Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2005922

Product Sales Forecast Analysis Based on Data Mining and Other Related Technologies

Summary

In order to help Sunshine Company set suitable sales strategies and succeed in these three new online marketplace product offerings, our team makes an analysis and prediction on the trend of these prducts sales depend on many factors such as star rating, review body and so on.

For Part1, we first explore the relationship in star ratings, reviews and helpfulness ratings. After analysis, we find that Most people tend to give good or bad reviews, with a few in the middle. At the same time, if we want to analyze the products, we need to start with good reviews and bad reviews. The distribution of the degree of help, most of the comments, and the help / total rate are more than 98%, which is intuitive. After all, whether it's to write good or bad reviews, it's the evaluation of goods, and it's also helpful to other customers, but there are also some comments with the help rate in the middle. After screening, the credibility of this part of comments should be high, but it may be controversial, and it is worth further study. For each product, the average star rating and average comment length are required. It is found that most of the comments are few, which is also in line with the actual life. Many people just default or directly comment, and are reluctant to write too many comments. At the same time, for each different star, some people wrote a lot of comments, some people wrote a little comments. And we see that the total number of helpfulnotes obtained from each star rating review, it is obvious that most helpfulnotes are concentrated at both ends of high and low stars.

For part2, we use the tool nltk to participle the words in text and use the TextBlob to make emotional analysis. With the help of TextBlob, we will get the value of the text and we can judge whether the text is positive or negative. We assume that the sales number are the number of review bodyBased on the value of text and star rating, we design a model and get the weights of text value and star rating by using multivariate least square method so that we can make a prediction by using this model. And then, we analyze the relationship between time and star rating. We use the triple exponential smoothing method to make a time series so that we can predict the trend of star rating changes. We find that the star rating will be stable during a long time. In order to obtain the influence of previous stars on reviews, we will stagger the corresponding stars one by one and the emotional value of reviews one by one, make the time series difference manually, take stars as independent variables and emotional value as dependent variables, and bring them into the formula of finding correlation coefficient so that we can find the correlation between star rating and review. Through the NLP package TextBlob and the word segmentation package nltk, the words with the highest frequency and higher positive and negative emotions are analyzed, and the correlation coefficient is calculated to reflect the impact of the key words of comments on the star rating.

Finally, depend on the above analysis and conclusion, because we find review has more powerful influence on the sales number so we recommend that we can give some products to vines before formal sale and get the review from them. We can modify products according to their opinions and try to get good review and high star rating at the beginning.

Contents

1 Introduction **2**
1.1 Problem Background . . . 2
1.2 Literature Review . . . 2
1.3 Our Work . . . 2

2 Preparation of the Models **3**
2.1 Assumptions . . . 3
2.2 Notations . . . 3

3 Data Preprocessing **4**
3.1 Illegal Data Cleaning . . . 4
3.2 Invalid Data Cleaning . . . 4
3.3 Data Processing . . . 5

4 Specify the Reasonable Sales Strategy **5**
4.1 Estimate of Product Review Confidence . . . 5
4.2 Comparative Analysis of Three Products . . . 6
4.2.1 Overall Word-of-mouth Perception . . . 6
4.2.2 Separate Statistics Based on Quarterly Product Ratings . . . 7
4.2.3 The Number and Annual Distribution of Ratings for the Three Categories 8

5 Improve Customer Satisfaction **9**
5.1 Comment Number Distribution of Help Rate . . . 9
5.2 The Average Valid Vote in the Distribution of Rating Stars . . . 10
5.3 The Distribution of the Average Number of Valid Votes . . . 11
5.4 The Scatter Distribution of Average Comment Length and Average Score . . . 11

6 Requirements **12**
6.1 A+C Design Suitable Model for Predicting Sales Number Based on Review and Star Rating . . . 12
6.2 Calculate the Correlation Coefficient to Find the Relationship Between Star Rating and Review . . . 14
6.3 Calculate the Correlation Coefficient to Judge the Relationship between Key Words and Star Rating . . . 14
6.4 Reputation Forecast Based on Time Series Analysis . . . 15
6.4.1 Exponential Smoothing Method . . . 16
6.4.2 Model Simulation and Analysis . . . 17

7 Strengths and Weaknesses **18**
7.1 Strengths . . . 18
7.2 Weaknesses . . . 18

8 Conclusion **19**

Memorandum **20**

References **21**

1 Introduction

1.1 Problem Background

As the intersection of the era of big data and e-commerce, online comments are becoming the most concerned factor for online shoppers. Online reviews are primarily by consumers, based on their experience of using the same or similar products and providing information about the experience of goods or services to potential buyers. This concept was first proposed by Chatterjee (2001) when he studied whether commodity review would have an impact on buyers' purchase intention. Cui (2012) believes that the correlation between online reviews and word-of-mouth spread of products is greater than 1. However, Mo zan (2015) believes that buyers can customize reasonable consumption strategies for individuals through emotional analysis of others' comments, so as to reduce purchase risks.

The function of additional comment is an important innovation of the star rating system in recent years, which further improves the credibility of the original comment. It enables consumers to judge the objectivity, authenticity and practicability of the goods through the feedback of others within a certain period of time when the transaction is completed. And the merchant also often gives this kind of comment higher attention, in order to dig out their own marketing focus.

1.2 Literature Review

Sunshine plans to launch three products in the online marketplace: microwave ovens, baby pacifiers and hair dryers. In addition, the consumption and scoring data of three consumer goods of amazon were provided to assist our team in analyzing their reasonable competitive sales strategies. How to qualitatively and quantitatively mine the mathematical relationship among scores, comments, valid votes, time and their special combination and use it to predict their sales trend is the key of ontology.

1.3 Our Work

- First, we find a few points in these questions How to find the relationship in star rating review and helpfulness ratings and how we calculate helpfulness ratings. How to identify data mesures based on ratings and reviews, and how to quantify reviews. How to analyse the relationship between time and star rating. How to design a model base on star rating and reviews. How to find the relationship between the star rating and reviews. How to get the key words through the large data.

- Secondly, we assume the helpfulness ratings by a formula designed by our selves. And we try to quantify reviews through the length of reviews and plot picture to analyze the trend.

- Thirdly, we combine a and c to design a model based on multivariate least square method. We quantify the review by nlp and get the value which reflect the feeling of customers. And we use multiple linear regression equation to present it.

- Then, we use time series to predict the trend of star rating changs, because we find that as time goes on, star will change. We assume that time will influnce the star rating.

- In the end, we find the correlation coefficient will reflect relationship between star rating and reviews. And we use nlp to get the key words so that we can calculate the emotional value of key words and find the relation with star rating.

2 Preparation of the Models

2.1 Assumptions

Since the content and amount of data is limited, we made the following assumptions to help us improve our model.

1. Not verified purchases are regarded not fully credit.

2. Amazon Vine members are honest and trustworthy. Their reviews are fully credit.

3. Data is corrupt, such as reviews that contain non-English characters, and need pre-processing to clean up the corrupt data.

4. Let's assume that each comment corresponds to a purchase, and the two correspond to each other.

5. We used the Y, N ratio in vine as the minimum confidence rate considering only this parameter.

6. The buyers of the three types of products can represent the overall distribution of the market buyers.

7. All the data provided are non-malicious comments, that is, all comments are based on personal practical considerations and have some influence on the decision. And each in the data after the cleaning of the score data have reference value.

8. The sum of total votes and review-id was taken as the total page views, that is, each comment data represented a page view, and the purchase behavior occurred. However, as the number of views of potential consumers, total votes did not purchase for the time being. However, there was no browsing for comments without total votes, but the browsing probability of all comments was assumed to be the same.

