Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2015224

A Wealth of Data

Summary

In the period of big data, data has become an important tool for business analysis. In this paper, we analyze and evaluate three kinds of products, and give some suggestions on business strategy to Sunshine Company based on the data of customer-supplied ratings and reviews.

For requirement 1, we clean the data first to get rid of the useless data. For example, we remove samples that are not relevant to a given product from the data. Then, we quantified reviews with sentiment analysis algorithm to extract the emotional tendency of reviews as the sentiment value. Then we do word frequency analysis of the reviews to find the keywords which reflect what people concern. We find that people are concerned about the temperature of the wind from the hairdryer, the size of the microwave oven and the pacifier. Furthermore, we draw star rating bar charts for descriptive statistics. We discover the pacifier obtaining the widespread high praise, but the microwave has a relatively negative rating.

For requirement 2(e). First, we use TF-IDF algorithm to further quantify review text. Then, according to reality, we classify different star ratings into 0-1 variables. Finally, we do logistic regression on the star rating and quantified review text, and we find that the emotion of people strongly associates with rating levels. Besides, we also select some advantages and disadvantages of the product. For example, a cute pacifier has more positive reviews, but some hairdryers have the disadvantage of being overheated, and a microwave oven has the disadvantage of being flimsy.

For requirement 2(a)(c), we set up an evaluation function including numerical variables and quantified text variables to evaluate the products. Based on this scoring criterion, we select several sub-products with the highest comprehensive scores of each product.

For requirement 2(b)(d), based on the scoring system we develop, we draw the time series graph of product reputation. The reputations of all three products are generally smooth, except that they fluctuate widely in the early time because of few reviews. Finally, based on the autoregressive distributed lag model, we carried out the Pearson correlation analysis on the sentiment values and star ratings in the lagged time series. It is found that the previous review has little effect on the present star rating.

For requirement 3, we offer some online sales strategies about the main market and main direction. And potential design features are also mentioned in the letter.

Keywords: Sentiment Analize; Logistic Regression; Statistical Comprehensive Evaluation; TF-IDF; Review Text-mining

Contents

1 Introduction **2**

2 Descriptive statistics and Data preprocessing **2**

2.1 Data preprocessing 2

2.1.1 Removal of meaningless data 2

2.1.2 Quantification of text data based on NLTK 2

2.2 Descriptive statistics 3

3 The relationship between star rating and review text based on Logistic Regreesion **7**

3.1 Quantification of review text based on TF-IDF Method 7

3.2 Logistic Regression 8

3.3 The result of Regression 9

4 Establishment of product evaluation function **10**

4.1 Integration of variables 10

4.2 The result of evaluation 12

5 The relationship between product reputation and time **12**

6 The relationship between the emotional tendency of reviews and historical star ratings **14**

7 Strengths and weaknesses **16**

7.1 Strengths 16

7.2 Weaknesses 16

8 The letter to Sunshine company **16**

Appendices **18**

Appendix A Data clean code **18**

Appendix B Modeling code **19**

1 Introduction

In the period of Internet, onlineshoping has become the main way for people to shop. The data of the online shopping platform has become an important basis for the online store to make business decisions.Sunshine Company is planning to introduce and sell three new products in the online marketplace: a microwave oven, a baby pacifier, and a hair dryer.Sunshine company has some online shopping data of customers based on time. Our team will analyze these data and solve the following problems:

i:Analyze the three product data set, with describing the meaningful variables and relationships among the star rating, review, and helpful votes.

ii:Identify data measures based on ratings and reviews that are most informative for Sunshine Company to track, once their three products are placed on sale in the online marketplace.

iii:Identify and discuss time-based measures and patterns within each data set that might suggest that a products reputation is increasing or decreasing in the online marketplace.

iv:Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product.

vi:Write a one- to two-page letter to the Marketing Director of Sunshine Company summarizing your teams analysis and results.

2 Descriptive statistics and Data preprocessing

2.1 Data preprocessing

For data preprocessing on the data provided by Sunshine company, our team mainly considers two aspects. The first is the removal of useless data, and the second is the quantification of text data.

2.1.1 Removal of meaningless data

By observing the data given in the question, we find that there is a lot of irrelevant data. For example, there are a lot of comments on non-pacifier items in the pacifier data, and we delete those comments. We also found that there were a lot of user comments that were meaningless, and we culled them out.

2.1.2 Quantification of text data based on NLTK

Comment data belongs to text type, which is not conducive to data analysis. So, we need to quantify the review data.The quantitative processing of this paper is based on Natural Language Toolkit algorithm.The purpose of the algorithm is to represent the positive or negative emotions and their degree in comments with a number in (-1,1).If the sentiment value is larger than zero,it means a positive sense,if the sentiment value is smaller than zero,it means negative ,if the sentiment value equal to zero,it means netrual sense.The absolute value is more large ,the greater the emotional tendency.The detail algorithm is as follows:

i.Read the comment data and clause the comment.

ii.Search for the sentiment words of clauses, record their position and whether it is positive or negative.

iii.Search for degree words near the sentiment words, set the weight for the degree words, times the emotional value and multiply the sentiment value by the weight.

iv.Search for negative words in front of sentiment word in the same clause. Count the number of negative words, and time -1 when the number is odd or 1 when the number is even.

vi.The corresponding sentiment value pluses a suitable value if there is an exclamation point at the end of one clause.

vii.Calculate the sentiment value of all clauses in a comment and the sentiment value of every comment.

The detail codes are in appendix. The sentiment values of some reviews of the three products are shown in the following table

Table 1: sentiments values tablek

Product	review	sentiments value
hairdrier	Awesome dryer, I have been looking for...	0.784
hairdrier	Only used it for 2 weeks. ...	-0.5972
microwave	What a great unit!	0.9365
microwave	This product was a terrible disappointment.	-0.8074
praciflier	terrible, leaks like crazy..	-0.5574
praciflier	Love them!!. My son loves these!	0.9771

2.2 Descriptive statistics

First, we make descriptive statistics on the comment data and make data visualization.Data processing and cleaning is an important step before starting text mining. In this step, we remove punctuation, stop words, etc., to make the comments as uniform as possible. Once that's done, you can check the most commonly used words in the data. So, let's define a function here that shows the n most common words in the data on a bar chart.Next, to further remove the noise from the text, we can use the word form reduction from the spacy library. It can restore words to their original form and reduce the repetition of words.

