Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2021021

Time Series Model Based On Sentiment Analysis Of Product Reviews

Summary

In order to solve the problem of informing the marketing manager of the online sales strategy, we make a set of plans to adapt to the commodity system, analyze and process the data, and establish a corresponding model to provide the plans.

In the first part, we use Apriori algorithm to analyze the characteristics of star ratings, reviews, and helpful ratings, and find out that the main factors influencing consumers' purchase decisions are the stars of reviews, the sales volume of products, and the usefulness of reviews.

In the second part, we grade the sentiment of each comment by natural language processing method, and establish a scoring system based on evaluation and scoring. In order to study the influence trend of time on product reputation, we established ARIMA time series model, observed the change of monthly sales volume and the average score of each month's products, and found that reputation will gradually accumulate over time. With the increase of time, the product rating and sales volume have obvious periodicity, which is similar to the service life of the product. In addition, seasonal changes affect the sales volume of important factors, for example, in winter people's demand for hair dryer is the largest. Then we use the N-gram model and Markov hypothesis to make word frequency statistics on the reviews, and get the key features that affect the success of each product. For example, the factors that affect the sales of hair dryer include the power and weight of hair dryer. Next, based on a small range of time series, we explore the impact of specific star rating on the follow-up evaluation, and find that there is a strong correlation among them. In the next period of time when the good reviews decline, the number of bad reviews will increase. Finally, we find that the high star rating is related to some positive energy words by TF-IDF algorithm based on logistic regression, and the low star rating is related to disappointment 、 waste and so on.

Keywords: Apriori;Arima;n-gram;Logistic Regression;TF-IDF

Contents

1 Introduction

1.1 Background

With the advent of the Internet era, online shopping has expanded rapidly, and online reviews have grown tremendously, and it has gradually become an important reference for online shopping decisions made by consumers. The number, quality, attitude, timeliness, and credibility of reviewers will affect the purchase decisions of online consumers, and this information will also give companies and enterprises Sunshine plans to launch and sell new microwave ovens, baby pacifiers and hair dryers on Amazon. They hope to also analyze existing reviews and sales information for these three types of products on the website. To help them acquire online sales strategies and improve their products to attract more customers. sunlight The company provided us with purchases of these three items sold on the Amazon market during the time period indicated by the data Ratings and reviews provided by authors, we will use this data to give them actionable recommendations

1.2 Our Goals

Based on our understanding of the problem, we set the following goals.

1. Establish a new scoring system based on ratings and reviews to help Sunshine get a discount after the product goes on sale Information on consumer preferences
2. Exploring the factors that affect the rise or fall of reputation in the online market based on time series
3. Based on our scoring system, determine the characteristics of each of the most successful or failed products.The conclusion can a ssist managers to develop products.
4. Investigate whether a specific star rating will affect other consumer reviews
5. Investigate whether specific emotional text in text comments is closely related to its rating

2 Notations

Table(1) are the notations and their meanings in our paper:

Table 1: Notation

Symbol	Meaning
Sc	Score comment
Po	positive
Neg	nagative
Ne	Neutral
St	star-rating
SeC	Sence-score
Hss	h-sense-score

3 Data Processing

For data-analysis problem,there are usually some incomplete and abnormal data in the large amount of raw data,which may seriously affect the efficiency of modeling and the accuracy of conclusions.So it is quite important to preprocess the data.

3.1 Data cleaning and normalization

There are some missing items in the given original sequence, and we choose to remove null values. For data normalization, we convert the review period into a time series, which is convenient for future discussion of time-based measurement models.In the given data set, the values of product-category, vine, and verified-purchase consist both uppercase and lowercase. In order to facilitate subsequent processing, we clean the data and change Y and N to lowercase to eliminate the case brought by the case when naming statistics error,and then convert vine and veritifed-purchase into numerical variables of type 0 and 1,we define y as 1 and n as 0. Then we change the names of product-category to lowercase to eliminate the error caused by case when counting. There is also a category of default praises in these comments, namely "No complaints" and "No comment". In view of the subsequent need for natural language processing, these comments are replaced with "good".

3.2 Feature Engineering

In order to facilitate the statistics of star rating and evaluation of the impact on products, we have calculated the total sales volume of each brand from 2002 in the data set according to verified-purchase, and added a column to the table. Based on helpful-votes and total-votes,We divided helpful-votes by total-votes to construct new attribute characteristics helpful-rating，which is used to measure the useful rate of voting.real-helpful is used to determine whether the comment is considered valid or invalid under certain circumstances.Based on the needs of subsequent discussions, we also build a new column of data h-sense-score to assign new weight for supportedcomments.Then we Construct new data attributes real-helpful and review-dis to store discrete variables.

3.3 Discretization of continuous attributes

In the following analysis, we will use one-hot coding, which requires data to be in categorical form, so we will use real-helpful to determine whether the comment is valid or invalid under certain circumstances.As for helpful-rating, from the analysis of the data, we know that when the support rate is greater than 0.8, it is usually accepted as a reference, and when it is less than 0.8, it is considered as invalid information by consumers. So we stipulate real-helpful-1 as a useful rate greater than 80%,real-helpful–1 represents comments as a useless rate, real-helpful-0 represents comment without vote。Then we Discreted the number of reviews to an integer between 1-6 with review-dis.

4 Correlation Analysis

4.1 Correlation of Statistical Numeric Data

Using covariance formula

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{1}$$

to find the correlation between any two variables， then we getting the heat map shown in the figure(1), you can see the three variables vine, verified-purchase, and reviews that are related to star-rating, so we will expand on these three dimensions in the subsequent discussion; helpful-votes and total-votes The correlation is 1, so we divide these two sets of data to get helpful-rating, which becomes an influencing factor。