9. If the header of the comment is consistent with the content of the topic, it will be treated as the default comment.

10. Suppose that all the people who participated in help-votes only participated once, and there was no case of multiple vote swipes.

11. It is assumed that the fluctuation of the data volume of the three types of products is little affected by the actual population growth. That is to say, the increase or decrease of product review volume is mainly determined by its sales strategy and quality, which is not affected by the increase or decrease of the purchase population.

12. Assume that the ratio of useful votes to the total number of comments is an elastic limit for the authenticity and usefulness of the comment.

2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
h	Helpful votes
t	Total votes
v	Verified purchase or non-verified purchase
γ	Amazon Vine member or not
α	Smoothing constant
S_t	Forecast based on previous period data
$MAPE$	Mean absolute percentage error

3 Data Preprocessing

3.1 Illegal Data Cleaning

Deletes rows that contain null data, as well as rows that contain illegal characters (non-english characters).

11464	US	42697102	R2EBUWR09JU	B000FS1W4U	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	I love it	love it	9/9/2014
11465	US	961562	RG08SVIA1MC	B00BT7UNLY	407404113	panasonic nano-e nano	Beauty	5	0	1	N	Y	Five Stars	love it锛?12/13/2014	
11466	US	8699272	R1DAPGUCAFI	B00BT7V4DA	407404113	panasonic nano-e nano	Beauty	5	0	1	N	Y	it is the good hair dryer	it is the good hair dryer锛岃 will continue use	
11467	US	843938	R13QIN3CJ4EY	B001QTW2FK	328811288	conair minipro tourmalin	Beauty	5	0	0	N	Y	Five Stars	really nice锛?10/19/2014	
11468	US	1191638	RDEN4FIA9I4KI	B000A3I2X4	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	Five Stars	great锛?8/31/2014	
11469	US	44583555	R2BM9N7S3R6	B000R80ZTQ	486774008	conair 1875 watt cord ke	Beauty	1	9	9	N	Y	It broke锛?"Owned it le	2/28/2013	
11470	US	41542353	R2TB402APP2Z	B0002G214U	685652978	conair soft bonnet hair d	Beauty	5	0	0	N	Y	Soft Bonnet Love It	Easy to use dry your hair fast easy to carried i	
11471	US	37484202	R66GQX14WVI	B000H0XV3G	963066492	vidal sassoon vs547 1875	Beauty	1	1	2	N	Y	quality problem锛佽鈴	1/10/2013	
11472															

Figure 1: The result of 3.1

3.2 Invalid Data Cleaning

- Because we assume that all data is referenced, all default comments (see assumption) (28 lines) are deleted.

1	marketplace	customer_id	review_id	product_id	product_par	product_title	produc	star_rat	helpful	total_vc	vine	verified	review_headline	review_body	review_date
1074	US	37179381	R23IDQOX198	B00GU2PWHS	466064538	conair 1875 watt turbo h	Beauty	5	0	0	N	Y	awesome look, good ha	awesome look, good hai	1/7/2015
1201	US	15023402	R26RKTB8A65Y	B000A3I2X4	235105995	revlon essentials 1875w f	Beauty	5	0	1	N	Y	Wife and daughter love	Wife and daughter love	1/9/2015
1298	US	2524412	RGJ968519KQH	B000A3I2X4	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	good	good	10/12/2014
1356	US	6343306	R184EVIXU22V	B003V264WW	732252283	remington ac2015 t|studi	Beauty	5	0	0	N	Y	EXCELLENT PRODUCT	EXCELLENT PRODUCT	10/14/2014
2043	US	3045622	R1CM6IIBIFJ6L	B000A3I2X4	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	Love it	Love it	11/13/2014
3832	US	26259230	R2YE7WK8D4H	B00005O0MZ	694290590	conair corp pers care 146	Beauty	5	0	0	N	Y	has held up for over ye	has held up for over yea	2/12/2015
3847	US	23463608	R21996X7302B	B003V264WW	732252283	remington ac2015 t|studi	Beauty	4	0	0	N	Y	Dries my long hair in ab	Dries my long hair in abc	2/12/2015
5138	US	39009863	RFP55RTPLXX6	B00BB8ZIW0	253917972	remington silk ceramic pi	Beauty	5	0	1	N	Y	i like	i like	3/15/2015
5210	US	39553333	R1RV648TSQO	B000FS1W4U	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	love	Love	3/17/2015
6735	US	9188895	R30HAYZICTLK	B00BB8ZIW0	253917972	remington silk ceramic pi	Beauty	5	0	1	N	Y	great value	great value	4/4/2015
7220	US	2879295	REWI9G5AZY9	B00MFBGATM	646149518	panasonic hair dryer nan	Beauty	5	1	1	N	Y	makes my hair smooth	makes my hair smooth	5/2/2015
8177	US	22457194	R2RO1EBBU52	B00BB8ZIW0	253917972	remington silk ceramic pi	Beauty	5	0	0	N	Y	excellent	excellent	6/20/2015
8259	US	11455170	R21HHNKFNEL	B00006IV22	357308868	conair 1875 watt dual vo	Beauty	5	0	0	N	Y	Great price.	Great price.	6/23/2015
8570	US	46557913	RJXX5U51ACVI	B001UE7D46	197856712	andis 1600w quiet hangu	Beauty	5	0	0	N	Y	Awesome	Awesome	6/6/2015
9396	US	14426496	RYO914LZ95O	B00006IV22	357308868	conair 1875 watt dual vo	Beauty	5	0	0	N	Y	Excellent product. Worl	Excellent product. Worke	7/28/2015
9635	US	50355569	RBVC0UK87C5	B00005O0MZ	694290590	conair corp pers care 146	Beauty	5	0	0	N	Y	good value	good value	7/8/2014
9751	US	42275461	R228TJT2VMB(	B000A3I2X4	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	Awesome	Awesome	8/10/2015
9882	US	50037859	R35CSZDSEON	B00005O0MZ	694290590	conair corp pers care 146	Beauty	1	1	1	N	Y	None available.	None available.	8/15/2014
9902	US	37383618	RTKNIPZHVDF	B007ROD3O2	614083399	salon sundry professiona	Beauty	4	0	0	N	Y	Good.	Good.	8/15/2015
9976	US	47439252	R347OYE5NGY	B0012DL8B4	897437023	conair vagabond folding	Beauty	5	0	0	N	Y	Small, compact, powerf	Small, compact, powerfu	8/17/2015
10000	US	44767434	R3FT3HHLO6C	B00005O0MZ	694290590	conair corp pers care 146	Beauty	5	0	0	N	Y	Great Product!! Thank y	Great Product!! Thank yc	8/18/2014
10372	US	16209825	R1WXXPRNQV	B0009XH6WI	197856712	andis 1600w quiet hangu	Beauty	5	0	0	N	Y	Perfect	Perfect	8/28/2015
10413	US	11306741	R2HRH424Q9C	B000FS1W4U	235105995	revlon essentials 1875w f	Beauty	5	0	0	N	Y	Excellent product!	Excellent product!	8/3/2014
10482	US	7588294	R2LF2GT610YC	B003V264WW	732252283	remington ac2015 t|studi	Beauty	5	0	0	N	Y	excellent product, work	excellent product, works	8/4/2014
10654	US	53017258	R3952WZRAX7	B003ICX810	415057628	pro beauty tools professi	Beauty	5	0	0	N	N	Great Seller! Shipped a	Great Seller! Shipped as	8/5/2014
11058	US	21006622	R2L35REZCE0E	B00132ZG3U	758099411	conair 1875 watt tourma	Beauty	5	0	0	N	Y	Girlfriend and lady frien	Girlfriend and lady frienc	9/19/2014
11060	US	41300006	R3G0ZNTDULV	B00132ZG3U	758099411	conair 1875 watt tourma	Beauty	5	0	0	N	Y	Excellent product!	Excellent product!	9/19/2014
11335	US	41074284	R20HFFHOBZZ	B0009PL4D8	959834931	turbo power twinturbo 2	Beauty	5	0	0	N	Y	Powerful but quiet.	Powerful but quiet.	9/30/2014
11472															

Figure 2: The result of 3.2

- Considering that baby pacifiers, hair dryers and microwave ovens are among the "priors" of commodities, customers cannot obtain the change information of product quality from the outdated data. In view of this phenomenon, we eliminated some data with a long time span. For example, the data table of the first hair dryer contains two outdated data of 2008, and the most recent year is 2011.