From the remaining vocabulary,the 30 words with high frequency for each product are shown in the figure below.High frequency words prove that the user is concerned about a certain performance of the product or that the product has a certain feature, which is worthy of the company's attention in product improvement

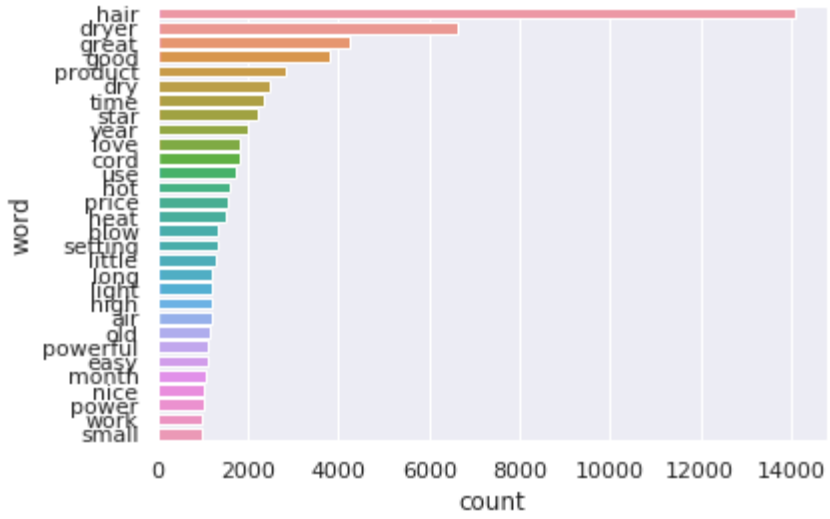


Figure 1: the top 30 words in the reviews of hairdryer

According to the image above," there are some keywords like "good","love","powerful","hot" and "heat" in the reviews.It can be seen that most users are approved of the hair dryer.Most hair dryers is powerful,but some hairdryers may be too hot to scald the users.The company should conduct further research on the temperature control of the hair dryers.

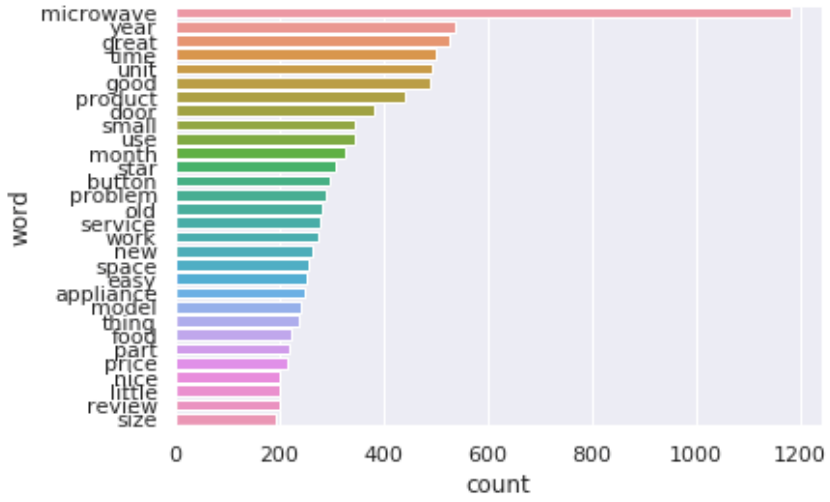


Figure 2: the top 30 words in the reviews of microwave

According to the figure 2,users are alsolikely to praise the microwave.However,there are a lot of words about the size and color of microwave oven in the keywords.Although it is impossible to tell from word frequency whether these features are positive or negative,We can find that customers are very concerned about the appearance of the microwave.It's worth the company's attention.

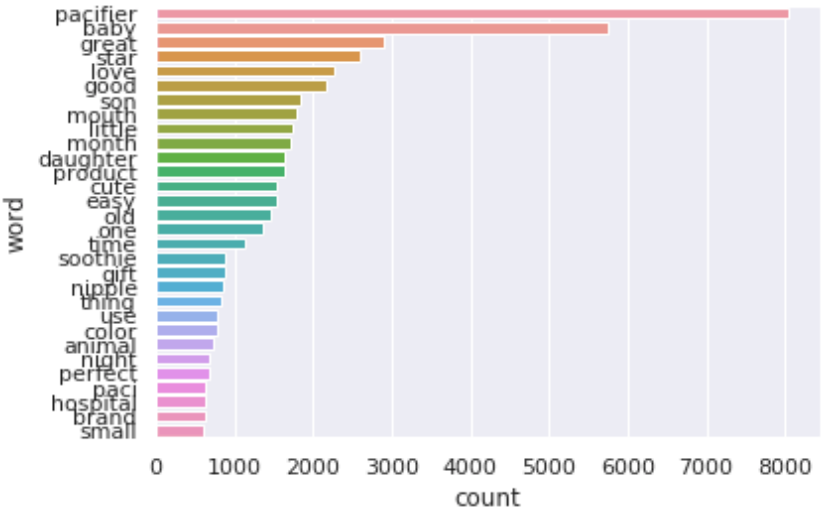


Figure 3: the top 30 words in the reviews of pacifier

According to figure 3,The positive words in the comment keywords of the pacifier were higher than the other two products.Customers are also concerned about the cuteness and color of the pacifiers.

Then,we discribe the important variable star rating combing with the sentiment value.For each product we divided it into positive and negative groups based on the review attitude.If the sentiment value is larger than 0,we consider it belongs to positive group,else it belongs to negative group. For each group of samples, we drew the bar graph as follows according to the star rating.