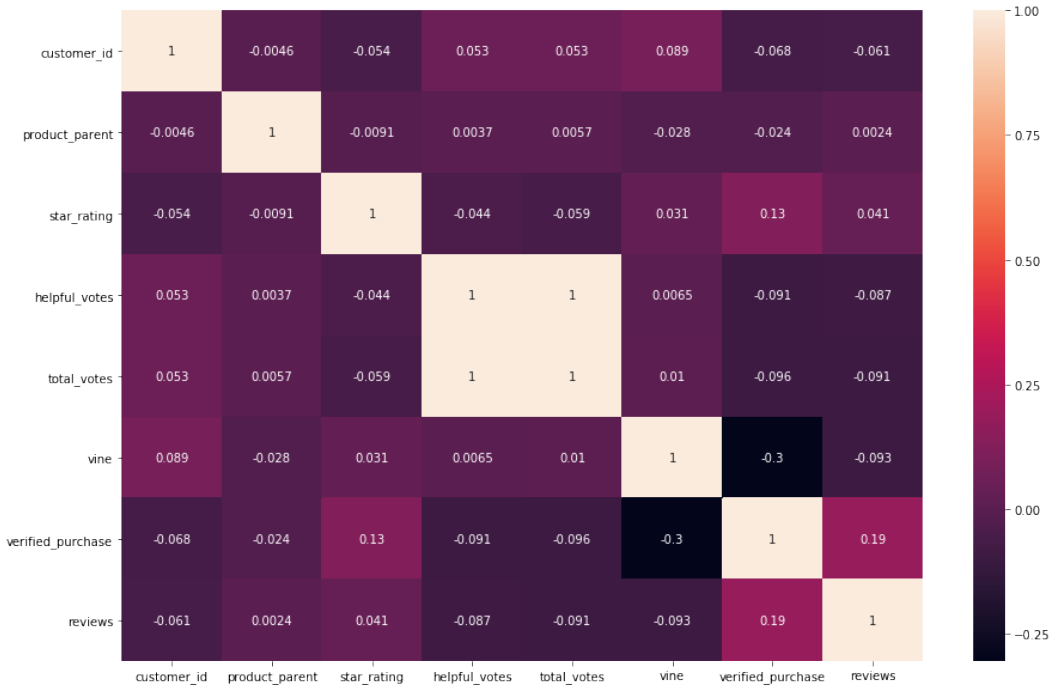

Figure 1: heat map

4.2 Correlation Analysis

4.2.1 One-Hot Encoding

In order to make the values of non-partially ordered variables not partial and to be equidistant to the dots, this paper uses one-hot coding to extend the value of discrete features to Euclidean space. The value of discrete features corresponds to The points in Euclidean space make the calculation of the correlation between features more reasonable.

In the data processing part we have converted the continuous helpful-rating into discrete real-helpful

Figure(2)is the data distribution chart and its standard deviation of helpfu-votes、total-votes and reviews,e can know that reviews are between 0-600, so we divided them into six parts: 1-100, 101-200, 201-300, 301-410, 401-500, and 500-600, and converted into discrete Variable review-dis1-6. We then use one-hot coding to encode these discrete variables.

4.2.2 Apriori Algorithm

In order to determine the closeness of the correlation and the direction of the linear correlation, we use the Apriori algorithm to perform the correlation analysis on the discrete variables obtained by the above method to determine whether it is the vine, the discrete value of the number of reviews, and the support for the evaluation (X) As a condition, whether to confirm the purchase is the result (Y)We use the formula eqref aa to calculate the lift, and use the

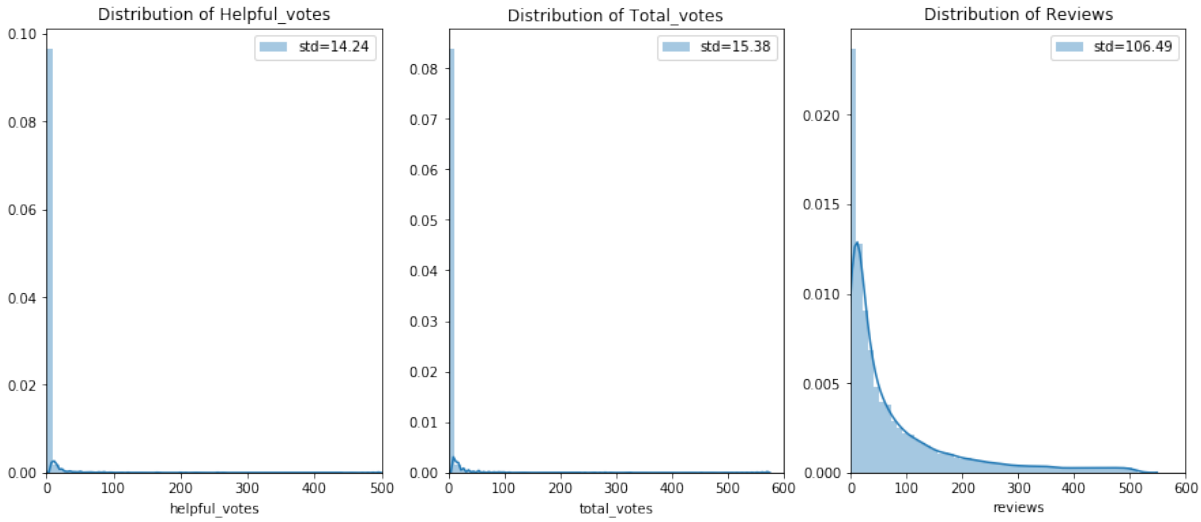

Figure 2: Distribution of helpful votes,total votes and reviews

formula eqref c to calculate the confidence.

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{2}$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{3}$$

Get the association rules of X and Y, and judge the strength of the association rules by calculating the degree of improvement and confidence, so as to determine which of the three factors, star rating, support rate and number of reviews, are the most direct factors that determine whether consumers purchase Whether there is a strong correlation between X and Y), considering that the confidence degree is used to determine how frequently Y appears in the transaction containing X, and the support degree is used to determine how often the rule can be used for a given data set. After the association analysis, we sort the confidence levels and use the lift as a reference to get the results shown in eqref fig: zz. Analyzing these ten rules with the highest confidence, we can know that star rating, which is more able to determine whether users purchase products or not than reviews and helpfulness ratings.