1	marketplace	customer_id	review_id	product_id	product_par	product_title	produc	star_rat	helpful_	total_vo	vine	verified_	review_headline	review_body	review_date
2	US	13199369	R31Z87MDTUD	B000LQB5YS	194533684	t3 bespoke labs 83888-se	Beauty	1	9	11	N	N	Greatly disappointed	My husband got me this	1/1/2008
3	US	13273034	RFUKTJXOAKW	B000BBS636	560455235	conair yb075w pro yellow	Beauty	5	7	8	N	Y	Simply The Best Blow D	This is the best standard	1/1/2008

Figure 3: The result of 3.3

3.3 Data Processing

- Textual affective analysis.

  This article directly preprocessed the word segmentation, part of speech tagging, and text features of the online comment text by calling the python package, and then calculated the emotion score with the emotion dictionary. Commenting on this basis, put forward in this paper, text sentiment analysis method, the term is used to analyze the emotional feelings dictionary with the method of combining point of information retrieval algorithm in an analysis of the emotional comments, considering the factors that affect the degree of emotional tendency direction or, these factors are language emotional comments on corresponding processing to improve the accuracy. For example, the two situations in vine and verified were combined as the four categories of people participating in the market. Then they were given different levels of confidence and used simple linear programming to combine them into weights of data to participate in the sentiment analysis. Based on the difference between comment topic and comment subject and the number of words of comment subject, the influence of comment quality on emotion tendency is considered into the emotion analysis model. Then the results of emotion analysis were applied to the prediction of commodity sales, and the autoregressive model was selected to predict the commodity sales, and the affective perception model for short-term prediction was proposed. The polarity of emotion analysis results will be characterized by the real number interval [-1, 1].

- Delete irrelevant parameters.

  The market and category attributes of the product are fixed. All products are aimed at the us market and fall into a specific category. Therefore, these two parameters are trivial for the establishment and analysis of the actual model and can be directly deleted to simplify the calculation.

4 Specify the Reasonable Sales Strategy

4.1 Estimate of Product Review Confidence

In order to solve the reasonable sales strategy, we first derive the formula of confidence. And explain it:

The confidence of a review can be described by equation (1):

$$\left(1-\frac{1}{e^{h+1}}\right)\times\frac{h+1}{t+1}\times v+\gamma = Confidence \tag{1}$$

where $v=\begin{cases}1, & \text{purchase is verified}\\ 0.7, & \text{purchase is not verified}\end{cases}$ and $\gamma=\begin{cases}1, & \text{review is written by vine member}\\ 0, & \text{review is not written by vine member}\end{cases}$

- A review from not verified purchase has less confidence than that from a verified purchase, since it is unknown to us whether the product is purchased on Amazon and the customer could be influenced by the deep discount, who miight write inaccurate reviews.

- Though a review gets more helpful votes are more likely to be useful and fair, it is also subject to **helpful_votes** - **total_votes** ratio. If ratio is low, the review probably is controversial and degraded in confidence.

- Parameter calculation: we used monte carlo algorithm to calculate the two unknown parameters, and tried to make the value of the confidence rate close to the proportion of Y people in vine. This pair of parameters satisfies the condition.

- Fractional terms: as an extra confidence weight caused by valid comments, the reason to add 1 to both numerator and denominator is to avoid illegal calculations of 0/0.

- Constant term: this section gives an extra weight to the amazon certified population to highlight their contribution to accurate reviews of amazon products.

- Exponential term: when the parameters take greater than or equal to 0 h value, the entire index item will take [1-1 / e, any value between 1). First the by + 1 at the same time also ensure the non-zero properties, and this is for the sake of universality. Because when the value of help - votes small, he only represent a limited view of part of the crowd, for us to study the integral design of the product supply and marketing, do not have much practical significance. And e chosen as the change trend, because when that increasing the number of effective comment, the credibility of the derivative increase gradually, that is to say, the USES of the comment can reaction gradually increase.

4.2 Comparative Analysis of Three Products

By using python's matplotlib library, we made a correlation analysis of the three products.

4.2.1 Overall Word-of-mouth Perception

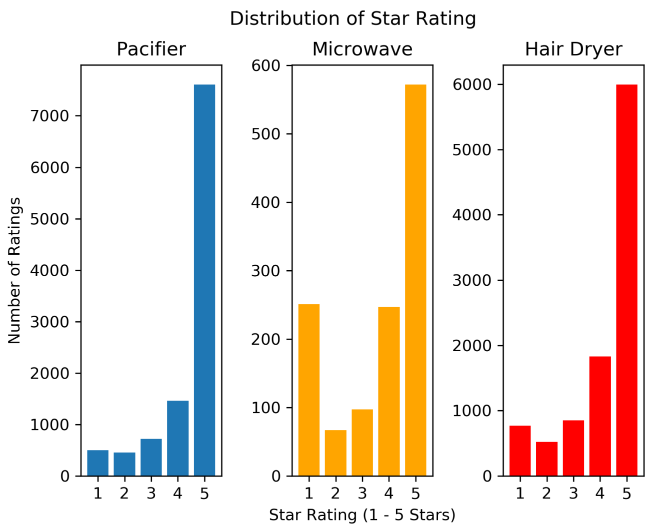

Figure 4: The result of 4.2.1

- First we from three kinds of products to evaluated the overall score, Bass model proposed by starting from the individual consumer value judgment to the new product, put forward to promote the new product diffusion of two factors, in addition to the mass media such as advertisement, and word of mouth or score. The latter with its high influence and strong commercial value, the impact of the role far more than newspapers, commercial advertising and personnel promotion. According to the polarity and its influence, the score can be divided into positive and negative. While the positive effect of a positive rating on consumer adoption appears to be the same as the negative effect of a negative rating on the spread of the product, the negative effect is actually more significant than the positive effect. In other words, when users face the same gains or losses, they will be more sensitive to losses.

- Based on the foreground theory, it is easy to find that pacifier (9.52%) and hair dryer (11.22%) in the low section (for the time being, comments rated 1 and 2 are considered as low sections and comments rated 4 and 5 as high sections) show less comment quantity distribution than in the situation of the situation, indicating that the overall quality of both products is higher than that of the situation. The 1 minute number of comments on the product is significantly higher than the 2 and 3 star number, and the 3 is not a continuous decreasing trend. It indicates that the reason for the 1 star bad rating may be not only the quality of the product itself, but also the conflict between the design concept of the product itself and the personal will of some buyers.