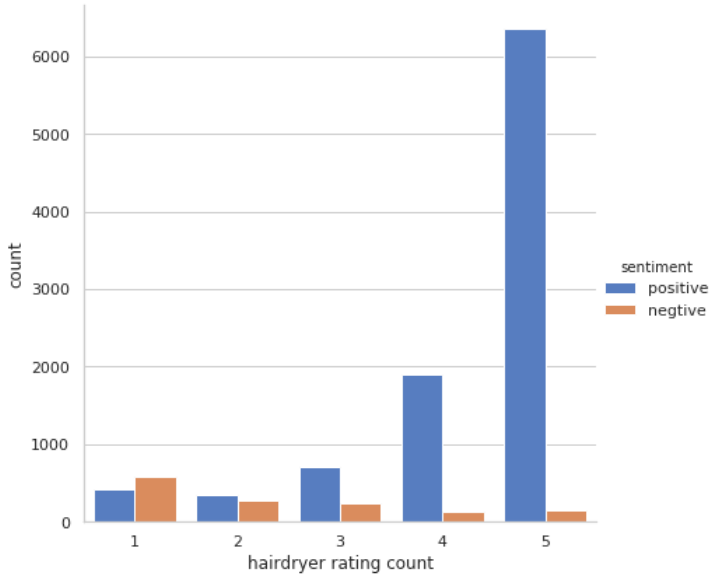


Figure 4: hairdryer star rating bar

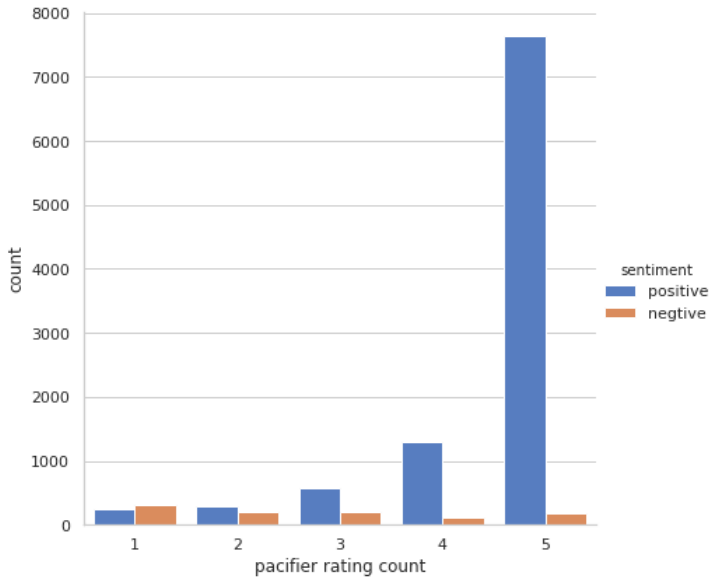


Figure 5: pacifier star rating bar

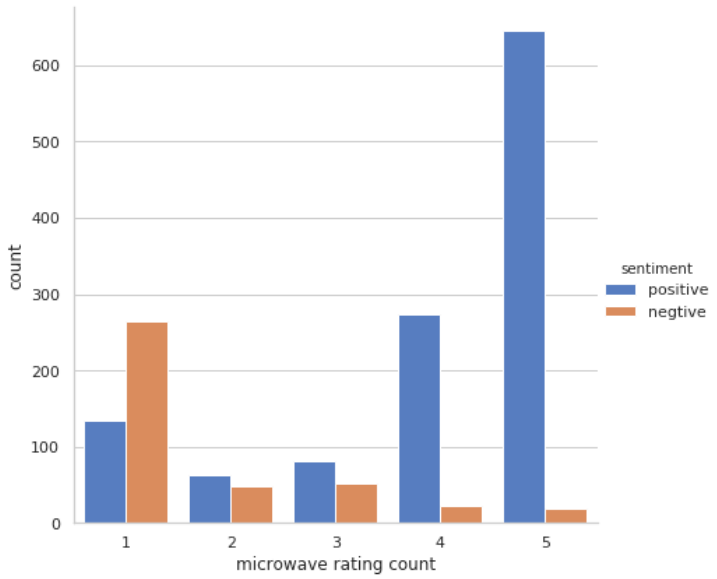


Figure 6: microwave star rating bar

According to the figure above, we found that five stars accounted for the majority of the stars in the three products. Among them,the proportion of five stars in pacifier is highest . We can consider that the pacifier as the most popular of the three products.In addition, the positive group had the highest proportion of five stars and the negative group had the highest proportion of one star.This rule applies to three products.To some extent,this also tests the accuracy of our emotional analysis of the text.

3 The relationship between star rating and review text based on Logistic Regreesion

3.1 Quantification of review text based on TF-IDF Method

Now, we want to research the relationship between reviews and star rating. Generally speaking, the more positive the comment is, the higher the star rating will be, and the lower the star rating is, the more negative the comment will be. On the one hand, we verify whether this conclusion is valid. On the other hand, we want to understand the features that people who give good reviews are more interested in and what features that people who give bad reviews complain about most. In order to achieve this purpose, we use TF-IDF technology to quantize review according to its importance in all reviews. Then, deal with the data of star rating. We denote 'one star','two star' , 'three star' as 0, and others as 1. Thus we gain a series data of 0-1 from star rating data. After that, we get some words that have a great influence on consumers' attitude through logistic regression between the above 0-1 series and those quantified review. Thereupon, some product features consumers interested in most can be concluded, which Sunshine Company need to track once their three products are placed on sale in the online marketplace.

Now, we want to research the relationship between reviews and star rating. Generally speaking, the more positive the comment is, the higher the star rating will be, and the lower the star rating is, the more negative the comment will be. On the one hand, we verify whether this conclusion is valid. On the other hand, we want to understand the features that people who give good reviews are more interested in and what features that people who give bad reviews complain about most. In order to achieve this purpose, we use TF-IDF technology to quantize review according to its importance in all reviews. Then, deal with the data of star rating. We denote 'one star','two star' , 'three star' as 0, and others as 1. Thus we gain a series data of 0-1 from star rating data. After that, we get some words that have a great influence on consumers' attitude through logistic regression between the above 0-1 series and those quantified review. Thereupon, some product features consumers interested in most can be concluded, which Sunshine Company need to track once their three products are placed on sale in the online marketplace.

TF-IDF is a statistical method to reflect the how important a word is to a document in a collection or corpus.The following is TF-IDF formula:

$$
\begin{aligned}
tf - idf(t,d) &= tf(t,d) \times idf(t) \\
tf(t,d) &= \frac{n_(t,d)}{\sum_k n_{k,d}} \\
idf(t) &= log\frac{n_d}{df(d,t)} + 1
\end{aligned}
$$

$tf(t,d)$ is tf value, indicating the frequency fo term t in a text d.
$idf(t)$ is the calculation formula of idf value of term t; n_d is the number of text in training set; $df(d,t)$ is the total number of documents containing term t. The idf value is an improvement of word weight, not only considering the frequency of words in the text, but also the frequency of words in the general text. To avoid 0 occurring in the denominator, the formula of idf we

actually use is

$$idf(t) = log(\frac{1+n_d}{1+df(d,t)}) + 1$$

TF-IDF is the product of tf and idf in fact and increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. So we can apply it in quantizing the text data.

3.2 Logistic Regression

Logistic regression is a classification model,whose principle is as follows.

For a given training data set,$T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$.Among them,$x_i \in R^n$,$y_i \in \{0, 1\}$Assume $z = -(wx + b)$.Then we call the following probability distribution logistic regression model:

$$P(Y = 0|x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{wx+b}}$$
$$P(Y = 1|x) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} = \frac{e^{wx+b}}{1+e^{wx+b}}$$

In the equation above ,we call the w the weight coefficient.To facilitate the representation of multiple variables,let's introduce the weight coefficent vector$W = (w_1, w_2, ..., w_n, b)'$.The sample vector $X = (x1, x2, ..., x_n, 1)'$.Now,the matrix of logistic regression is represented as follows.

$$P(Y = 0|X) = \frac{1}{1+e^{WX}}$$
$$P(Y = 1|X) = \frac{e^{WX}}{1+e^{WX}}$$

In this model, we need to estimate the weight coefficient vector W.The method of estimation is maximum likelihood estimation(MLE).