5 Model Construction

5.1 Scoring System Construction

We measure the quality of a product based on its ratings and reviews. The star rating of the product is relatively intuitive. For reviews, we will score each review through sentiment analysis. Our approach is to review each review. The emotions in semantics are divided into positive, neutral, and negative through natural language processing. Calculate the percentage of each comment in this comment, and assign them a weight of 1, 0, and -1. Each review is rated positive, intermediate, and negative. For the weight division between reviews and ratings, after repeated experiments, we found that after normalizing the star-rating, setting the star weight to 0.7 and the review content weight to 0.3, the resulting comprehensive product The score (sense-score) most closely matches the actual data, so we use this allocation scheme to build a new scoring system, and the following figure is the result we got.

$$\mathrm{Sc} = 1 \times Po + 0 \times \mathrm{Ne} - 1 \times Neg \tag{4}$$

[58]:

	antecedents	consequents	confidence	lift
89	(vine_0, reviews_dis_2.0, star_rating_5)	(verified_purchase_1)	0.953210	1.114394
91	(reviews_dis_2.0, star_rating_5)	(vine_0, verified_purchase_1)	0.953210	1.114508
46	(reviews_dis_2.0, star_rating_5)	(verified_purchase_1)	0.953210	1.114394
98	(reviews_dis_2.0, star_rating_5, real_helpful_0)	(verified_purchase_1)	0.953165	1.114341
116	(vine_0, reviews_dis_2.0, star_rating_5, real_...	(verified_purchase_1)	0.953165	1.114341
119	(reviews_dis_2.0, star_rating_5, real_helpful_0)	(vine_0, verified_purchase_1)	0.953165	1.114454
70	(reviews_dis_2.0, real_helpful_0)	(verified_purchase_1)	0.952978	1.114123
106	(vine_0, reviews_dis_2.0, real_helpful_0)	(verified_purchase_1)	0.952978	1.114123
108	(reviews_dis_2.0, real_helpful_0)	(vine_0, verified_purchase_1)	0.952978	1.114236
60	(reviews_dis_2.0)	(vine_0, verified_purchase_1)	0.949196	1.109814

Figure 3: Sort Rules by Confidence

$$\mathrm{SeC} = 0.7 \times St + 0.3 \times Sc \tag{5}$$

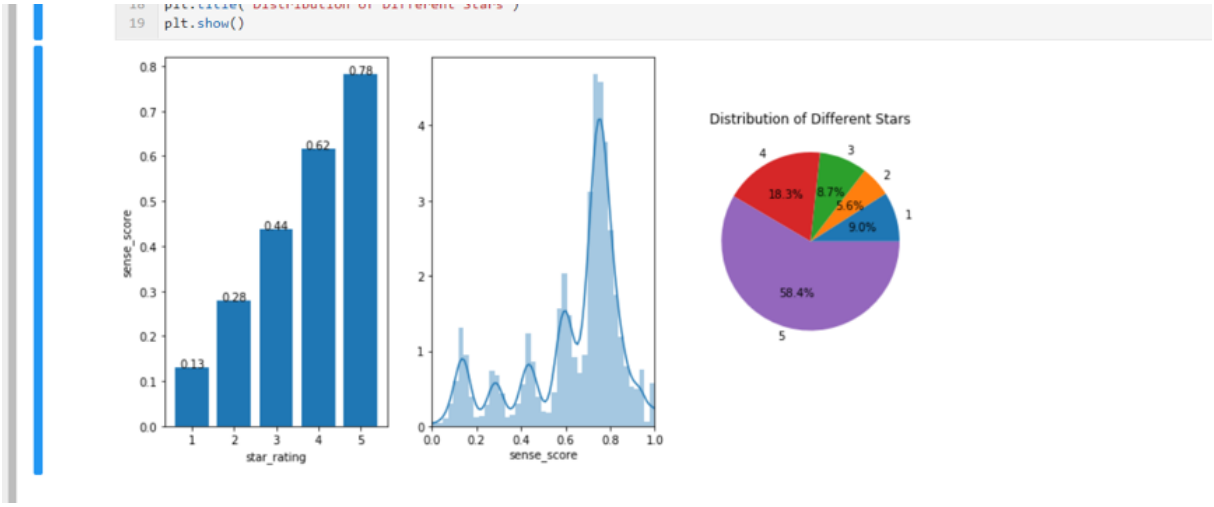

Figure 4: The Results of Scoring System

5.2 Timing Analysis Model

Figure(5)epresents factors that affect consumer decision-making regardless of time. We will assign weights to these factors and verify the correctness of our allocation scheme based on a time series model. Considering that to identify and discuss time-based metrics, we will use time series analysis to solve this problem and establish an Arima (differential autoregressive moving average) model. AR is autoregressive, p is the autoregressive term, and MA is the moving average. q is the number of moving average terms, and d is the number of differences made when the time series is stationary.

5.2.1 Data exploration

Based on our customized scoring system, the total annual sales of the product and the average score of the product are calculated in units of years. Then we display the average score of each

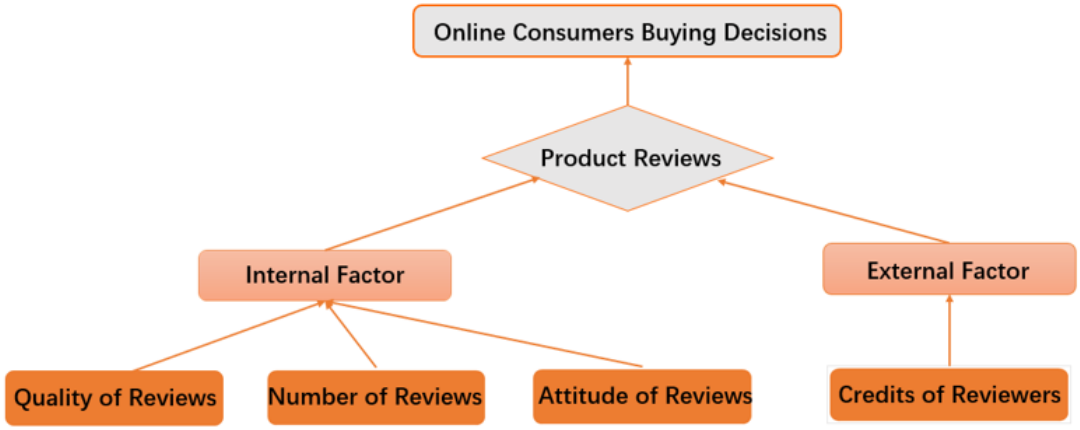

Figure 5: Diagram of the impact of product reviews on online consumer purchase decisions

year in a line chart as Figure eqref fig: 21, you can intuitively see the change in annual sales and the average score of the product each year. Analyzing by month gives the result shown in