- Compared with pacifier and hair dryer, the score number between [1,4] of the latter was significantly higher than that of the former, and the range of the latter was lower than that of the former. Therefore, the 5-star comment of the former is more representative, and the hair dryer is more likely to have certain quality problems in all kinds of people. Moreover, the score distribution of pacifier is close to the exponential distribution, which makes it a good marketing case. We can dig out the reason of pacifier's good reputation from the 5-star praise.

4.2.2 Separate Statistics Based on Quarterly Product Ratings

Separate statistics based on quarterly product ratings

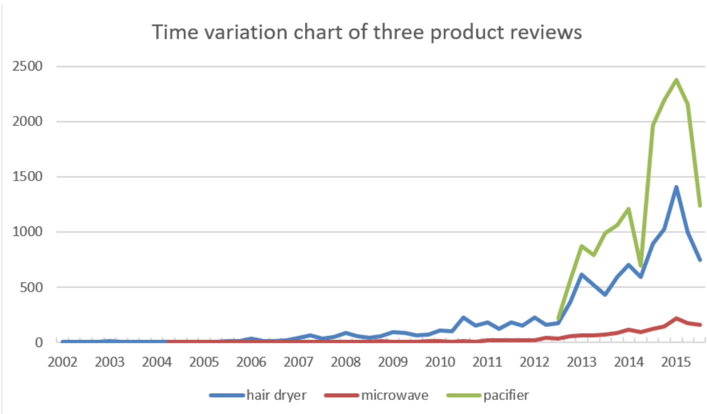

Figure 5: The result of 4.2.2

- But the previous analysis was problematic. Because this chart only shows the distribution of the overall historical score of the product, if a product enters the market with a bad reputation for various reasons, but after a period of strategic adjustment, the product is revitalized and occupies the market. We can not unilaterally rely on its relatively high negative rating ratio to analyze it, so as to make the wrong market strategy for the product.

- Therefore, we introduce the time series and analyze the change trend of three products with the help of the line graph. Here we use quarters to simplify the timeline and temporarily analyze it in terms of the number of reviews, or sales.
- The number of online reviews represents how much users like the new product and how popular it is. It is easy for buyers to follow the crowd and choose the product according to the intuitive feeling brought by the number. Chen's empirical research on books shows that there is a significant positive correlation between the number of book reviews and their promotion speed.
- Based on the above theory, we can intuitively see that the development trend of the three types of products is increasing. However, after curve fitting by least square method, we find that although all three of them are the trend of concave functions, the curvature changes the most slowly. Our analysis shows that while all three are the "search" in our product, they are relatively stable because they have significantly longer durability and tend to last longer than the others.
- It is worth noting that in the second half of 2012, the curve of hair dryer showed an inflection point and the volatility value began to get larger, which may be an important time for us to pay attention to the content of his comments. Similarly, for pacifier, we should extract its review data from the first half of 2014 and investigate the reason for its significant decline.

4.2.3 The Number and Annual Distribution of Ratings for the Three Categories

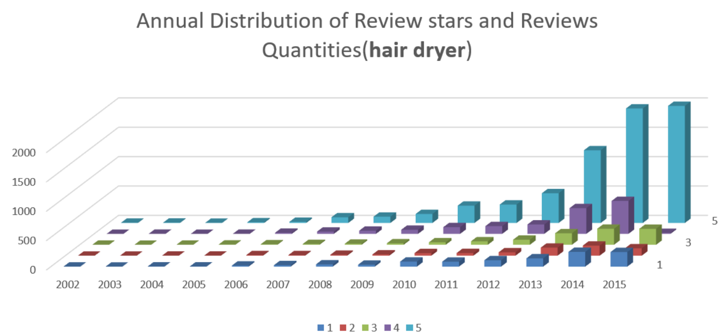

Figure 6: The result of 4.2.3

- The above model only stops from the overall trend to judge its development trend, is not perfect. Therefore, on the basis of the number of comments per year, we added the distinction of rating stars and analyzed them respectively.
- First of all, we can see that the trend of 5-star evaluation for the three products has been going up, but only the growth rate is relatively stable. However, when we analyzed the star rating data of the last two years laterally, the range between the high and low grades was still very small. The reason is not the low trend of the high section, but the high proportion of the low section review in the past five years, although it has a downward trend. So when we were making our decision for microwave, it was time to focus on solving the problem of its one-star newsletter, which has been around for nearly five years.

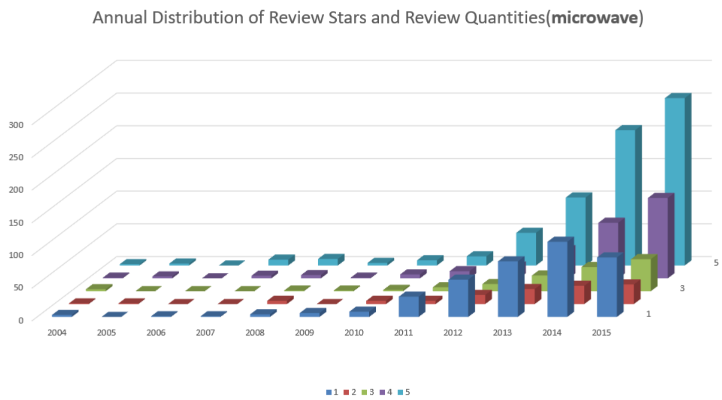

Figure 7: The result of 4.2.3

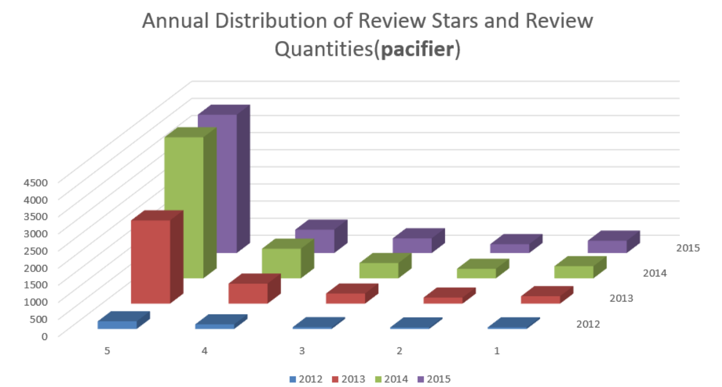

Figure 8: The result of 4.2.3

- For products such as pacifier, although the star rating is concentrated on 5-star review every year, the 5-star review has increased and stagnated in 2013 and 2015 respectively. The focus of the study is to pay attention to 5-star review in these two years.
- It seems that Hair dryer has a good growth trend and focuses on 5-star comments. However, comparing the data in 2014 and 2015, it can be found that 4-star comments plunged in 2015 and the proportion of 4-star and 5-star comments in that year decreased for the first time. So the 2015 4-star review will be the key to developing a reasonable sales strategy for such products.

