly reducing the amount of calculation

lags	1	2	3	4
AR params	-1.3647	0.092675	0.554139	/
MA params	0.735827	-1.10166	-0.739466	0.124298

Using MATLAB software, input the original data to get its variance 8.5556 and covariance matrix, and get the prediction result of this product (Figure 8). From the predicted image, we conclude that the reputation of the product's pacifier in the online market shows a certain upward trend over time, which is also consistent with the characteristics shown by the time series of the product in time (Figure 9).

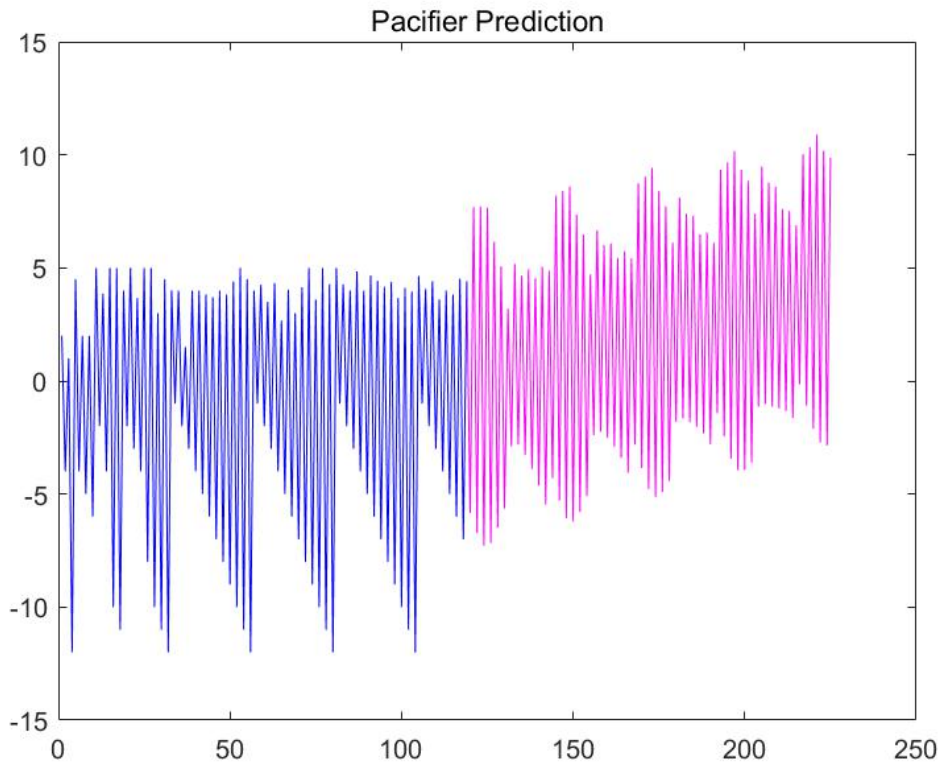


Figure 8

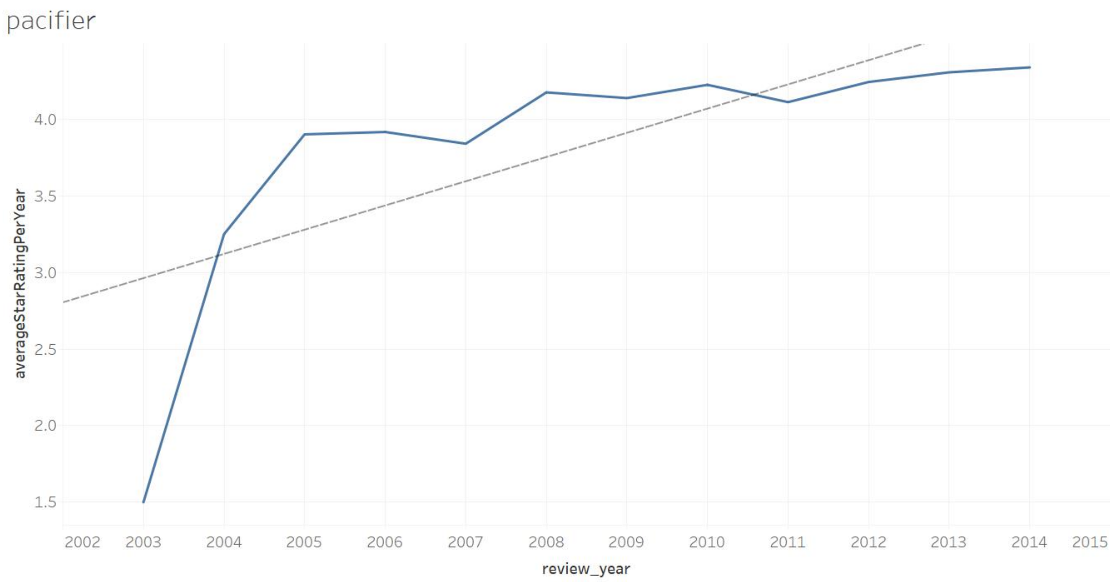


Figure 9

2c – Reputation Value

We have previously identified a text-based measure that effectively reflects the helpfulness of reviews, HelpfulnessRating. Further, we reasonably define a combination of text-based and rating-based measures to indicate the potential success or failure of a product:

$$ReputationValue = HelpfulnessRating \times StarRating$$

Its value range is $[0, 5]$, The higher the reputation value, the higher the potential success of the product, otherwise, the lower the reputation value, the lower the potential failure of the product.