We assume:

$$P(Y = 1|x) = \pi(x)$$
$$P(Y = 0|x) = 1 - \pi(x)$$

The logarithmic likelihood function

$$\begin{aligned} L(W) &= \sum_{i=1}^{n}(y_i log(\pi(x_i)) + (1 - y_i)log(1 - \pi(x_i))) \\ &= \sum_{i=1}^{n}(y_i log(\frac{\pi(x_i)}{1-\pi(x_i)}) + log(1 - \pi(x_i))) \\ &= \sum_{i=1}^{n}(y_i(w_i x_i - log(1 + exp(w_i * x_i)))) \end{aligned}$$

Let $\frac{\partial L}{\partial w_i} = 0$,We can get the maximum of the likelihood function.The vector $\widehat{W}$ that we get from that is the maximum likelihood estimation of the weight parameter.According to statistical

study, maximum likelihood estimation has good statistical properties.So we now have a logistic regression estimation model as follows:

$$P(Y=0|X) = \frac{1}{1+e^{\widehat{W}X}}$$

$$P(Y=1|X) = \frac{e^{\widehat{W}X}}{1+e^{\widehat{W}X}}$$

3.3 The result of Regression

Previously, we have introduced the TF IDF algorithm to deal with the word frequency of text. We applied this algorithm to the comments of three products and took the data obtained as the independent variable of logistic regression. For example, in the hair dryer product, we extract words such as " quiet,spark' in user comments as independent variables, and the value of independent variables is obtained by the TF IDF algorithm above.As for the treatment of dependent variables ,we denote 'one star','two star' , 'three star' as 0, and others as 1.

The training sets of the obtained independent variables and dependent variables are carried out logistic regression.We used test sets to verify the accuracy of the model. The overall regression results of the three products and their accuracy are shown in the following table.

Table 2: Logistic regression accuracy results

	features	train records	test records	Model Accuracy
hairdryer	156971	8602	2868	0.880753138
microwave	51778	1211	404	0.841584158
pacifier	251940	14204	4735	0.855543823

According to the above table, the accuracy of logistic regression estimation of the three products is above 80%, so it can be considered that the results of logistic regression estimation are statistically satisfactory.

Next, we will show the coefficient estimation obtained by logistic regression and its specific meaning.Since there are too many independent variables in logistic regression to show them all, we only show some representative variables in the following table.

Table 3: The coeffient of some variables

hairdryer	positive	heat setting	hair fast	lightweight	quiet
		3.697376	2.375626	1.794518	1.268872
	nagetive	spark	hot	heavy	loud
		-2.4839	-2.29064	-1.9728	-1.8672
microwave	positive	easy	space	price	easy clean
		2.4074	1.83266	0.8172	0.6544
	nagetive	warranty	repair	quality	whirlpool
		-1.967804	-1.2249	-1.0242	-1.4451
pacifier	positive	gift	soft	cute	sturdy
		3.0069	2.3177	2.0372	1.7048
	nagetive	price	poor quality	flimsy	pink
		-4.4512	-2.5574	-2.3267	-2.0981

The specific meaning of the coefficient of the variable obtained by logistic regression is the influence of the word frequency of the variable appearing in the comment on the star rating.If the variable is a more positive term, the coefficient is positive and larger,if the variable is a more negative term, the coefficient is negative and smaller.According to this principle, combined with the results obtained in the above table, we can find the following rules.

People generally like the heat setting in the hair dryer, and some hair dryers has the advantages of fast, light, small sound and so on. Among them, heat setting is the most attractive.But some brands of hair dryer and there will be sparks, too hot, too heavy, too loud and other shortcomings.We can see that customers are very concerned about the temperature and lightness of the hair dryer.

Customers generally concern about the advantage of microwave oven,which is simple, occupying small space , clean and the low price.But the whirlpool microwave got a lot of bad reviews, and users hated the constant need for repairs and worried about its quality.

In the eyes of customers, the pacifier is a perfect gift, soft but tough, and very cute.However, some customers think that the price of pacifier is too high, some of the quality of the pacifier is very poor.Even some customers hate pink pacifiers.

4 Establishment of product evaluation function

According to the data provided by the title, we select the useful variables "star rating", "help votes", "total votes", "vine", "verify purchase", "review headline", "review body" to construct the evaluation system of the product.The two text variables "review headline" and "review body" have been processed into quantized sentiment values above.According to the above analysis of emotional value, it can be known that emotional value is a positive indicator, which means that the greater the emotional value, the higher the utility of the product.Now, we're going to deal with the few remaining numerical variables.

4.1 Integration of variables

Now let's deal with the variables "help votes", "total votes" and "vine". We assume varialbe HV as the number of "help votes",variable TV as the number of "total votes". The basic idea behind our approach to these two variables is to construct a function $VOTE = g(TV, HV)$ to fit the value of the comment.One commonly used indicator is the effective review rate, that is $\frac{HV}{TV} \times 100\%$.But there are drawbacks.When $TV = HV = 0$,this indicator is meaningless.Considering the practical implications, the more total votes there are, the worthier if two comment has the same valid comment rate $\frac{HV}{TV} \times 100\%$.So we need to fix the effective comment rate. According to the statistical method, we can smooth the effective comment rate.The function we built is as follows:

$$VOTE = g(TV, HV) = \frac{HV + \alpha}{TV + \beta}$$

Based on empirical research, we finally determined $\alpha =$ According to the review of some literature, at the same time, the data for many experiments.Finally we made $\alpha = 1$,$\beta = 2$.In this case,$g(TV = 10, HV = 10) = \frac{11}{12}$,$g(TV = 5, HV = 5) = \frac{6}{7}$. Although the effective comment rate is the same in both cases, the 10 "total votes" represent more attention and indicate that the comment is more reliable. In our model ,$g(10) > g(5)$,To some extent, the rationality

of the fitting function is explained.

According to the knowledge of statistics, only the standardized variables have the meaning of phase addition and subtraction. So we normalize the affective value and star rating.The normalization formula is as follows

$$x_{std} = \frac{x - \overline{x}}{s}$$

Inside this,$\overline{x}$ means the average of samples,s means the standard deviation of the samples. We call the standardized variable "sentiment value" as SV_{std},the standardized variable"star rating" as SR_{std}.