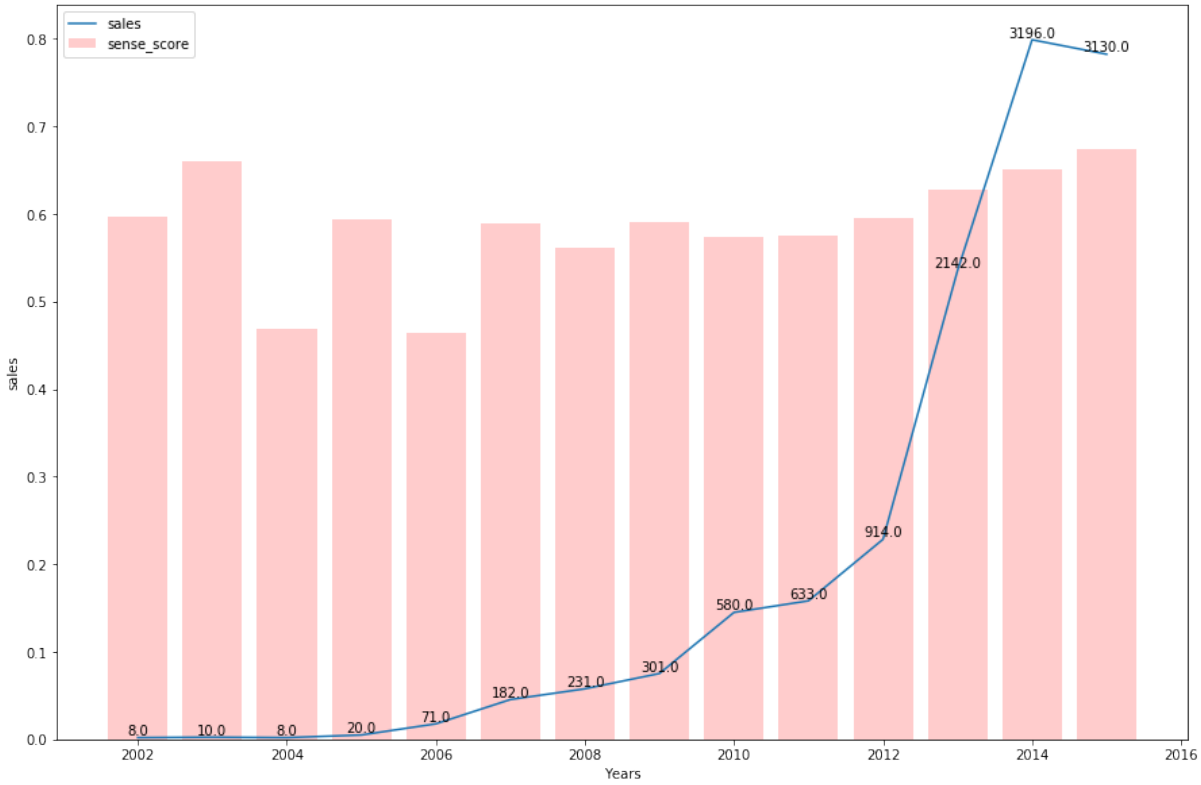

Figure 6: Changes in annual sales and average score for hair dryers per year

the figure eqref 1111. Analyzing by quarter, we can find that in the data of each year, the sales volume of the first quarter is always the highest in the year, as shown in the red dot of eqref 5213, so we know that the season also affects the sales The important factor is that people have the greatest demand for hair dryers in winter.

We build a new column of data h-sense-score, and build a model as shown by the formula eqref 8 and formula eqref 7 to assign new weights to whether a little like comment is useful.

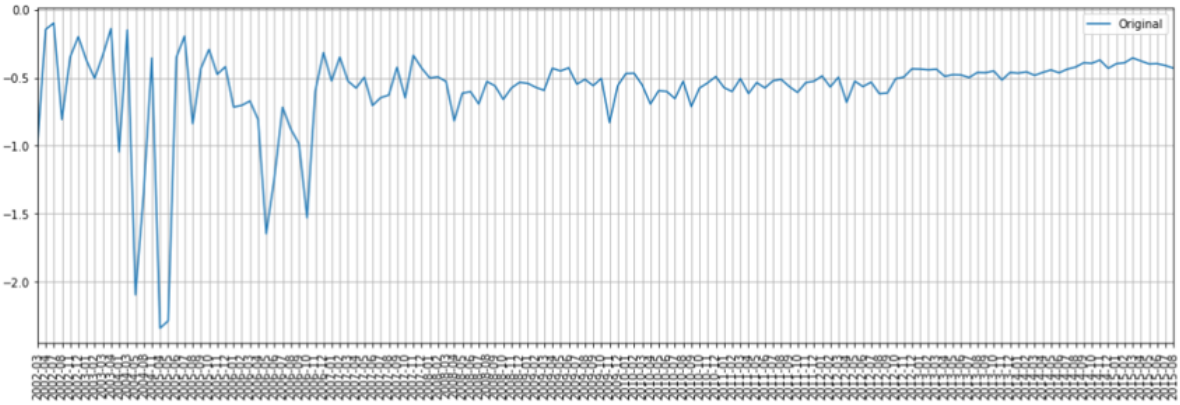

Figure 7: Analysis by Months

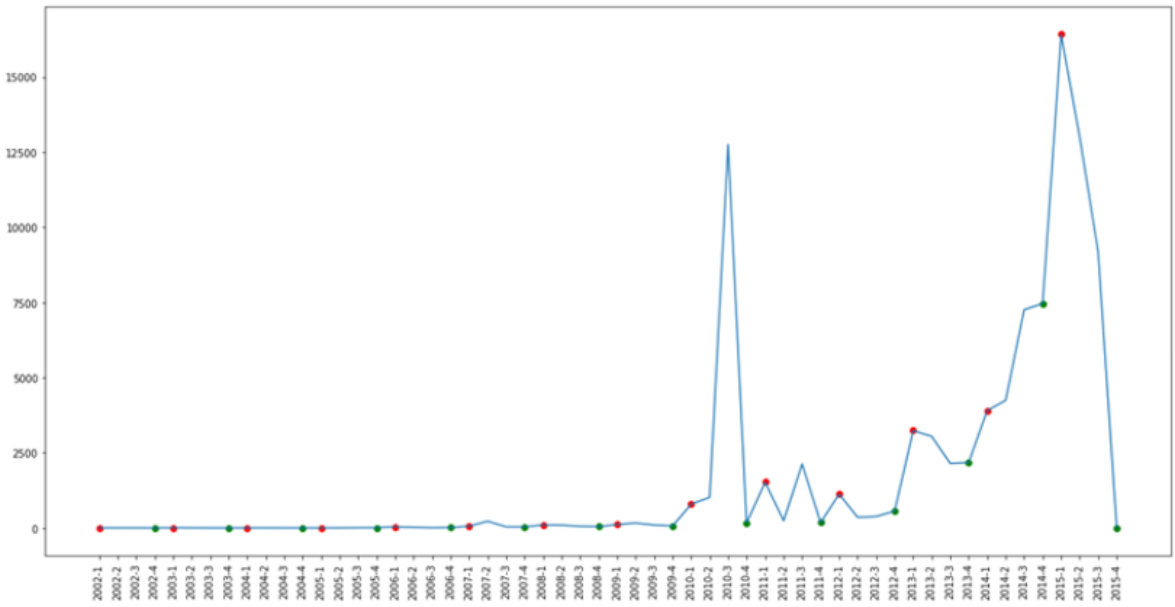

Figure 8: Analysis by Quarter

We assume that when be liked, and the comment is useful, use eqref 6 to assign weight. When the comment is a be liked, and the content of the comment is useless, use eqref 7 to assign weight. When the comment is not be liked, then h- sense-score = sense-score. Next, we will build a model based on this and verify the rationality of this assumption.