We have plotted the statistics of the different models of the three products (Figure 1), and marked the maximum and the minimum value in each image.

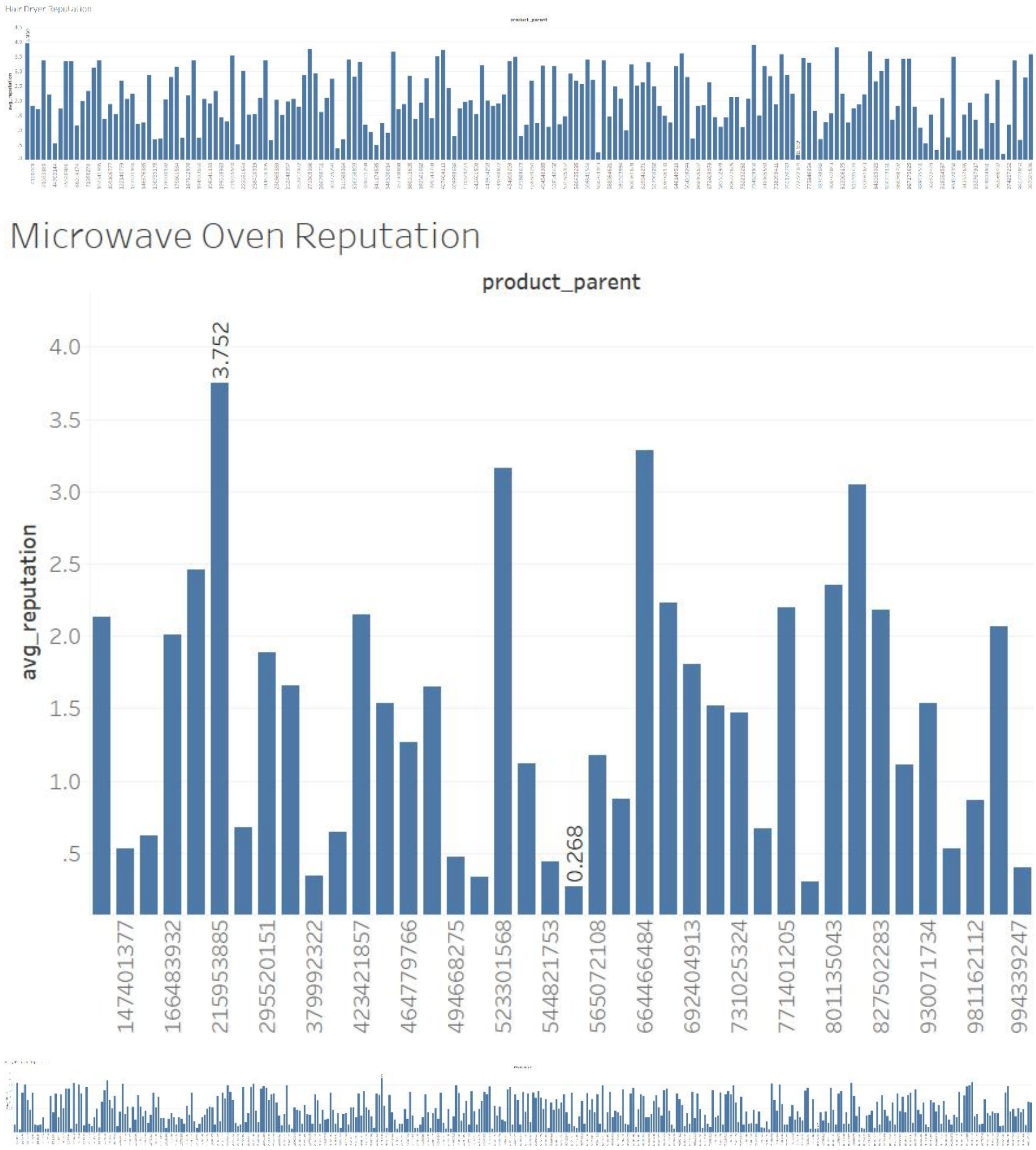


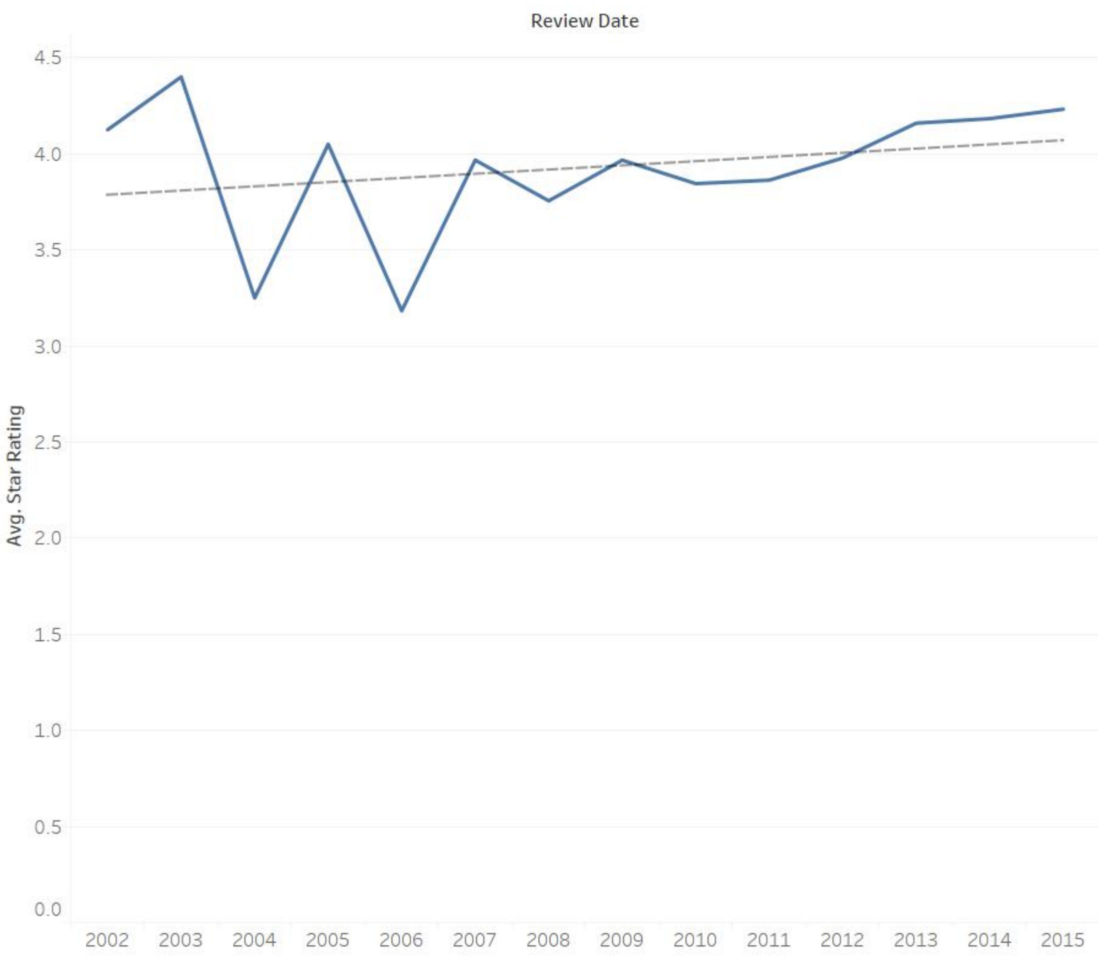
Figure 1

Obviously, for the hair dryer, the product 4120409 has the highest reputation of 3.960, and the product 772722324 has the lowest reputation of 0.152; for the microwave oven, the product 21953885 has the highest reputation of 3.752 and the product 55562680 has the lowest reputation of 0.268; for the baby pacifier, the product 379901061 has the highest reputation of 4.583 and product 801167869 has the lowest reputation of 0.217.