Now let's set up a function to evaluate each product.Here, we draw lessons from the principles and methods of statistical comprehensive evaluation commonly used in statistics.Building a function so that all of the sample values fall into (0,100).It's like making a rating system.The independent variables of our evaluation function include $VOTE, SV_{std}, SR_{std}, "VINE"$," PURCHASE".Using the statistical scoring method .Let's build a score first,for each sample:

$$\begin{aligned} SCORE_i = VOTE \times (\gamma(SV_{std}^i(1 + h(VINE_i) \times \theta_1 + z(PURCHASE_i) \times \theta_2)) \\ +(1-\gamma)(SR_{std}^i(1 + h(VINE_i) \times \theta_3 + z(PURCHASE_i) \times \theta_4))) \end{aligned}$$

inside the founction ,

$$h(VINE) = \begin{cases} 1, VINE = 1 \\ 0, VINE = 0 \end{cases}$$

$$z(PURCHASE) = \begin{cases} 1, PURCHASE = 1 \\ 0, PURCHASE = 0 \end{cases}$$

γ is the weight of sentiment value which means the contribution to score of sentiment value.$1-\gamma$ is the weight of star rating,which means the contribution to score of sentiment value.θ_1 means the contribution to sentiment value of vine.θ_2 means the contribution to sentiment value of the vertified purchaser.θ_3 means the contribution to standarized star rating of vine.θ_4 means the contribution to standarized star rating of vertified purchaser.According to the literature review, we let $\theta_1 = \theta_3 = 0.25$,$\theta_2 = \theta_4 = 0.2$

We use the above function to calculate all filtered data.Finally, we need to do a hundred differentiation for all scores.The process of differentiation is as follows: First step, we calculate and work out the maximum and minimum values of the score.We denote the maximum score $SCORE_{max}$,the minmum score $SCORE_{min}$.Then,We calculate the standardized score by the following formula.We denote the final scoring function as f:

$$f_i = \frac{SOCORE_i - SCORE_{min}}{SOCORE_{max} - SCORE_{min}} \times 100$$

After normalization, the product with the highest score will have a score of 100, while the product with the lowest score will have a score of 0.The higher the score, the better the product quality.

4.2 The result of evaluation

We categorize the scores by product category to look for some kinds of product with good or bad reputation . It's a good choice for Sunshine Company to pay more attention on those products with high reputation that are more likely to be hot-selling. Inherit high-reputation products' good and popular features and avoid similar defects of low-reputation products. The top five and bottom five products respectively in hair dryer, microwave, pacifier are as follows.

Table 4: The score of The top five and bottom five products

product	score	product_title
pacifier	90.38746946	nuk 8 count silicone pacifier
	90.36089667	2 mam ulti trends baby pacifiers for age 6+ months
	90.23408989	#1/16 serta sheep baby with pacifier
	90.15993999	mam crystal print silicone pacifier - green (6+ months)
	89.84272737	nuk 1 pack genius silicone bpa free pacifier, size 2, 6+ months
hair dryer	92.25517376	turbo power twin turbo 2600 professional hair dryer
	87.98337449	t3 featherweight hair dryer tourmaline ionic bespoke labs 83808
	87.47621602	babyliss pro babfrv2 volare ferrari designed professional luxury ...
	87.33843455	remington d3190a damage control ceramic hair dryer,...
	86.31218888	turbo power 1500 hair dryer
microwave	78.63851266	sharp r1874t 1.1 cu. ft. stainless steel over-the-range microwave
	77.61202409	whirlpool stainless look countertop microwave, 0.5 cu. feet, wmc20005yd
	76.91473121	pem31dmww%2d profile spacemaker ii%2dcountertop microwave
	76.36588093	sharp kb6524p 24 microwave drawer oven ...
	75.11139959	jx7227sfss - deluxe built-in trim kit for 2.2 microwave ovens...

The above products generally received high praise. Reviews on them are highly positive and star rating is almost five stars. All consumers who place orders confirm their purchase and have a good experience on using them. However, those products are purchased in small quantities and consumers pay less attention on them. Generally speaking, the above products are high-quality, high-potential and it's a wise choince to make further efforts on them. Sunshine Company can learn more from them, at the same time , promote own product, enhance service quality and even adopt price promotion to drop more attention on own products. Compare microwaves, hair dryers and pacifiers together, and it can be seen that pacifiers market has the best reputation with fierce competition; microwave market has the worst reputation. Hence Sunshine Company should carefully consider which market to focus on.

These products with low scores are products with low reputation and likely to be of poor quality and service. Therefore, it is of great helpness for Sunshine Company to aviod the low-score products' sales strategy and product features.

5 The relationship between product reputation and time

Then, sort the reviews in order of time to study whether a product's reputation is increasing or decreasing in the online marketplace. The following is the time series graph of hairdryer scores.

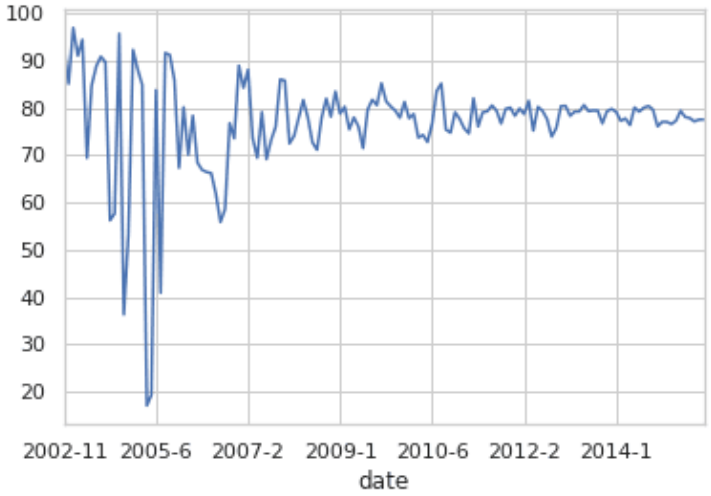
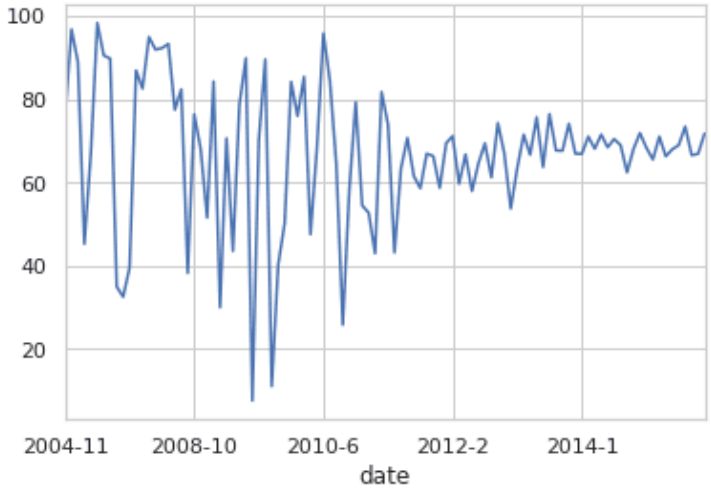