$$Hss = (1 + 20\%) \times \text{Sec} \tag{6}$$

$$Hss = (1 - 20\%) \times \text{Sec} \tag{7}$$

5.2.2 Stationary Sequence

Here we establish the Arima model. Since this type of prediction model has certain requirements for stationarity, that is, the mean and variance of the sequence do not change significantly, we will first use the unit root test method of ADF to determine whether our original sequence is Smooth, got the result as shown in Table(2). We found that Test Ststistic <Critical Value (55 %), so the original data is stable and stable within the confidence interval of 95 %, but we further decompose the observation data on the time series and find that it is extremely polar Strong periodicity, as shown in eqref 5221 In order to reduce the error caused by this periodicity, we

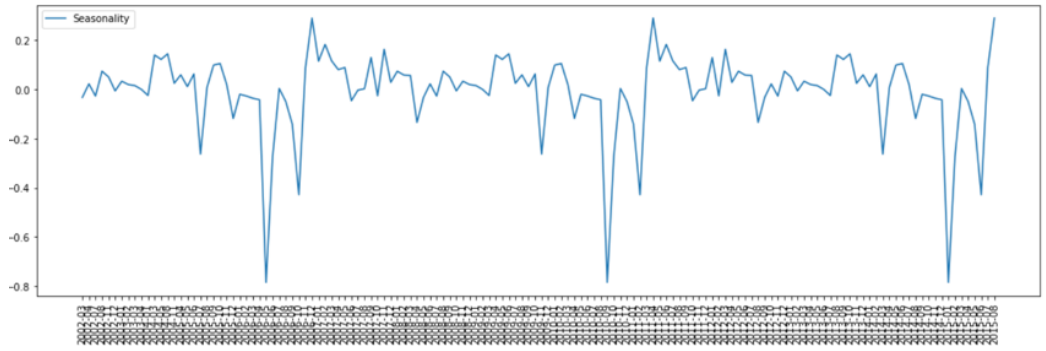

Figure 9: Periodicity of Data

perform logarithmic smoothing of the data and a sliding window of 12. In order to further reduce volatility, we use the difference method to find the time series at time t and t-1. The difference, after performing a first-order difference, it can be known through ADF test result(Table(3)) that the confidence of the obtained data is greater than 99 %, which has strong stability, so the value of d is 1.

Table 2: Notation

Test Statistic	-3.318860
p-value	0.014056
#Lags-Used	8.000000
Number of Observations Used	131.000000
Critical Value(1%)	-3.481282
Critical Value(5%)	-2.883868
Critical Value(10%)	-2.578677
dtype:float64	

Table 3: Notation

Test Statistic	-4.696776
p-value	0.000085
#Lags-Used	12.000000
Number of Observations Used	115.000000
Critical Value(1%)	-3.488535
Critical Value(5%)	-2.887020
Critical Value(10%)	-2.580360
dtype:float64	

5.2.3 Determination of p and q Orders

Establish an autoregressive model (AR) and a moving average model (MA). We will predict our own future through the historical data of the variable itself. The moving average model (MA) is used to calculate the accumulation of error terms in the autoregressive model. Eliminate random fluctuations in predictions. ARMA model formula (8)

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{8}$$

Yt is the current value, u is the constant term, p is the order, yi is the autocorrelation coefficient, and et is the error.

The values of p and q are defined by ourselves. We use the autocorrelation function ACF and the partial autocorrelation function PACF to find the accurate values of p and q. In the ACF model, MA falls within the confidence interval after the q order. (Confidence is greater than 95 %). In the PACF model, AR falls to the confidence interval after the p order. From the figure eqref 5231, we can know that the value of p should be 5, and the value of q should be 6.

Then we train the model, and use the mean square error RMSE to fit the prediction curve and the original curve to get the results shown in the figure eqref 5232. It can be seen that the calculated root mean square error is 0.2154, which is very small. Prove that our above assumption is valid, the model can predict the rising and falling trend of product reputation.

5.3 N–gram Model and Markov hypothesis

Our goal is to determine the combination of text-based metrics and rating-based metrics that best indicate the potential success or failure of a product. Therefore, we must first define the meaning of success and failure. Based on our custom scoring system, the score-score is greater than An evaluation of 0.5 is considered successful, and less than 0.5 is considered a failure. Because the data set is too large, we select a value of 1 for real-helpfu, that is, reviews with a review useful rate greater than 0.5 as the research object, analyze the words that describe the product feature most frequently in such reviews, and then obtain the potential successful product feature. Markov Assumption mentioned that the appearance of a word is only related to a few words before it. We use the N-gram model to complete the word frequency statistics, segment the comments, and use the formula eqref 10 to select a noun to establish a 2-gram model. The review contains information about the product itself, such as dryer and hair that frequently appear in hair dryers. Useful information cannot be extracted from such words, so we will discard it.