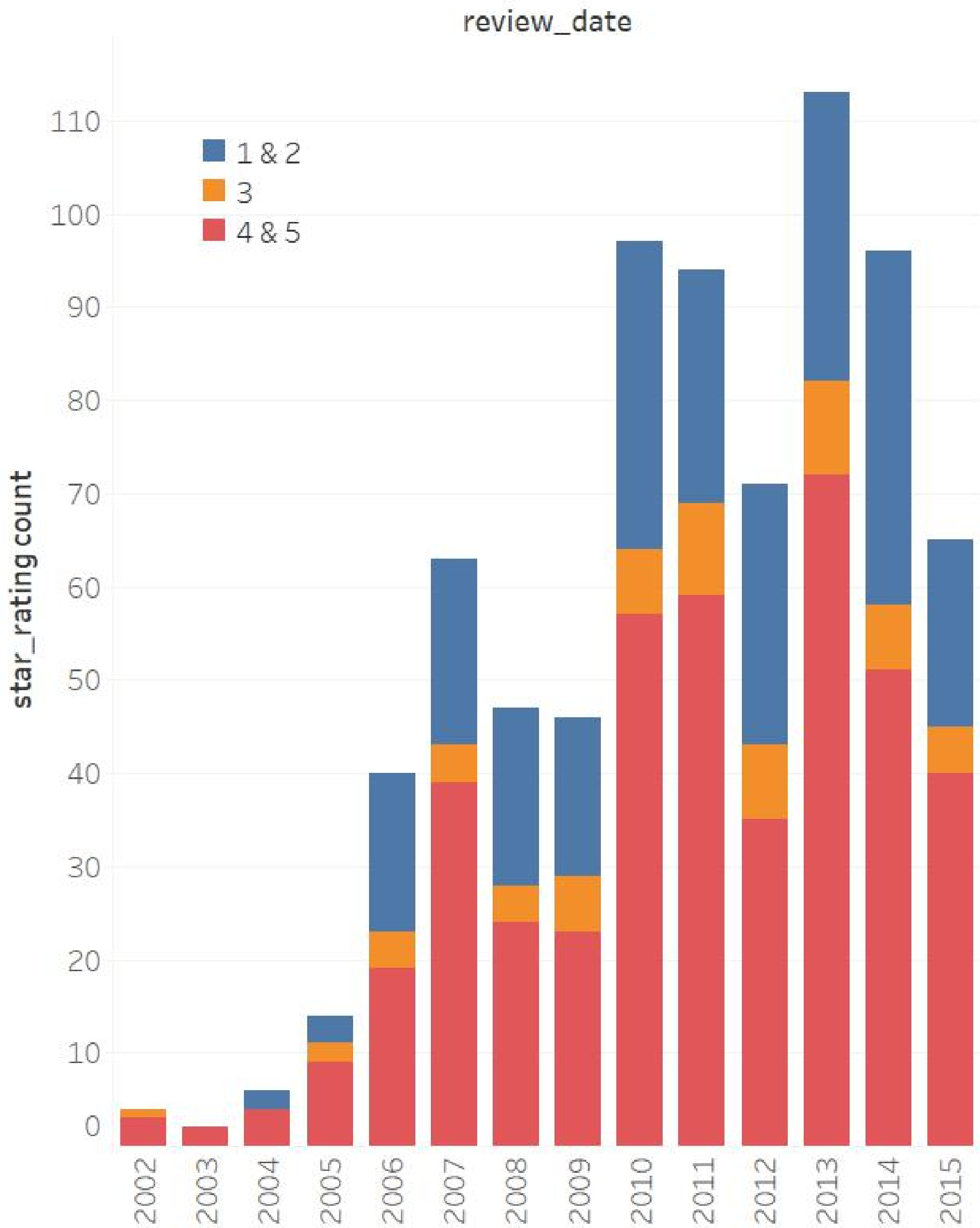
4.4 Star ratings model

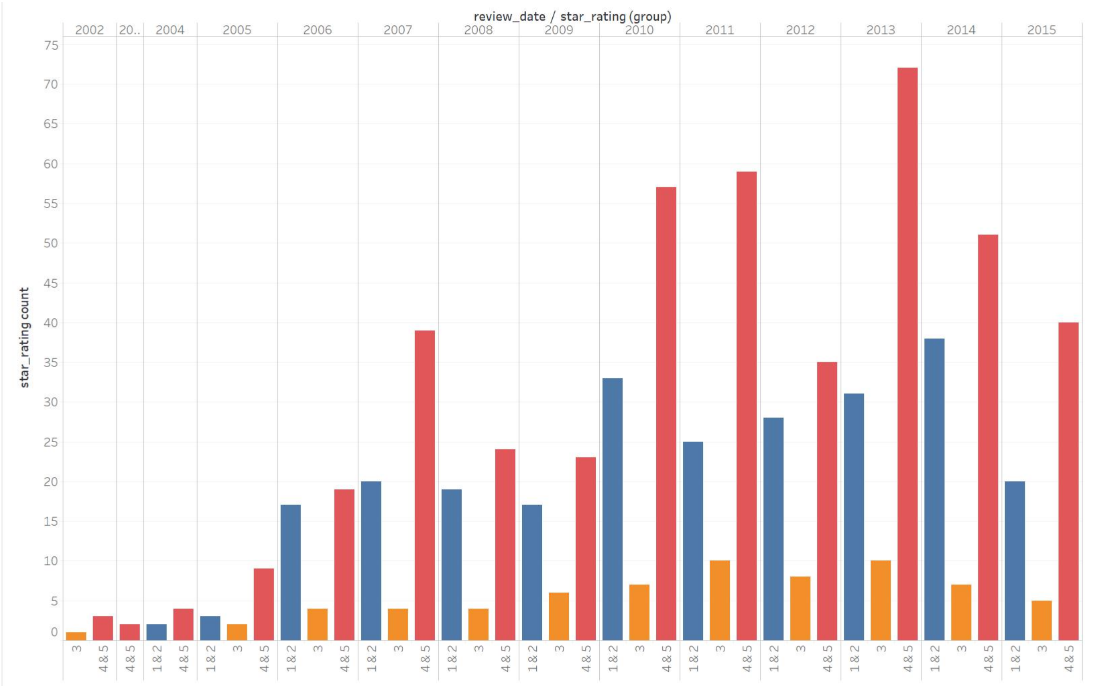
During model Analyzing, we observed that the histograms of the star rating stayed more or less constant over time,the number of all stars is increasing, the rate of increase of four or five stars is about five times that of one or two stars.(shown in the figure below).