5 Improve Customer Satisfaction

5.1 Comment Number Distribution of Help Rate

- Here the help rate is equal to the ratio of the available votes to the total votes cast.
- It can be found that the distribution curves of the three kinds of goods show a high similarity. Most of the reviews had a help rate of over 98%, and produced a second higher peak at 0.5. The reason for this phenomenon may be that most of the comments granted valid votes do have some reference value to the buyers themselves, which is intuitive. After all, whether you write good or bad reviews, there is always a certain degree of objectivity in the evaluation of a product, so that other customers can get help from it. But there is also a portion of the help rate in the middle of the comment. After screening, the

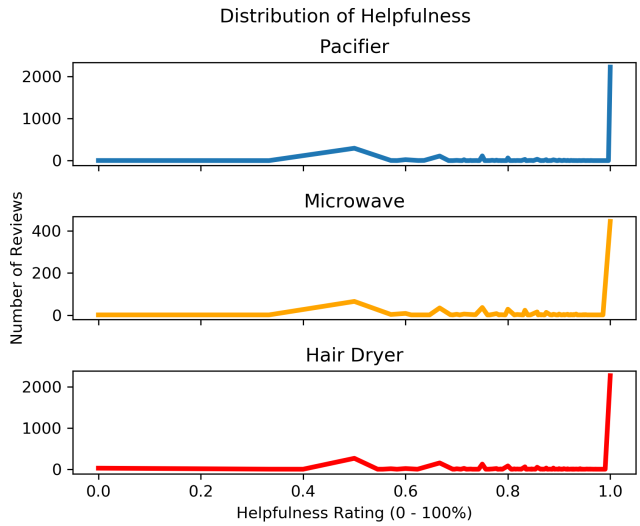

Figure 9: The result of 5.1

credibility of this part of comments should be high, but it may itself be controversial. The comments themselves are mixed with high subjectivity, thus resulting in the diversion of the browsing crowd.

5.2 The Average Valid Vote in the Distribution of Rating Stars

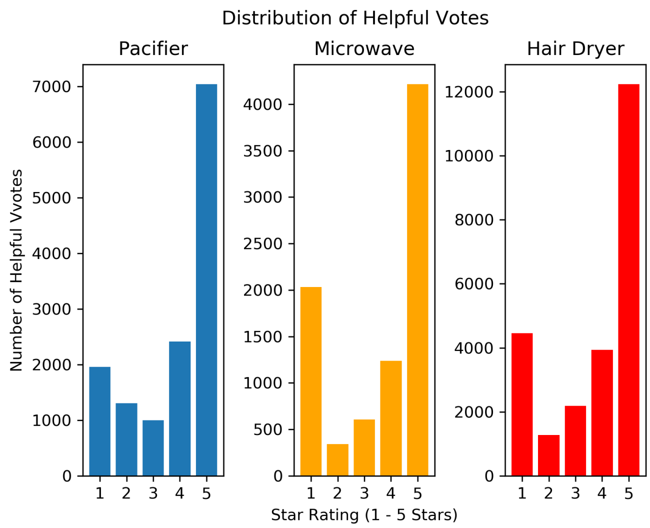

Figure 10: The result of 5.2

- 5.1 the model diagram is very limited in that it can only summarize the number distribution of valid comments in our three data sets, and we cannot get any content related to decisions from it. Therefore, we introduced the model diagram of 5.2.
- By comparing the relationship between the number of comments and the rating of stars, it can be seen that pacifier and hair dryer have a larger frequency rebound in the lower segment. For example, the effective vote of pacifier in the 1-star category has a nearly 300% rebound compared with that in the 5-star category. There is no doubt that this does not meet or even far exceed the validity and practicability of the negative and positive

evaluation in our understanding. Because while negative comments tend to influence buyers' buying intentions more than positive ones, almost all bounces above 100% are abnormal. However, it has performed well in this type of distribution. Although the 4-star reviews have shrunk in the amount of feedback, the amount of feedback has not exceeded 100%. Therefore, the utility of the 5 ratings of this product is basically the same, and the comments are more consistent with the public.

5.3 The Distribution of the Average Number of Valid Votes

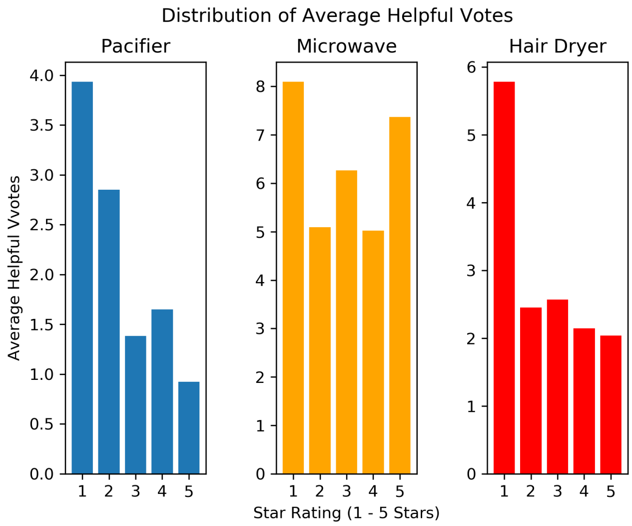

Figure 11: The result of 5.3

- This diagram largely validates the 5.2 model diagram. Because all the scores that had a rebound point, such as 1 star, 2 stars for the pacifier, 2 stars, 4 stars for the dryer, and 1 star for the hair dryer, all experienced an elastic increase and decrease. Because if you divide these ratings by the previous elastic limit, and omit the factors that make positive and negative ratings more likely to attract attention, the charts should all show a horizontal trend.

5.4 The Scatter Distribution of Average Comment Length and Average Score

- Based on the observation of the model, the distribution of pacifier and hair dryer has certain similarities, while the distribution of comment words in the dryer is relatively irregular. The relative irregularity of the Microwave distribution map is closely related to the small elastic limit presented in the model diagram above. Therefore, in the whole star-rating axis, the average comment length will show randomness, because the purchase intention of the reviewer is little affected by the subjective intention in the comments of the star-rating people.

- The comment length of Pacifier and hair dryer showed a regular and concentrated distribution trend. The average comment length for both is mainly in the low number segment, which is in line with our general expectations. But the former's ratings were concentrated at 4.5 stars, and the longest reviews were around 5 stars. The latter rating was concentrated at four stars, with the longest review appearing at four stars. Moreover, the length

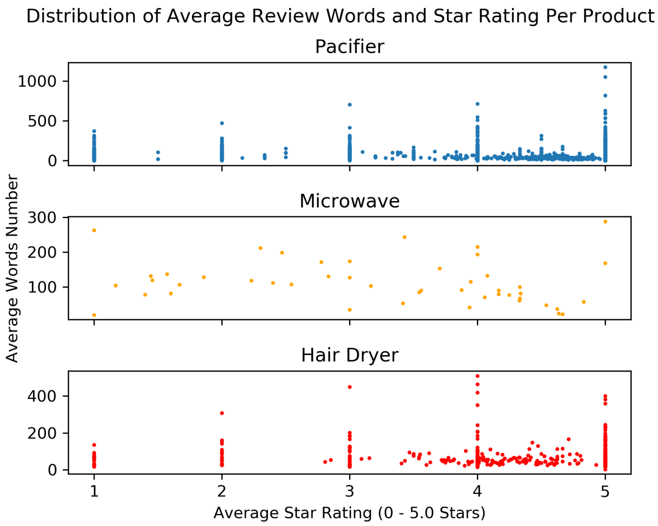

Figure 12: The result of 5.4

of pacifier's longest comment is much longer than that of hair dryer. That is to say, the comments made by the 5-star reviewer group of pacifier and the 4-star reviewer group of hair dryer are more realistic for the product decision-making.