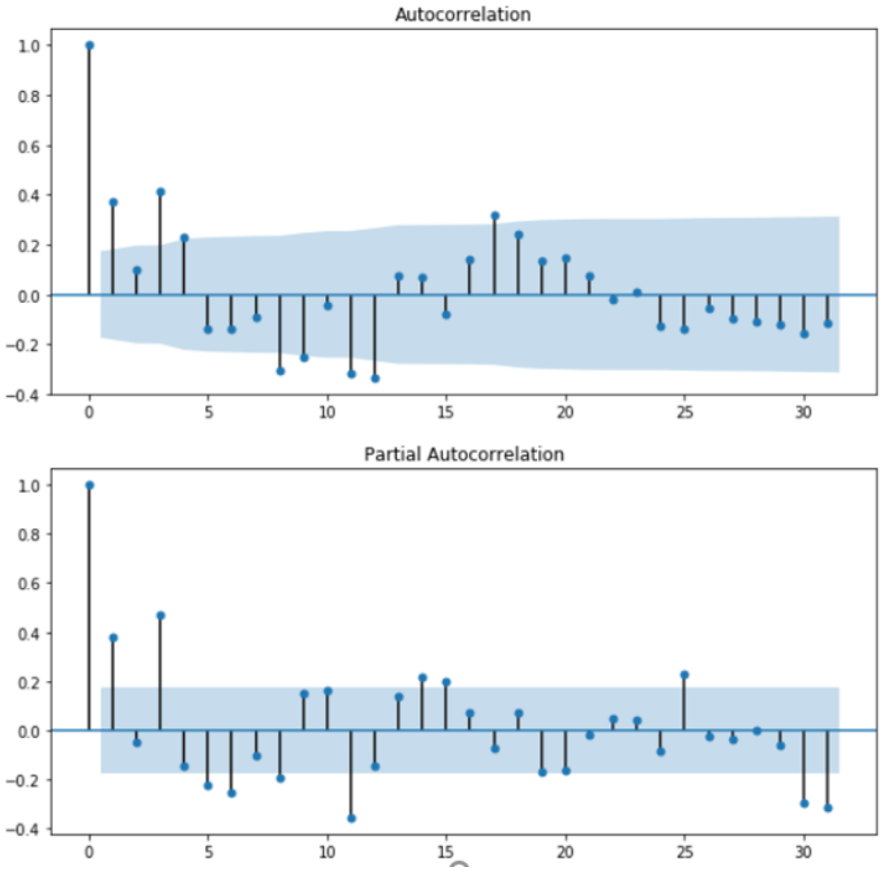

Figure 10:

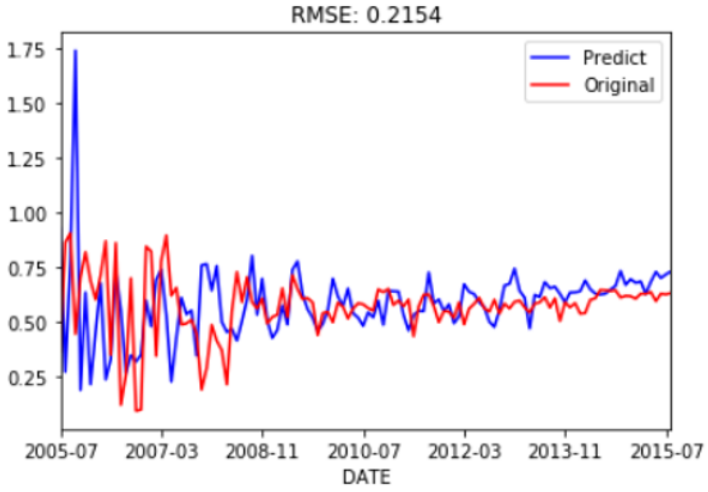

Figure 11:

$$p(w_1 \cdots w_n) = \prod p(w_i|w_{i-1} \cdots w_1) \approx \prod p(w_i|w_{i-1} \cdots w_{i-N+1}) \tag{9}$$

Immediately afterwards, the nouns after removing the useless keywords are used to establish a 3-gram model using formulas, and word frequency statistics are performed on the mentioned keywords to find the factors that determine success and failure. Analyzing the feature words of the successful product and the failed product separately, we have obtained their word frequency statistics table. As shown in the Figure(12)and Figure(13), it is found that the high-frequency words have a large degree of overlap.So we judge that wind speed, heating performance, whether wired are the characteristics that influence the potential success or failure of hair dryer,From Figure(14)and Figure(15),we can also judge that water resistance,after-sales, size, control system, service life are the characteristics that influence the potential success or failure of microwave.Similarly,from Figure(16) and Figure (17) we can judge that adapter and stroller are the characteristics that influence the potential success or failure of pacifier.

5.4 Collateral effects of Specific reviews

In order to investigate whether a particular star rating will cause more reviews of the same type, we have analyzed from a time series perspective to see if the increase in praise over a period of time will lead to an increase in praise after a period of time. Because the number of positive and negative reviews in terms of year has a great relationship with sales volume, and it is not analyzable, here we will explore this question in monthly units. Taking January 2015 as an example, the star distribution map is shown in figure(18), in this figure, blue dots represent five stars, purple dots represent four stars, red dots represent Samsung, green dots represent two stars, and orange dots represent one star. In the data, we can't see the obvious influence of the trend, so we use the difference operation to draw the trend chart, and the result after the difference is shown in the figure(19). Here we list Samsung, Four Stars and Five Stars as positive reviews, and one star and two stars Negative evaluation, taking the dotted line in the figure as an example, we can use the sliding time observation method to intuitively see that the number of negative evaluations starts to increase at the same time or after a short interval; The numbers started to fall, so we came to the conclusion that this commentary has an influential trend over time.

	key_words	frequence
0	(speeds, heat, settings)	7
1	(heat, speed, settings)	7
2	(settings, shot, button)	6
3	(heat, settings, speed)	6
4	(speed, heat, settings)	5
5	(air, flow, heat)	5
6	(speed, medium, heat)	4
7	(folding, handle, cord)	4
8	(heat, settings, button)	4
9	(settings, heat, settings)	4
10	(drying, time, half)	4

Figure 12: Successful Features of Hair Dryers

	key_words	frequence
0	(speeds, heat, settings)	5
1	(Dont, waste, money)	4
2	(heat, settings, speeds)	3
3	(Ive, s, years)	3
4	(heating, element, air)	3
5	(months, heating, element)	3
6	(DO, NOT, THIS)	3
7	(heat, settings, cord)	3
8	(air, flow, heat)	3
9	(product, warranty, product)	2
10	(Ive, s, past)	2

Figure 13: Failed Features of Hair Dryer

	key_words	frequence
0	(back, cabinet, front)	4
1	(size, dinner, plate)	4
2	(control, panel, buttons)	3
3	(start, button, seconds)	3
4	(button, seconds, time)	3
5	(wall, back, cabinet)	3
6	(time, start, button)	3
7	(speed, cook, function)	3
8	(water, vapor, food)	2
9	(vapor, food, temperature)	2
10	(food, temperature, cook)	2

Figure 14: Successful Features of Microwave

	key_words	frequence
0	(service, calls, service)	3
1	(parts, issue, Parts)	3
2	(issue, Parts, parts)	3
3	(service, call, parts)	3
4	(relay, control, board)	3
5	(lineup, metal, frame)	3
6	(metal, frame, stainless)	3
7	(frame, stainless, steel)	3
8	(stainless, steel, covers)	3
9	(model, stainless, steel)	2
10	(stove, stainless, steel)	2