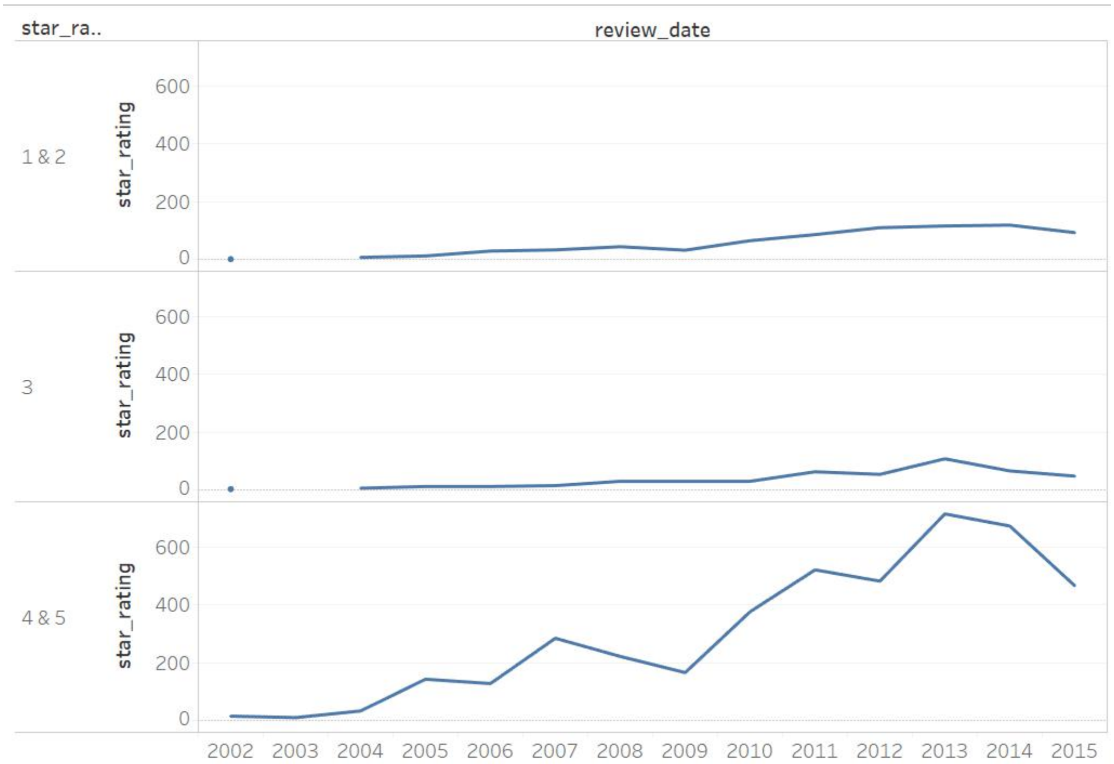
Star average:



Star count:


review_date
star_rating count
1 & 2
3
4 & 5
0
10
20
30
40
50
60
70
80
90
100
110
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015





The data provided to us has 15 related factors. Taking Amazon as an example, considering the real e-commerce review system reviews are sorted from high to low in number of helpful votes, and the ranking order affects the star status observed by customers. Specific star ratings will also cause consumers' emotional changes, so here we mainly choose 2 factors (helpful votes, review date) to develop a model and predict future star ratings. IIt will simplify our problem and enhance interpretability.

4.4.1 power

As shown in the figure, the best fitting effect is the power function. We select these two factors(helpful votes, review date) to predict the star amount X_t in year t. Then, our problem is that of estimating parameters A and B such that:

$$\ln(X_t) = A\ln t + B$$

The image analysis program solves the optimal function and obtains:

$$\ln(X_t) = 1028.43 * \ln t - 7816.1$$

4.4.2 Time series analysis

Calculating star ratings for each month we obtain the the average,and then we arrange them in chronological order to get a series of average star ratings, and then predict the predicted value of the average star rating for each month in year t quantitatively:

For example:

2016 average hair dryer star forecas:

Jan	Feb	Mar	Apr	May	
5.000000	4.000000	4.000000	4.000000	4.500000	5.0 0
Jul	Aug	Sep	Oct	Nov	De
5.000000	2.500000	1.000000	2.000000	4.000000	5.0 0

4.4.3 Random forest

We formulate a radom forest model to account for the influencing factors of star ratings.Using historical data from the United States, we determine initial conditions for our model, this model leads to a computer simulation of randomize the use of variables (columns) and the use of data (rows), generate many classification trees, and then summarize the results of the classification trees.