6 Requirements

6.1 A+C Design Suitable Model for Predicting Sales Number Based on Review and Star Rating

- Under the common sense of judgement, we believe that the evaluation of the emotional value is positive, the higher the rating, the more sales will (assuming the linear growth), so we assume that a regression equation Y = W1 W2 * * X1 + X2 (1, 2, here is the subscript), we through the least square method to find the corresponding weights W1, W2, the reason for using multiple least-squares method, because we hope that by building on existing data sets is a linear model to fitting case data sets the relationship between the feature vector of each component. We expect reviews to be more heavily weighted, because reviews contain more information than ratings, so they should be a bigger determinant of sales than ratings. The following is the formula of multivariate least square method.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, i = 1, 2, \cdots, n$$

Figure 13: The result of 6.1

- Take the partial derivative of each parameter
- In order to get the parameters our goal is to minimize the sum of the squares of the errors
- The normal system can be obtained
- Let me write it in matrix form as
- So we get the solution of the parameter

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) = 0 \\ \dfrac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}) x_{ij} = 0 \end{cases}$$

Figure 14: The result of 6.1

$$\min_{\beta_0, \beta_1, \cdots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$
$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

Figure 15: The result of 6.1

$$\begin{cases} n\beta_0 + \sum x_{i1}\beta_1 + \cdots + \sum x_{ip}\beta_p = \sum y_i \\ \sum x_{i1} \cdot \beta_0 + \sum x_{i1}^2 \beta_1 + \cdots + \sum x_{i1} x_{ip} \beta_p = \sum x_{i1} y_i \\ \cdots\cdots\cdots\cdots \\ \sum x_{ip} \cdot \beta_0 + \sum x_{ip} x_{i1} \beta_1 + \cdots + \sum x_{ip}^2 \beta_p = \sum x_{ip} y_i \end{cases}$$

Figure 16: The result of 6.1

$$X'X\beta = X'Y$$

Figure 17: The result of 6.1

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Figure 18: The result of 6.1

- These are the conclusions we use to calculate the corresponding weight.
- Finally, for the three products, the weight pacifier we calculated was w1 = 1.2 w2 = 4.7, and hair dryer's w1 = 3.9 w2 = 30.0 dryer's w1 = 3.29 w2 = 82.31.

6.2 Calculate the Correlation Coefficient to Find the Relationship Between Star Rating and Review

- Correlation coefficient is a statistical index to reflect the close correlation between variables. The correlation coefficient is calculated by the product difference method, which is also based on the deviation between two variables and their respective mean value, and the correlation degree between two variables is reflected by multiplying two deviations. This paper focuses on the linear single-phase relations. In order to obtain the influence of previous stars on comments, we stagger the corresponding stars and comment emotion values one by one, artificially create the time series difference, and take star value as the independent variable and emotion value as the dependent variable into the formula to find the correlation coefficient

$$r = \frac{S^2_{xy}}{S_x S_y} = \frac{\sum (x-\bar{x})(y-\bar{y})/n}{\sqrt{\sum (x-\bar{x})^2/n} \cdot \sqrt{\sum (y-\bar{y})^2/n}}$$

$$= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Figure 19: The result of 6.2

- For pacifier, approximately 2/3 of the data reflected a linear relationship between star rating and rating, approximately 1/3 had a positive correlation, and 1/3 had a negative correlation.
- Almost all the data of hair dryer show that there is a linear relationship between the star rating and the rating, about 1/2 has a positive correlation and 1/2 has a negative correlation.
- The situation with the receiver and with the hair dryer was about the same. About 2/3 of the data reflect that there is a linear relationship between star rating and rating, about 1/3 has a positive correlation, and 1/3 has a negative correlation.

6.3 Calculate the Correlation Coefficient to Judge the Relationship between Key Words and Star Rating

- Through the analysis of NLP package TextBlob and word segmentation package NLTK, the words with the highest frequency and high positive and negative emotions were found, the words with strong emotional power were visualized through the word cloud, and the correlation coefficient was obtained to reflect the influence of comment keywords on star evaluation.

- For pacifier, the higher the rating, the more positive the review, the higher the rating, and the lower the rating, the negative correlation between the review sentiment and the rating.
- For hair dryer, in the product with higher rating, the comment emotion value was negatively correlated with the rating, while in the product with lower rating, there was a positive correlation between the comment emotion value and the rating.
- The situation is similar to that of pacifier, but the linear relationship is more obvious.
- The following is a partial selection of high-frequency emotional words

Figure 20: The result of 6.31

Figure 21: The result of 6.32

6.4 Reputation Forecast Based on Time Series Analysis

Firstly, we define that the reputation of a product is determined by its star rating, which contains less ambiguity as stated above and is more accurate. The next step is to find out how the star rating can be analyzed base on time series and construct a reliable model to help Sunshine company predict whether reputation is increasing or decreasing.

To grasp a better understanding of the relationship between reputation, i.e. star rating, and time, we analyzed the average star rating of each month from the oldest review respectively, as well as the latest average star rating of 14 days (2 weeks). The result is shown in Figure 22, where t denotes the time in months or days, and y value refers to the average star rating of the time.

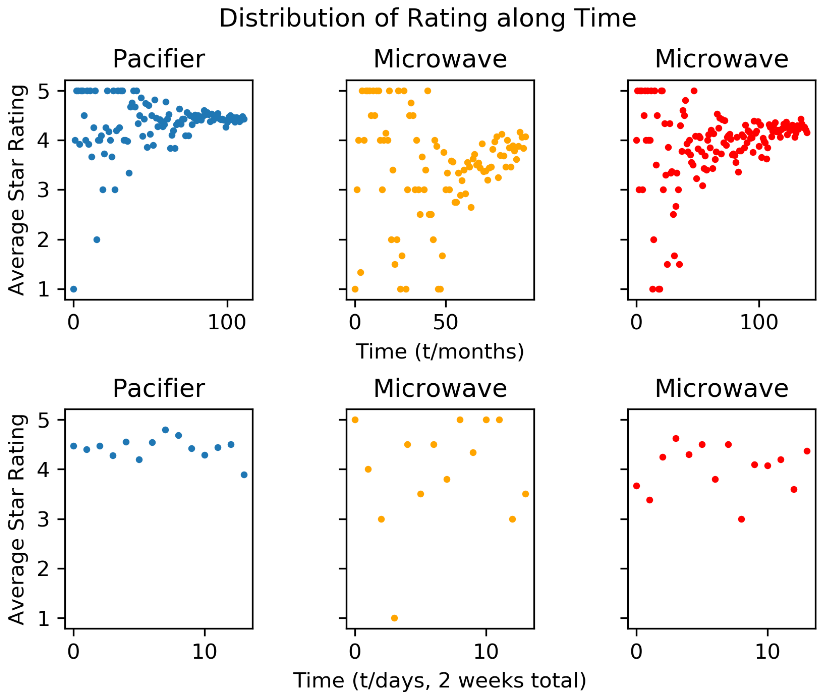

Figure 22: The result of analysis

From the data set, and as shown in the figure, we found that the average star rating has been becoming more stable, concentrated around 4 stars, which is reflected by the analysis above. However, in the early days, due to the limited amount of data, the rating fluctuated so much that it had no reference value. On the other hand, from the figures depicting latest star ratings of 14 days, we found that the reputation, or the star rating, is basically a stationary series, which means that we could use **Time Series Analysis** to forecast the trend of reputation increasing or decreasing.

6.4.1 Exponential Smoothing Method

Since the forecast only considers relatively short period of time, and given the fact that data is non-lieaner, we use the **Triple Exponential Smoothing** of the latest 2 weeks of rating data to help Sunshine company predict the reputation.

The equation of this methods is (2):

$$\begin{cases} S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \\ S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha) S_{t-1}^{(3)} \end{cases} \tag{2}$$

where α is the smoothing constant determine the weights for observations, as impact to the future degrades as time passes by. α varies from 0 to 1. y_t is the actual value of the period while S_{t-1} is the forecast for previous period. After third smoothing, we have the third smoothing value $S_t^{(3)}$.