Figure 7: score of hairdrers



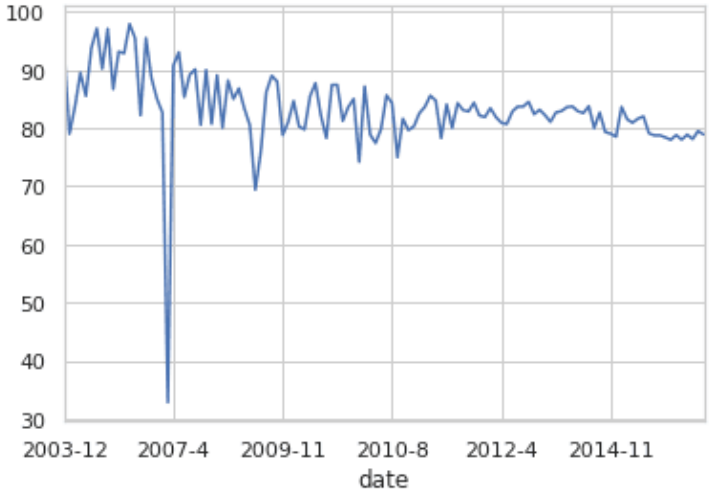


Figure 8: time series

As a whole, the score of the hairdryer fluctuates but generally stable. From November 2002 to February 2007, hairdryer scores fluctuated significantly and smooth between February 2007 and August 2015.The average score of hairdryer in March 2007 is a abnormal number. There were two reviews in March 2007, with scores of 28 and 46. It occurred for the reason that in the early years, the Internet was not popular and people seldom chose to shop online. Therefore, we mainly analyze the change of pacifier'reputation by its scores in the next few years. In general, the reputation of hair dryers is smooth. The following is the time series graph of microwave scores and pacifier scores. It can be seen that the reputation of microwave is slightly increasing and the pacifier's reputation is slightly decreasing recently.

6 The relationship between the emotional tendency of reviews and historical star ratings

Now, our team is ready to reaserch the relationship between the emotional tendency of customer reviews and historical star ratings.In statistical terms,that is to study whether there is a autocorrelation of customer's emotional tendency in time series.We use the methods of visual contrast and statistical test to study this problem.

First, we selected the product with the highest score among the pacifiers, drew a time series graph of its star rating and sentiment value, and placed the two time series graphs on the same coordinate graph as following:

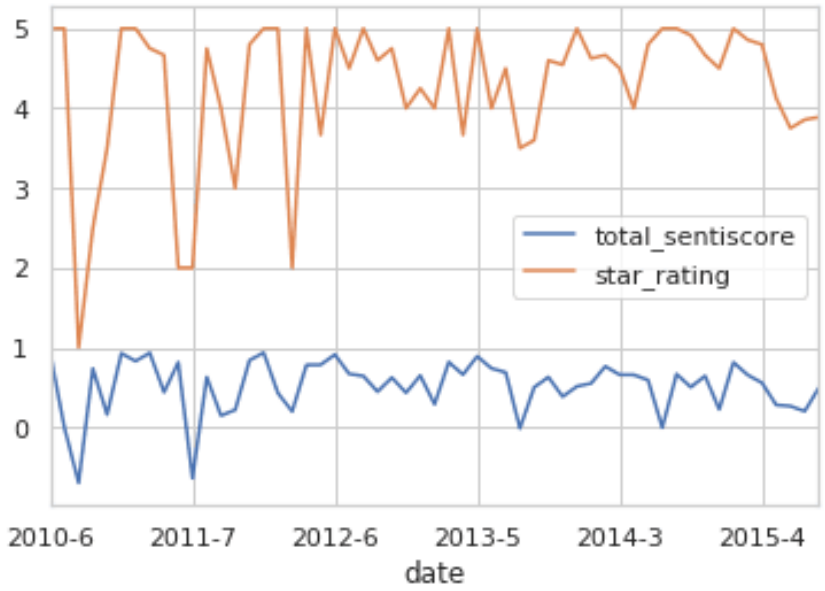


Figure 9: Time series graph of star rating and sentiment value

According to figure 7, we found that the trend of emotion value and star value is roughly the same.This shows that the correlation between star rating and emotion value is high at the same time.But if we stagger the time by one unit, we can see that the correlation between stars and emotion value is greatly reduced.

Next, we use more rigorous mathematical statistical methods to test the hypothesis.The test method we use is Pearson correlation test.We do two tests.One is the correlation test of star rating and sentiment values in the condition of time - to - time .The others is the correlation

test of star rating and sentiment values in the condition of hysteretic time series.The specific treatment method is as follows:first we calculated the monthly average sentiment value and the average star rating.We denote the monthly average sentiment value vector as X,denote the monthly average star rating vector as Y.Now,we need to find the correlation of X_t and Y_t, and the correlation of X_{t-1} and Y_t.

We calculate the correlation coefficients as follows:

$$\rho_1 = Corr(X_t, Y_t) = \frac{\sum(X_t - \overline{X})(Y_t - \overline{Y})}{\sqrt{\sum(X_t - \overline{X})^2 \sum(Y_t - \overline{Y})^2}}$$

$$\rho_2 = Corr(X_{t-1}, Y_t) = \frac{\sum(X_{t-1} - \overline{X})(Y_t - \overline{Y})}{\sqrt{\sum(X_{t-1} - \overline{X})^2 \sum(Y_t - \overline{Y})^2}}$$

We not only have to figure out the correlation coefficient, but we have to check whether it's zero.The hypothesis testing process is as follows:

First,present test statistics:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

Second,build and calculate test statistics:

$$T = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \sim t(n-2)$$

n is the number of samples.

Thrid,calculate the rejection region or P value.Make a decision whether or not to accept the null hypothesis in a certain level of significance.

The results of hypothesis testing are shown in the following table:

Table 5: Correlation and P value

	variable	corr	p-value
hairdryer	(X_t, Y_t)	0.6872	0.0000
	(X_{t-1}, Y_t)	0.1804	0.3358
pacifier	(X_t, Y_t)	0.3341	0.0002
	(X_{t-1}, Y_t)	-0.0188	0.8409
microwave	(X_t, Y_t)	0.6183	0.0000
	(X_{t-1}, Y_t)	-0.0281	0.7828

According to the table 4 ,we can find that sentiment value and star rating are highly correlated time to time,but almost irrelevant in the condition of hysteretic time.Assume the significance level $\alpha = 0.05$,the correlation of The sentimental tendency of historical star rating and current evaluation is statistically nonsignificant.

In fact,people tend to be rational when they shop.People tend to judge products based on their own experience of the product rather than others' comments.Therefore, it is correct that people's attitude towards products is almost irrelevant to the historical star rating to a certain extent.

7 Strengths and weaknesses

7.1 Strengths

- **Applies widely**
  The model we build can be widely used in various areas of e-commerce, not only can help merchants track consumer reviews to improve their products and marketing strategies, but also help e-commerce platforms (such as Amazon) to optimize product rankings, Products recommended to consumers are more personalized.

- **Simple and efficient**
  The model we built takes into account various factors and establishes a simple and efficient scoring system that can effectively evaluate the score of each review, and then based on the score of each review, the evaluation of a certain period of time or the product Overall evaluation human resources for the airline.

7.2 Weaknesses

- **Insufficient data**
  In the earlier period, people used online shopping less often, so there were very few reviews during that time, which led to the lack of that part of the data, so it could not reflect the people's evaluation of product quality

- **Incomplete text mining**
  In the text mining process, due to lack of time, although we have screened out some words that can reflect the characteristics of the product, they tend to be homogeneous and cannot fully reflect the characteristics of the product.