We fit the model to the modified data and get the following data:

randomForest (formula = starRating ~ reviewDate + helpfulVotes , data = data)

Type of random forest: regression Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 1.618631 % Var explained: 4.27

IncNodePurity reviewDate 348.35

helpfulVotes 1090.07

It can be seen that helpfulVotes has a greater impact, so the star count cannot simply be described with time. The main reason is Amazon's special evaluation system, that affects the place reviews appear, stars customers browse to is also affected.

Star distribution.

4.4.4 Multiple linear regression

Because helpful votes have a greater impact, coupled with specific star ratings will affect consumer sentiment, and most online customer do not browse all reviews, plus reputation mechanisms such as Amazon platform vine, making users more willing to believe that the sorted review (star rating). We assume that n1, n2, n3, n4, and n5 are the count of 1 to 5 stars that already existed before time t. With the increase of time, the number of stars and the order of reviews constantly change. We can calculate the weight of each star by establishing a multiple linear regression model, to judge the emotional impact of particular star ratings on users , and predict what review users will give after browse some specific star ratings .

First arrange all data according to time order, then calculate the count of every star rating before each review, we set the optimal function as the following formula, and then calculate the predicted value of X_t compared with real data to observe the impact of a specific star.

$$X\_t = a * n1 + b * n2 + c * n3 + d * n4 + e * n5$$

We apply this strategy to three products and list the Coefficients obtained by fitting the data of three products:

hair dryer

(Intercept) n1 n2 n3 n4 n5 3.726e + 00 9.140e-04 1.962e-03 -3.038e-03 6.548e-04 -1.006e-05

Microwave oven

(Intercept) n1 n2 n3 n4 n5 3.113352 0.005134 -0.016323 -0.022364 -0.007976 0.008983

Baby pacifier

(Intercept) n1 n2 n3 n4 n5 4.036e + 00 1.930e-03 -6.127e-05 -2.695e-03 1.101e-03 -8.931e-05

It can be seen that the impact of specific star ratings is different for different product categories. The lower the star value of higher product value, the greater the impact.In reality, we measure more when buying high-priced products.Theoretically, the model is fit to truth.

4.4.5 Improve:

In consideration of true situation, this strategy is not optimal but can be improved .We can sort comments according to the actual helpful votes, then set a number n , represent the mean number of comments per consumer browses , and then calculate the number of each star in the front n star ratings.Last we establish the multiple regression model to calculate the weight of each star.

4.5 Affective word recognition model based on majority vote algorithm

We believe that there is a strong correlation between specific star ratings and some quality descriptors. Obviously, the sentiment expressed by the star rating should be consistent with the sentiment expressed by the review text. In other words, the comment corresponding to a one-star rating should not support the product, but should be opposed. This problem can be transfered to the emotion recognition task in NLP, and more specifically, to build an emotion dictionary applied to the current domain.

This article divides emotions into "positive" and "negative" two categories, using the star_rating of comments as the emotion label, ignoring 3-star reviews to denoise, 4-5 reviews marked as "positive", and 1-2 reviews marked "negative". We think that adjectives have a stronger emotional tendency than other parts of speech such as nouns and verbs, so we only consider adjectives in the review.

We use the Boyer–Moore majority vote algorithm to learn the sentiment of words. At the beginning, we defined the model as:

$$\text{vote}_i = n_{pos, i} - n_{neg, i}$$

Where vote_i is the vote for the word w_i, $n_{pos, i}$ is the number of times the word w_i appears in a 4-5 star review, and $n_{neg, i}$ is the word w_i appears in a 1-2 star review times.

If a word has a positive vote, it indicates that the word is associated with a 4-5 star rating, with positive emotions, and if it's vote is negative, it indicates that the word is associated with a 1-2 star rating, with a negative emotion. If the absolute value of a word's votes is greater, it indicates that the word's emotional tendency is stronger.

Table x shows the results of the model solution. The output of the 20 words with the largest votes, that is, the strongest 20 positive emotions. Similarly, the 20 words with the smallest votes, that is, the strongest 20 negative emotions. It can be seen that the voting value of the positive word is much larger than the absolute value of the voting value of the negative word. At the same time, the recognition effect of the positive word is significant, but the recognition effect of the negative word is not satisfactory.