Figure 15: Failed Features of Microwave

	key_words	frequence
0	(car, seat, seat)	7
1	(stuffed, anmal, mouth)	6
2	(car, seat, stroller)	5
3	(car, seat, daughter)	4
4	(car, seat, car)	4
5	(pack, play, mattress)	4
6	(mouth, stuffed, anmal)	4
7	(frog, turtle, months)	4
8	(turtle, months, purchase)	4
9	(months, purchase, rpper)	4
10	(purchase, rpper, threads)	4

Figure 16: Successful Features of Pacifier

	key_words	frequence
0	(car, seat, adapter)	7
1	(car, seat, stroller)	5
2	(adapter, car, seat)	5
3	(seat, adapter, car)	4
4	(seat, car, seat)	3
5	(npple, shape, npple)	3
6	(bumblerde, nde, twn)	3
7	(blah, blah, blah)	3
8	(months, everythng, mouth)	2
9	(bag, daper, bag)	2
10	(head, rest, straps)	2

Figure 17: Failed Features of Pacifier

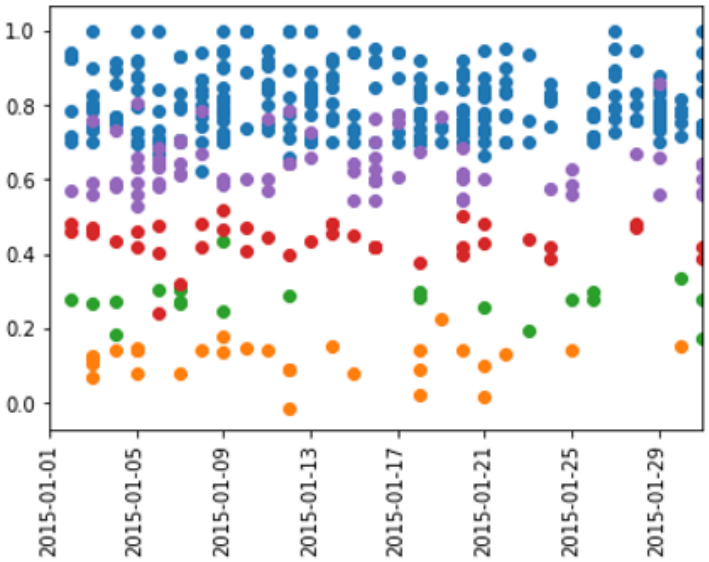

Figure 18: Star scatter chart for January 2015

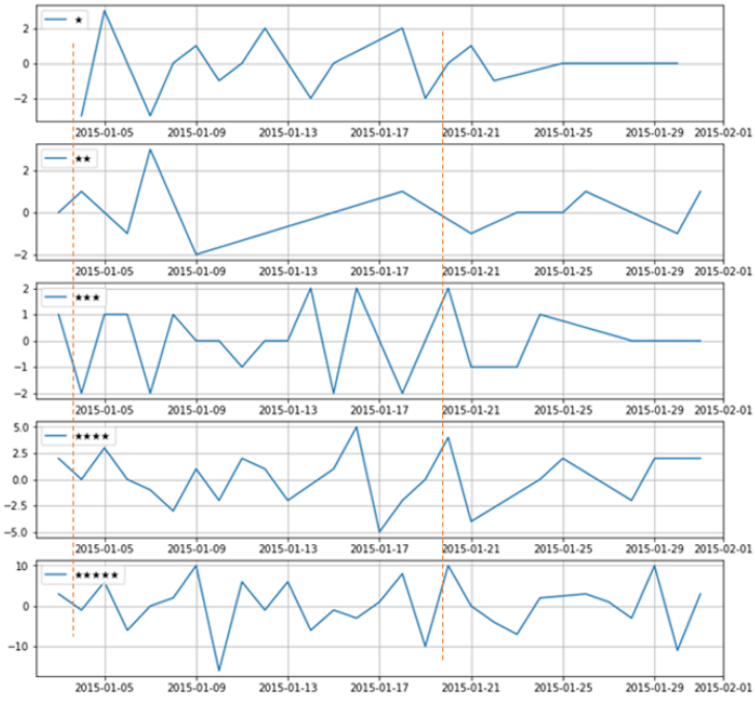

Figure 19: Star difference chart for February 2015 after difference

5.5 Logistic regression model on TFIDF

5.5.1 Model establishment

In order to map predicted values to probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. Here, we use sigmoid to map predictions to probabilities.(10)

$$S(z) = \frac{1}{1 + e^{-z}} \tag{10}$$

Decision boundary Our current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class (true/false, cat/dog), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.

$$if \mathrm{p} \geq 0.5, class = 1 \tag{11}$$

$$if \mathrm{p} < 0.5, class = 0 \tag{12}$$

we use a cost function called Cross-Entropy, also known as Log Loss. Cross-entropy loss can be divided into two separate cost functions: one for y=1y=1 and one for y=0y=0.

$$\begin{array}{ll} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \operatorname{cost}\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) & \\ \operatorname{cost}\left(h_\theta(x), y\right) = -\log\left(h_\theta(x)\right) & \text{if } y = 1 \\ \operatorname{cost}\left(h_\theta(x), y\right) = -\log\left(1 - h_\theta(x)\right) & \text{if } y = 0 \end{array} \tag{13}$$

Now we make the above functions compressed into one

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right) \right] \tag{14}$$

Then vectorized cost function.

$$\begin{array}{l} h = g(X\theta) \\ J(\theta) = \frac{1}{m} \cdot \left(-y^T \log(h) - (1 - y)^T \log(1 - h)\right) \end{array} \tag{15}$$

To minimize our cost, we use Gradient Descent One of the neat properties of the sigmoid function is its derivative is easy to calculate.

$$s'(z) = s(z)(1 - s(z)) \tag{16}$$

Which leads to an equally beautiful and convenient cost function derivative:

$$C' = x(s(z) - y) \tag{17}$$

Meaning of letters in the above formula are showed in Table(4)

C′ C′ is the derivative of cost with respect to weights yy is the actual class label (0 or 1) s(z)s(z) is your model’ s prediction xx is your feature or feature vector.