8 The letter to Sunshine company

Dear Sunshine Company:

After filtering, sorting and quantifying the data, our team built a scoring model for reviews. A higher review score indicates higher acceptance of this product and better product quality and service. Our team analyzes the microwave, hair dryer and pacifier market based on the model and gives some online sales strategy and a few of potentially design features that might enhance product desirability.

Comparing microwaves, hair dryers and pacifiers together, the store of pacifiers market generally provide products with higher quality and better services and therefore, the market of pacifiers gains highter reputation with fierce competition followed, which places higher requirements on the quality, features and service of products provided by merchants. If Sunshine Company enters the pacifier market, it will face more fierce competition but greater demand. Relatively speaking, there are more complaints about the quality, features and service of products in the microwave market, which is both a challenge and an opportunity for Sunshine Company when entering this market. It's a good choice to focus more resources on the microwave market if Sunshine Company is confident in providing products and services of higher quality than those in the microwave market now.

Through the model we get the features of different products that consumers pay attention to. The product should be designed as much as possible to meet the needs of consumers. So here are some product features designs that companies should pay attention to.

People generally care about the heat setting in the hair dryer, and some hair dryers has the advantages of fast, light, small sound and so on. Among them, heat setting is the most attractive. But there are some shortcomings which consumers complain about a lot, like sparks, smell bad, too hot, too heavy, too loud and so on. It can be seen that customers are very concerned about the temperature and lightness of the hair dryer.

Customers generally concern about the advantage of microwave oven, which is simple, occupying small space , clean and the low price.But the whirlpool microwave got a lot of bad reviews, and users hated the constant need for repairs and worried about its quality.

In the eyes of customers, the pacifier is a perfect gift, soft but tough, and very cute. The pacifier Sunshine Company designed should consider those features. Some customers think that the price of pacifier is too high, and some of the quality of the pacifier is very poor. So the design of pacifier should choose suitable material and control the price.

Yours,

Team:2015224

References

[1] Zhang K, Cheng Y, Liao W, et al. Mining millions of reviews: a technique to rank products based on importance of reviews[C]//Proceedings of the 13th international conference on electronic commerce. 2011: 1-8.

[2] Wes M K. Python for data analysis[J]. 2012.

[3] Dhanasobhon S, Chen P Y, Smith M, et al. An analysis of the differential impact of reviews and reviewers at Amazon. com[J]. ICIS 2007 Proceedings, 2007: 94.

[4] Miao Q, Li Q, Dai R. AMAZING: A sentiment mining and retrieval system[J]. Expert Systems with Applications, 2009, 36(3): 7192-7198.

[5] NLTK 3.5b1 documentation http://www.nltk.org/

[6] tfidf - Wikipedia https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[7] Critical Assessment - Amazon reviews on kindlehttps://www.kaggle.com/adityapatil673/critical-assessment-amazon-reviews-on-kindle

Appendices

Appendix A Data clean code

```
import pandas as pd
import numpy as np
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer


#read the data
hairdryer=pd.read_csv('Problem_C_Data/hair_dryer.tsv',sep='\t')
microwave=pd.read_csv('Problem_C_Data/microwave.tsv',sep='\t')
pacifier=pd.read_csv('Problem_C_Data/pacifier.tsv',sep='\t')

#remove unrelated data
microwave=microwave[microwave.product_title.str.contains('microwave')]
hairdryer=hairdryer[hairdryer.product_title.str.contains('dryer')]
pacifier=pacifier[pacifier.product_title.str.contains('pacifier')]

#function: convert review to sentiment value
sid = SentimentIntensityAnalyzer()
def sentiquantify(sen):
    score = sid.polarity_scores(sen)
    score=score['compound']
    return score

#dataprocess
def dataprocess(df):
    df['review_headline']=df['review_headline'].apply(str)
```

```
    df['review_body']=df['review_body'].apply(str)
    df['review']=df['review_headline']+'. '+df['review_body']
    df['total_sentiscore']=df['review'].apply(sentiquantify)
    df['review']=df['review'].str.replace("[^a-zA-Z#]", " ")

dataprocess(hairdryer)
dataprocess(microwave)
dataprocess(pacifier)

hairdryer['sentiment']=hairdryer['total_sentiscore'].apply(lambda x: 'positive' if x>=0
microwave['sentiment']=microwave['total_sentiscore'].apply(lambda x: 'positive' if x>=0
pacifier['sentiment']=pacifier['total_sentiscore'].apply(lambda x: 'positive' if x>=0 el

#save data
pacifier.to_excel('pacifier_2.xlsx')
hairdryer.to_excel('hairdryer_2.xlsx')
microwave.to_excel('microwave_2.xlsx')
```

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

Input matlab source:

```
#%%
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


# %%
data=pd.read_excel('MCM_NFLIS_Data.xlsx','Data')
data

# %%
data['rate']=data.TotalDrugReportsCounty/data.TotalDrugReportsState
data.head()

# %%
data.pivot_table(['rate'], aggfunc=[sum], columns=['YYYY'], index=['FIPS_State'])

# %%
data.describe()

# %%
data.isnull()

# %%
```

Appendix B Modeling code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
```

```
sns.set()
import warnings
warnings.filterwarnings("ignore")
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.dummy import DummyClassifier
from string import punctuation
from sklearn import svm
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk import ngrams
from itertools import chain
from wordcloud import WordCloud
from nltk import FreqDist
import spacy

#read data
hairdryer=pd.read_excel('Problem_C_Data/hair_dryer_2.xlsx')
microwave=pd.read_excel('Problem_C_Data/microwave_2.xlsx')
pacifier=pd.read_excel('Problem_C_Data/pacifier_2.xlsx')

 # selecting top 20 most frequent words
def freq_words(x, terms = 30):
  all_words = ' '.join([text for text in x])
  all_words = all_words.split()
  fdist = FreqDist(all_words)
  words_df = pd.DataFrame({'word':list(fdist.keys()), 'count':list(fdist.values())})

  d = words_df.nlargest(columns="count", n = terms)
  plt.figure()
  ax = sns.barplot(data=d, x="count" , y = "word")
  ax.set(ylabel = 'word')

stop_words = stopwords.words('english')

# function to remove stopwords
def remove_stopwords(rev):
    rev = rev.apply(lambda x: ' '.join([w for w in x.split() if len(w)>2]))
    rev_new = rev.apply(lambda x:" ".join([w for w in x.split() if w not in stop_words]))
    rev_new=rev_new.apply(str.lower)
    return rev_new

# filter noun and adjective
def lemmatization(texts, tags=['NOUN','ADJ']):
       output = []
       for sent in texts:
             doc = nlp(" ".join(sent))
             output.append([token.lemma_ for token in doc if token.pos_ in tags])
       return output