Table / picture majority vote algorithm the strongest 20 positive words and strongest 20 negative words

[[('easy', 'ADJ'), 522], [('nice', 'ADJ'), 286], [('clean', 'ADJ'), 204], [('perfect', 'ADJ'), 186], [('happy', 'ADJ'), 164], [('fine', 'ADJ'), 156], [('soft', 'ADJ'), 136], [('heavy', 'ADJ'), 125], [('hard', 'ADJ'), 125], [('powerful', 'ADJ'), 121], [('worth', 'ADJ'), 114], [('light', 'ADJ'), 98], [('regular', 'ADJ'), 97], [('natural', 'ADJ'), 95], [('short', 'ADJ'), 94], [('expensive', 'ADJ'), 92], [('professional', 'ADJ'), 91], [('quiet', 'ADJ'), 90], [('compact', 'ADJ'), 86], [('excellent', 'ADJ'), 84]]

[[('caught', 'ADJ'), -1], [('awful', 'ADJ'), -1], [('low_quality', 'ADJ'), -1], [('poorly_designed', 'ADJ'), -1], [('purchaser', 'ADJ'), -1], [('europe_plugged', 'ADJ'), -1], [('theory', 'ADJ'), -1], [('days_later', 'ADJ'), -1], [('lemon', 'ADJ'), -1], [('touchpad', 'ADJ'), -1], [('johnny', 'ADJ'), -1], [('embarrassed', 'ADJ'), -1], [('days_ago', 'ADJ'), -1], [('fine_print', 'ADJ'), -1], [('caught_fire', 'ADJ'), -1], [('recieved', 'ADJ'), -1], [('undependable', 'ADJ'), 0], [('dose', 'ADJ'), 0], [('matress', 'ADJ'), 0], [('dumb', 'ADJ'), 0]]

One of the reasons for the above problem may be the lack of negative samples and the difference in the length of the reviews. So we improved the voting model to eliminate the lack of comment length, and considered that 1-star reviews have a stronger negative emotional tendency than 2-star reviews, and 5-star reviews have a stronger positive emotional tendency than 4-star reviews.

\ text {vote} _ {i} = \ sum \ limits_j ^ N (s_j-3) \ cdot \ frac {n_ {i, j}} {L_j}

s_j indicates the star rating of the comment r_j, n_ {i, j} indicates the number of times the word w_i appears in the comment r_j, L_j indicates the length of the comment r_j, and N indicates the total number of comments in the corpus.

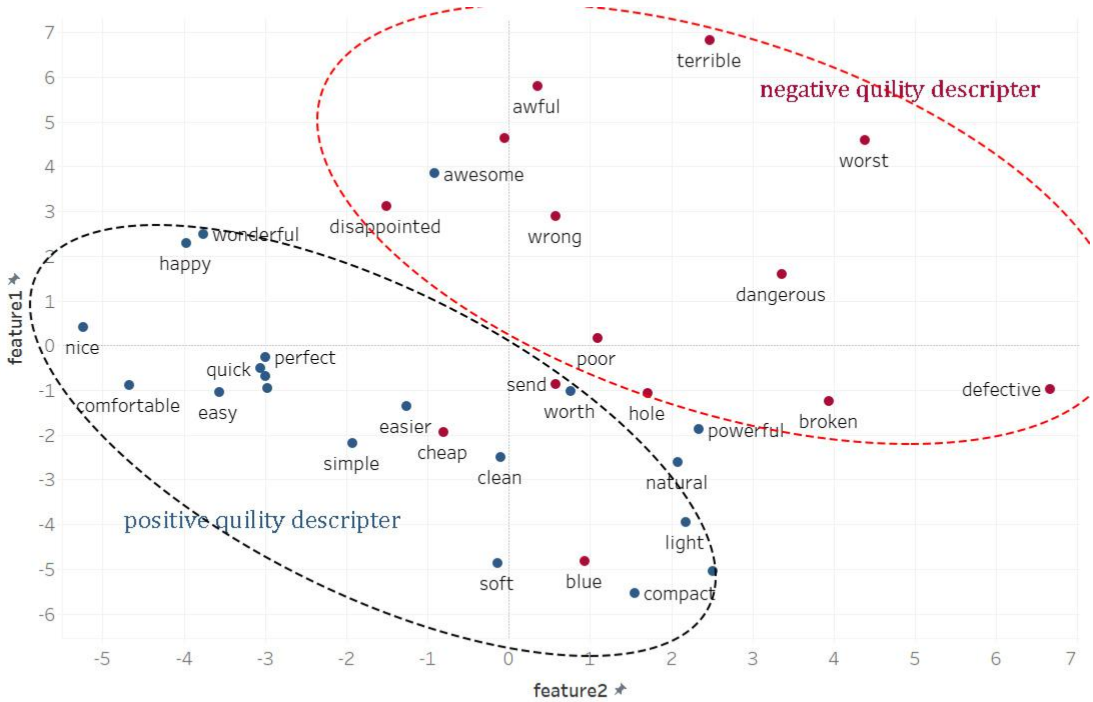
Table x shows the results of the improved model solution. Similarly, the 20 words with the largest and smallest votes are output respectively. It can be seen that the recognition effect of negative word is significantly improved.