Table 4: Notation

C'	the derivative of cost with respect to weights
y	the actual class label (0 or 1)
s(z)	your model’ s prediction
x	your feature or feature vector
Test Statistic	-3.318860

5.5.2 Solving Process

First we classify all words into two categories (1, 2 stars, 1 category; 3, 4 stars, 1 category).Then we convert the words into word vectors (TF-IDF).We substitute the word vector into the logistic regression model for training, and obtain the correlation coefficients. The correlation coefficients are ranked from high to low, that is, a scoring result of the words. The first 20 are positive and the last 20 are negative. Figure(20)shows the positive meaning while Figure(21) shows the negative meaning.So we come to the conclusion that text-based descriptions are very relevant to their star rating

6 Model Analysis

6.1 Arima model evaluation

In time series models, we make first-order differences to evaluate the correctness of our model. We use the root mean square error (RMSE) to determine the rationality of our model. We measure the error between our model measurement observation and the true value. We find that The calculated result is 0.2154, as shown in eqref fig: ana1. Through this, the trend of the time series can be fitted. The rationality of the model used is 87.5 %, which represents the expected value of the square of the error. ,we derive the following formula.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} |\,\text{actual}(t) - \text{forecast}(t)|^2}{n}} \tag{18}$$

6.2 Logistic Regression

The analysis report of the Logistic Regression model is shown in Figure(23) We get the model’s comprehensive accuracy rate of 87.5%

7 Strengths and Weakness

7.1 Strengths of Our Model

- Evaluation model established through semantic analysis
- Based on time series analysis, can effectively determine product trends
- Keyword analysis based on n-gram can evaluate product evaluation indicators and key parameters

	Word	Coefficient
11657	love	7.971665
11671	loves	7.044402
8969	great	6.889141
14192	perfect	5.995891
6734	easy	5.232032
2555	best	4.384276
7332	excellent	4.263403
9304	happy	4.118806
9564	highly	3.496373
14552	pleased	3.381403
1977	awesome	3.303229
8736	glad	3.148119
7706	favorite	3.089935
12978	nice	3.051845
21904	wonderful	3.045538
1404	amazing	2.973404
11659	loved	2.860681
16929	satisfied	2.684600
6512	dries	2.515207
21958	works	2.487304

Figure 20: Positive words ordering

	Word	Coefficient
6049	disappointed	-5.745907
16406	return	-4.777641
18813	stopped	-4.289956
21479	waste	-4.197564
3130	broke	-3.896316
20976	useless	-3.721655
16409	returned	-3.717348
5888	didn	-3.609829
6050	disappointing	-3.595231
9904	idea	-3.268725
17240	send	-3.159962
21986	wouldn	-3.147939
14689	poor	-3.066133
5893	died	-3.058467
8041	flimsy	-3.021879
16410	returning	-3.009948
2288	barely	-2.974877
5914	difficult	-2.959832
3298	burned	-2.827146
9719	horrible	-2.813796

Figure 21: Negative words ordering

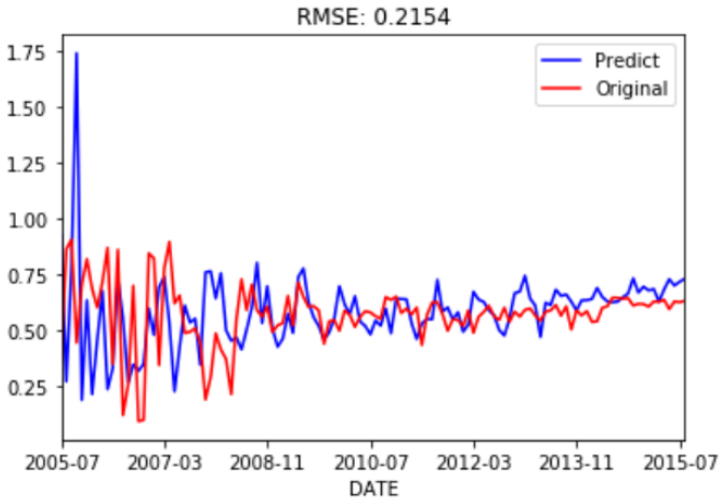

Figure 22: Error analysis with RMSE

```
classification_report(left: labels):
              precision    recall  f1-score   support

         0.0       0.81      0.53      0.64      1701
         1.0       0.88      0.97      0.92      6304

    accuracy                           0.88      8005
   macro avg       0.85      0.75      0.78      8005
weighted avg       0.87      0.88      0.86      8005

Model Accuracy: 0.8750780762023735
```

Figure 23: Logistic Regression Model Evaluation Report

7.2 Weakness of Our Model

- Failure to take into account the timeliness of comments, unable to accurately determine malicious comments

References

[1] Dong Sheng. Application of Natural Language Processing in News Classification [J] .Science Consulting (Education and Scientific Research), 2019 (11): 12-14.

[2] Liu Hua.Association Analysis Based on Apriori Algorithm [J] .Information and Computer (Theoretical Edition), 2019,31 (19)

[3] Fei Jiabao.Research on the Impact of Product Reviews on the Purchase Decisions of Internet Consumers [J] .Modern Commercial Industry, 2018,39 (25)

[4] Xia Yuqin, Shan Xuewei.Simple text sentiment analysis based on Python [J] .Yinshan Academic Journal (Natural Science Edition), 2018,32 (04)

[5] Li Baoku, Zhao Bo, Guo Tingting.Effects of Negative Online Review Contents and Sources on Consumers' Willingness to Buy——Regulatory Effects of Product Categories [J] .Science and Technology, 2019,32 (03): 96-99 + 105.

To: The marketing director of Sunshine Company

From: Team 2021021

Date: March 10, 2020

Subject: A Wealth of Data

First we set up a scoring system based on each product. Through reviews and star ratings, we can determine the final product score.

Then we set up a time series model under the scoring system, observe the changes in annual sales and the average score of the products each year and find that the season is also an important factor affecting sales. In winter, people have the greatest demand for hair dryers, and they can appropriately increase winter production. To get higher benefits. Then established a model that can predict the rise and fall of product reputation. You can predict the praise of the product and the reputation of the product in the market according to the model we established.

Finally we conducted word frequency statistics on the reviews and found out why people prefer certain products and why people don't like certain products. Analyzing the feature words of the successful product and the failed product separately, we have obtained their word frequency statistics table.Here is our conclusion

- wind speed, heating performance, whether wired are the characteristics that influence the potential success or failure of hair dryer;
- water resistance,after-sales, size, control system, service life are the characteristics that influence the potential success or failure of microwave.
- adapter and stroller are the characteristics that influence the potential success or failure of pacifier.

So we suggest you to improve the product from these aspects to get more consumers' favor

The above is the summary of our study.We sincerely hope that it will provide you with useful information.

Thank you!