# output the word freq
def out(review):
    review=remove_stopwords(review)
    tokenized_reviews = pd.Series(review).apply(lambda x: x.split())
    review=lemmatization(tokenized_reviews)
```

```
    for i in range(len(review)):
        review[i]=' '.join(review[i])
    freq_words(review, 30)
    return review

hairdryer['review']=out(hairdryer['review'])
microwave['review']=out(microwave['review'])
pacifier['review']=out(pacifier['review'])

#
g1 = sns.catplot(x="star_rating", hue="sentiment", data=hairdryer,
                 height=6, kind="count", palette="muted")
g1.set_xlabels('hairdryer rating count')

g2 = sns.catplot(x="star_rating", hue="sentiment", data=microwave,
                 height=6, kind="count", palette="muted",hue_order=['positive','negtive']
g2.set_xlabels('microwave rating count')
g3 = sns.catplot(x="star_rating", hue="sentiment", data=pacifier,
                 height=6, kind="count", palette="muted")
g3.set_xlabels('pacifier rating count')

tfidf_n = TfidfVectorizer(ngram_range=(1,2),stop_words = 'english')

# model
def text_fit(X, y, model,clf_model,coef_show=1):

    X_c = model.fit_transform(X)
    print('# features: {}'.format(X_c.shape[1]))
    X_train, X_test, y_train, y_test = train_test_split(X_c, y, random_state=0)
    print('# train records: {}'.format(X_train.shape[0]))
    print('# test records: {}'.format(X_test.shape[0]))
    clf = clf_model.fit(X_train, y_train)
    acc = clf.score(X_test, y_test)
    print ('Model Accuracy: {}'.format(acc))

    if coef_show == 1:
        w = model.get_feature_names()
        coef = clf.coef_.tolist()[0]
        coeff_df = pd.DataFrame({'Word' : w, 'Coefficient' : coef})
        coeff_df = coeff_df.sort_values(['Coefficient', 'Word'], ascending=[0, 1])
        print('')
        print('-Top 20 positive-')
        print(coeff_df.head(50).to_string(index=False))
        print('')
        print('-Top 20 negative-')
        print(coeff_df.tail(50).to_string(index=False))
    return coeff_df

# data preprocess
y_dict={1:0,2:0,3:0,4:1,5:1}
y1=hairdryer['star_rating'].map(y_dict)
y2=microwave['star_rating'].map(y_dict)
y3=pacifier['star_rating'].map(y_dict)
x1=hairdryer['review']
x2=microwave['review']
x3=pacifier['review']

#fit the model
```

```
text_fit(x1, y1, tfidf_n,LogisticRegression(),coef_show=1)
text_fit(x2, y2, tfidf_n,LogisticRegression(),coef_show=1)
text_fit(x3, y3, tfidf_n,LogisticRegression(),coef_show=1)

text_fit(hairdryer.loc[:,'review'][hairdryer['vine']=='Y'], y1[hairdryer['vine']=='Y'],
text_fit(microwave.loc[:,'review'][microwave['vine']=='Y'], y2[microwave['vine']=='Y'],
text_fit(pacifier.loc[:,'review'][pacifier['vine']=='Y'], y3[pacifier['vine']=='Y'], tfi

def dict(x):
    if x in ['y','Y']:
        return 1
    else:
        return 0

def convert(df):
    df['helping_rate']=(df['helpful_votes']+1)/(df['total_votes']+2)
    #df.loc[:,'helping_rate'][df['total_votes']>=10]*=2

hairdryer['verified_purchase']=hairdryer['verified_purchase'].apply(dict)
hairdryer['vine']=hairdryer['vine'].apply(dict)
pacifier['verified_purchase']=pacifier['verified_purchase'].apply(dict)
pacifier['vine']=pacifier['vine'].apply(dict)
microwave['verified_purchase']=microwave['verified_purchase'].apply(dict)
microwave['vine']=microwave['vine'].apply(dict)

convert(pacifier)
convert(microwave)
convert(hairdryer)

def convert2(df):
    df['vine']=df['star_rating'].apply(lambda x:1.25 if x==2 else 1)
    df['verified_purchase']=df['verified_purchase'].apply(lambda x:0.8 if x==0.5 else 1)

convert2(pacifier)
convert2(microwave)
convert2(hairdryer)

# caculate the score
def caculatescore(df):
    df['total_sentiscore']=(df['total_sentiscore']-df['total_sentiscore'].mean())/df['to
    df['score']=(df['total_sentiscore']+df['star'])*df['helping_rate']*(df['vine']+df['v
    df['score']=(df['score']-df['score'].min())*100/(df['score'].max()-df['score'].min()

caculatescore(hairdryer)
caculatescore(microwave)
caculatescore(pacifier)

# convert to datetime format
microwave['review_date']=microwave['review_date'].apply(lambda x : pd.to_datetime(x))
hairdryer['review_date']=hairdryer['review_date'].apply(lambda x : pd.to_datetime(x))
pacifier['review_date']=pacifier['review_date'].apply(lambda x : pd.to_datetime(x))

pacifier['score'].groupby(pacifier['date']).mean().plot()

def convertdate(x):
    return str(x.year)+'-'+str(x.month)

microwave['date']=microwave['review_date'].apply(convertdate)
```

```
hairdryer['date']=hairdryer['review_date'].apply(convertdate)

# plot some figures
microwave['score'].groupby(microwave['date']).mean().plot()
hairdryer['score'].groupby(hairdryer['date']).mean().plot()
pacifier['total_sentiscore'][pacifier['product_title']=='philips avent bpa free contempo
pacifier['star_rating'][pacifier['product_title']=='philips avent bpa free contemporary

pacifier['product_title'].groupby(pacifier['product_title'],).count().sort_values(ascend

microwave['total_sentiscore'].groupby(microwave['date']).mean().plot()
microwave['star_rating'].groupby(microwave['date']).mean().plot()

pacifier['total_sentiscore'][pacifier['product_title']=='philips avent bpa free contempo
pacifier['star_rating'][pacifier['product_title']=='philips avent bpa free contemporary
plt.legend()


import scipy.stats as stats
stats.kendalltau(data1.values,data2.values)

#output the correlate results
stats.stats.pearsonr(pacifier['star_rating'].groupby(pacifier['date']).mean().values,pac
stats.stats.pearsonr(hairdryer['star_rating'].groupby(hairdryer['date']).mean().values,h
stats.stats.pearsonr(microwave['star_rating'].groupby(microwave['date']).mean().values,m
stats.stats.pearsonr(pacifier['star_rating'].groupby(pacifier['date']).mean().values[:-1
stats.stats.pearsonr(hairdryer['star_rating'].groupby(hairdryer['date']).mean().values[:
stats.stats.pearsonr(microwave['star_rating'].groupby(microwave['date']).mean().values[:
```