Table / picture advanced majority vote algorithm find the 20 strongest positive words and 20 strongest negative words

[[('easy', 'ADJ'), 43.97199138748491], [('perfect', 'ADJ'), 24.473897686945012], [('nice', 'ADJ'), 21.303982846330708], [('awesome', 'ADJ'), 15.345700451385353], [('soft', 'ADJ'), 11.255412469432821], [('compact', 'ADJ'), 10.871837445086985], [('worth', 'ADJ'), 10.632973788003023], [('happy', 'ADJ'), 10.449347589281722], [('powerful', 'ADJ'), 10.247582098549556], [('excellent', 'ADJ'), 10.07544841695822], [('clean', 'ADJ'), 9.585166671571448], [('quiet', 'ADJ'), 8.92705633751037], [('comfortable', 'ADJ'), 8.846822594119873], [('easier', 'ADJ'), 8.217011399875933], [('light', 'ADJ'), 8.03204052335096], [('wonderful', 'ADJ'), 7.029998907747026], [('natural', 'ADJ'), 6.768304206186862], [('quick', 'ADJ'), 6.495497192895513], [('thick', 'ADJ'), 6.397459608370886], [('simple', 'ADJ'), 6.032317818355954]]

[[('disappointed', 'ADJ'), -17.233553701390132], [('terrible', 'ADJ'), -10.173947694612764], [('dangerous', 'ADJ'), -9.867731005969816], [('awful', 'ADJ'), -8.498114849187935], [('horrible', 'ADJ'), -8.060109503227293], [('defective', 'ADJ'), -7.858060442335373], [('don_t_waste_your', 'ADJ'), -7.623259860788863], [('less_than', 'ADJ'), -7.382524757369065], [('wrong', 'ADJ'), -6.825983158609728], [('worst', 'ADJ'), -4.804552058536977], [('broken', 'ADJ'), -4.5329504133818475], [('turntable', 'ADJ'), -4.335550270975752], [('blue', 'ADJ'), -4.315849679451768], [('cheap', 'ADJ'), -4.276342897794765], [('poor', 'ADJ'), -3.954190607584395], [('no_longer', 'ADJ'), -3.7408589897982316], [('save_your', 'ADJ'), -3.5269077081722093], [('hole', 'ADJ'), -3.501432928935955], [('scary', 'ADJ'), -3.495400326664828], [('send', 'ADJ'), -3.47027520705727]]

The 20 words with the most significant positive emotions and 20 words with the most significant negative emotions are represented as corresponding GloVe words, and then reduced to 2D features using PCA for visualization. Figure x shows that in the 2-dimensional word embedding space, the positions of homogeneous sentiment words are clustered, and different types of sentiment words are highly separable. On the one hand, it illustrates the rationality of the improved word sentiment model, and on the other hand, it proves the effectiveness of GloVe word embedding.



References

[1] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[3] Bojanowski, Piotr, et al. "Enriching

word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5 (2017):

135-146.

[4] Pennington J, Socher R, Manning C D. Glove:Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[5] Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

[6] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

[7] Wan, Yun, and Makoto Nakayama. "Are Amazon. com Online Review Helpfulness Ratings Biased or Not?." *Workshop on E-Business*. Springer, Berlin, Heidelberg, 2011.

[8] Wang, Bo-Chun, Wen-Yuan Zhu, and Ling-Jyh Chen. "Improving the amazon review system by exploiting the credibility and time-decay of public reviews." *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 3. IEEE, 2008.

[9] Nguy, Bobby. "Evaluate helpfulness in amazon reviews using deep learning." *Stanford University*. 2016.

[10] Tang, Jiliang, et al. "Context-aware review helpfulness rating prediction." *Proceedings of the 7th ACM conference on Recommender systems*. 2013.

[11] 汉密尔·顿·史密斯. 时间序列分析[M]. 夏晓华, 译. 北京: 中国人民大学出版社,2014:58-68

[12]薛冬梅.ARIMA 模型及其在时间序列分析中的应用[J].吉林化工学院学报,2010,27(03):80-83.

[13]李文涛,姜海波,王雪琴.自回归模型选择的多准则方法[J].统计与决策,2010(18):24-25.

[14]Y. Liu, X. Huang, A. An, and X. Yu, “Arsa: a sentiment-aware model

for predicting sales performance using blogs,” in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007, pp. 607–614.

[15]“Marketwatch.com. (2019). amazon.com inc..[online] available at:

https://www.marketwatch.com/investing/stock/amzn/financials [accessed

10 mar. 2019].”

[16]B. Fang, Q. Ye, D. Kucukusta, and R. Law, “Analysis of the perceived

value of online tourism reviews: Influence of readability and reviewer

characteristics," Tourism Management, vol. 52, pp. 498–506, 2016.

[17]Shadi,Abdalraheem, etc. "A Brief Analysis of Amazon Online Reviews",Sixth International Conference on Social Networks Analysis, Management and Security.2019.

Appendices

Appendix A