The forecast moel of triple exponential smoothing is (3):

$$\hat{y}_{t+m} = a_t + b_t m + C_t m^2 \text{ , } m = 1, 2, \ldots \tag{3}$$

where (4):

$$\begin{cases} a_t = 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)} \\ b_t = \frac{\alpha}{2(1-\alpha)^2}[(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}] \\ c_t = \frac{\alpha^2}{2(1-\alpha)^2}[S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}] \end{cases} \tag{4}$$

6.4.2 Model Simulation and Analysis

Firstly, we need to determine the smoothing parameter α for the model with best performance. The performance of the model is determined by **Mean Absolute Percentage Error**, or **MAPE**, which can be calculated by (5):

$$MAPE = \frac{\sum_{i=1}^{n} \left(\frac{|\hat{y}_i - S_t|}{\hat{y}_i} \times 100 \right)}{n} \tag{5}$$

where n is the number of forecasts. After trying different α in analyzing the latest 14 days of average star rating of 3 types of products respectively, we have the following result Table 2:

Apparently, when α equals to 0.3, exponential smoothing gets best result with the almost lowest MAPE. After using $\alpha = 0.3$ as parameter, we use the triple exponential smoothing method to analyze the data and furthermore, predict the reputation in the future. The result is shown below Figure 23:

Now, we can use the equation 3 to predict the star rating in next few days. If rating gets lower, reputation is decreasing. On the contrary, star rating getting higher indicates that reputation of the product is increasing.

Here, we give a simple example. We forecast the data in next 2 days for pacifiers, using the data calculated above. The calculation 6 shows that the reputation will increase a little bit then falls into a period of decreasing.

Table 2: Mean Absolute Percentage Error

α	Pacifier	Microwave	Hair Dryer
0.1	3.06	33.57	7.78
0.2	2.23	23.58	6.98
0.3	1.49	24.10	5.29
0.5	2.28	33.75	6.10
0.8	6.58	74.58	20.03

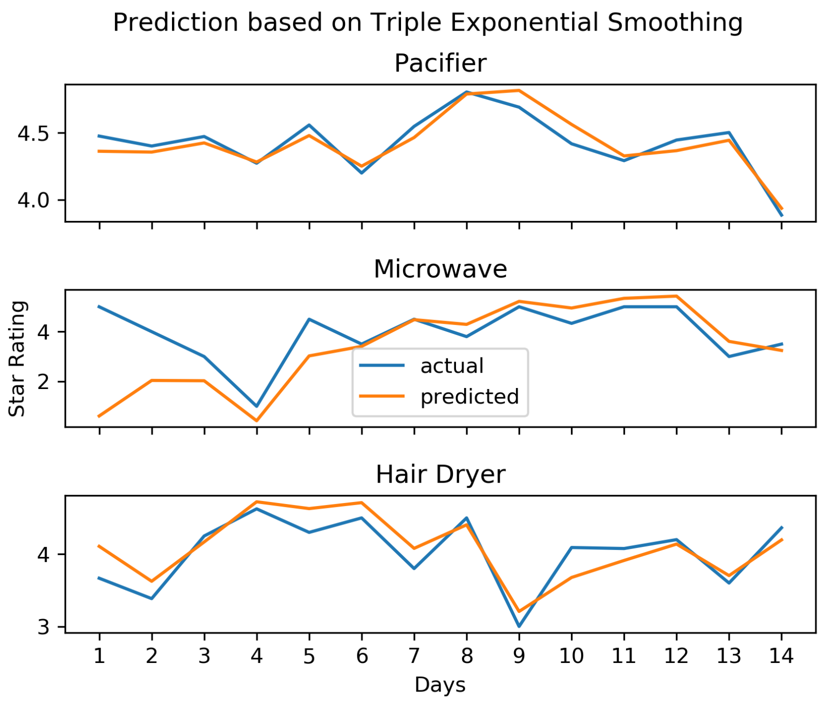

Figure 23: The result of Prediction

$$\begin{cases} y_t & = 3.89 \\ \hat{y}_{t+1} & = a_t + b_t \times 1 + C_t \times 1^2 = 3.94 \\ \hat{y}_{t+2} & = a_t + b_t \times 2 + C_t \times 2^2 = 3.79 \\ \hat{y}_{t+3} & = a_t + b_t \times 3 + C_t \times 3^2 = 3.65 \\ \ldots \end{cases} \tag{6}$$

7 Strengths and Weaknesses

7.1 Strengths

- We have designed a more scientific and helpful rating.
- According to the time series, the trend of product evaluation can be predicted accurately.
- It is more efficient to extract important keywords according to the way that natural language processing comes with its own package. The evaluation value is not a simple two number representing good or bad, but the value ranges from -1 to 1, which can better reflect an emotion. Trend and extent

7.2 Weaknesses

- we design a model which's reviews has bigger weight. In general, reviews are more informative, but they are not as intuitive as rstar atings, and they may be more ambiguous.
- The correlation coefficient can only show whether there is a linear relationship between the two variables. Even if the correlation coefficient is 0, it cannot be judged that there

is no relationship between the two variables. It can only indicate that the two variables have no linear relationship. There may be other variables between the two variables. relationship.

8 Conclusion

- In this paper, we make an analysis and prediction on these three products sales status. With the analysis, we have a suggestion to Sun shine company.
- We find that star ratings and reviews will affect sales, and reviews may have more impact, so we suggest that we should lay a good foundation for public opinion and quality control for all three products at first. We can first hand over the products to vines and let them evaluate the products. Based on their evaluation, we can better improve our products and lay a good basis for comments In order to make comments and ratings interact with each other, we found that the more positive the evaluation is, the higher the rating is, and the more sales volume can be promoted. The preliminary work is very important. With the passage of time, the evaluation of products will tend to ease. The preliminary work done well will also play a great role in the continuous sales of products.
- At last, we hope these suggestion will provide helps to you to sales these products.

Memorandum

To: Marketing Director
From: Team 2005922
Date: 03 10th, 2020
Subject: Report about the analysis and result of the products

Dear Sir/Madam:

It is our honor to help you analyze the data sets of three products. We are writing this letter to report our analysis and results.

Firstly, we did a data cleaning to pre-process the data before formally dive into the analysis job. An equation is created to calculate confidence of every single review based on votes helpfulness, star rating, purchase verification,and vine member, A basic analysis was carried on after data cleaning, in which we found some characteristics between star rating, reviews and helpfulness ratings that could help your products succeed. We found that a vast majority of review is concentrated on high star rating and low star rating, which also get top two amount of helpful votes. Yet, many of reviewers gives high star raintg while leaving very short review, which is not that helpful. Thus I would suggest that you keep your focus on the high star rating review as well as low star rating review, especially those who get a lot of helpful votes, find what is importan t to customers, what makes them hate the product or love the product, etc. Meantime, it is better if you could send some products to Amazon Vine members and invite them to test your products and publish their reviews, which helps advertise your product.

Secondly, Under the judgment of common sense, we think that the more positive the sentiment value is, the higher the rating, the more sales will be, so we assume a regression equation, and we use the multiple least square method to find the corresponding weight The reason to use the multivariate least squares method is because we want to fit a relationship between the components of the feature vector of the data set by constructing a linear model on the existing data set. We predict that the weight of reviews should be greater, because reviews contain more information than ratings, so the decisive factor for sales should be greater than the rating stars. The following is the formula for multivariate least squares. The correlation coefficient is a statistical index used to reflect the closeness of the correlation between variables. The correlation coefficient is calculated by the product difference method, which is also based on the deviation of the two variables and their respective averages. The degree of correlation between the two variables is reflected by multiplying the two deviations; the focus is on the linear single correlation coefficient. In order to obtain the influence of previous star ratings on reviews, we will stagger one-to-one corresponding star ratings and comment sentiment values with each other, artificially create time series differences, use star ratings as independent variables, and sentiment values as dependent variables, and bring them into the correlation In the formula of the coefficient, the correlation coefficient obtained can reflect the relationship between the two.

Additionally, we established a time series model based on the Triple Exponential Smoothing method, which you could use to forecast the trend of your product's reputation, given just a small amount of review data from the past 2 weeks. It must be useful to control the reputation and development of your product, for you will be informed ahead of the actual changes and get the opportunity to find the causes and stop the trend of decreasing or keep it increasing.

In order to make comments and ratings interact with each other, we found that the more positive the evaluation is, the higher the rating is, and the more sales volume can be promoted. The preliminary work is very important. With the passage of time, the evaluation of products will tend to ease. The preliminary work done well will also play a great role in the continuous sales of products.

At last, we hope these suggestion will provide helps to you to sales these products. Thank you!

Best wishes, Team 2005922

References

[1] Susan M. Mudambi, David Schuff, Zhewei Zhang. Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings *Hawaii International Conference on System Science*, 47th, 2014.

[2] NIE Guihua, ZHANG Bo. The Design and Implementation on Chinese E-commerce Semantic Lexicon ECNET. *International Conference on Computer Science and Software Engineering*, 2008.

[3] Paul Manuel, Omer Demirel, Rifat Gorener. Application of eCommerce for SMEs by using NLP principles. *IEEE*, 0-7803-8150-5/03, 2003.

[4] Shilin Shang, Hui Wang. Automatic User Preferences Acquirement in Chinese Commercial Web Sites With NLP and DM Techniques. *International Conference onon Asian Languages Processing*, 2009.

[5] Shilin Shang, Hui Wang. Automatic User Preferences Acquirement in Chinese Commercial Web Sites With NLP and DM Techniques. *International Conference onon Asian Languages Processing*, 2009.

[6] Chandrasekhar Rangu,Shuvojit Chatterjee,Srinivasa Rao Valluru. Text Mining Approach for Product Quality Enhancement (Improving Product Quality through Machine Learning) *International Advance Computing Conference*, 7th, IEEE 2017.

[7] *A Statistical Analysis of 1.2 Million Amazon Reviews*. (2014). Retrieved March 9, 2020, from `https://minimaxir.com/2014/06/reviewing-reviews/`

Appendices A: Basic Analysis Code

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import nltk
import math

# Libraries for text preprocessing
import re
# nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import RegexpTokenizer

# read data
df_pacifier = pd.read_csv('./pacifier.tsv', sep='\t')
df_microwave = pd.read_csv('./microwave.tsv', sep='\t')
df_hair_dryer = pd.read_csv('./hair_dryer.tsv', sep='\t')

# Data preprocessing
# 1. Delete irrelevant data, e.g. marketplace, product_category.
# 2. Delete empty data
# 3. Delete corrupt reviews, i.e. reviews that contain non-English characters.
def data_cleaning(df):
    df = df.drop(columns=['marketplace', 'product_category'])
    df = df.dropna(how='any')

    def isEnglish(s):
        try:
            s.encode(encoding='utf-8').decode('ascii')
        except UnicodeDecodeError:
            return False
        else:
            return True

    for index, row in df.iterrows():
        if isEnglish(row['review_body']) == False:
            df.drop(index, inplace=True)

    return df

# Clean all data sets
df_pacifier = data_cleaning(df_pacifier)
df_microwave = data_cleaning(df_microwave)
df_hair_dryer = data_cleaning(df_hair_dryer)

# Function that add confidence to each data set
def add_confidence(df):
    def calc_confidence(row):
        if row.verified_purchase == 'Y':
            verified = 1
        else:
            verified = 0.7

        if row.vine == 'Y':
            is_vine = 1
        else:
            is_vine = 0;
        return (1-math.exp(-(row.helpful_votes+1)))*(row.helpful_votes+1)/(row.total_vot

    col = df.apply(calc_confidence, axis=1) # get column data with an index
    df = df.assign(confidence=col.values) # assign values to column 'confidence'
    return df

# Add confidence to each data set
```

```
df_pacifier = add_confidence(df_pacifier)
df_microwave = add_confidence(df_microwave)
df_hair_dryer = add_confidence(df_hair_dryer)

# Cleanup data with confidence <= 0.5
def delete_low_confidence_data(df):
    df.drop(df[df['confidence'] <= 0.5].index, inplace=True)
    return df

df_pacifier = delete_low_confidence_data(df_pacifier)
df_microwave = delete_low_confidence_data(df_microwave)
df_hair_dryer = delete_low_confidence_data(df_hair_dryer)

# Basic statistics - distribution of star rating
def count_products(df):
    return df['product_parent'].nunique()

def count_rating(df):
    return df.groupby('star_rating').size()

pacifier_count = count_products(df_pacifier)
microwave_count = count_products(df_microwave)
hair_dryer_count = count_products(df_hair_dryer)

pacifier_rating_distribution = count_rating(df_pacifier)
microwave_rating_distribution = count_rating(df_microwave)
hair_dryer_rating_distribution = count_rating(df_hair_dryer)
```

Appendices B: Time Series Analysis Code

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math

# read data
df_pacifier = pd.read_csv('./pacifier.tsv', sep='\t')
df_microwave = pd.read_csv('./microwave.tsv', sep='\t')
df_hair_dryer = pd.read_csv('./hair_dryer.tsv', sep='\t')

# Exponential smoothing
def triple_exponential_smoothing(alpha, st):
    st1 = np.zeros(st.shape)
    st2 = np.zeros(st.shape)
    st3 = np.zeros(st.shape)
    st1[0] = np.mean(st[0:3])
    st2[0] = st[0]
    st3[0] = st[0]
    for i in range(1, len(st1)):
        st1[i] = alpha*st[i]+(1-alpha)*st1[i-1]
        st2[i] = alpha*st1[i]+(1-alpha)*st2[i-1]
        st3[i] = alpha*st2[i]+(1-alpha)*st3[i-1]
    return st1, st2, st3

def predict_rating(data, alpha = 0.30):
    # weight alpha
    # data
    st = data.tail(14).values
    # triple exponential smoothing
    st1, st2, st3 = triple_exponential_smoothing(alpha, st)
    # prediction
    a = 3 * st1 - 3 * st2 + st3
    b = 0.5*alpha / (1-alpha)**2 * ((6-5*alpha)*st1 - 2*(5-4*alpha)*st2 + (4-3*alpha)*st
    c = 0.5*alpha**2 / 2*(1-alpha)**2 * (st1-2*st2+st3)
    s_predict = np.zeros(st.shape)
    s_predict = a + b + c
    temp = a[-1] + b[-1] *3 + c[-1] *9
    print(s_predict[-1])
```

```
    print(temp)
    return st, s_predict

st_pacifier, s_predict_pacifier = predict_rating(pacifier_daily_rating)
st_microwaver, s_predict_microwaver = predict_rating(microwave_daily_rating)
st_hair_dryer, s_predict_hair_dryer = predict_rating(hair_dryer_daily_rating)
```