Studies on the influencing factors and prediction of product star change in the process of e-commerce transaction based on BP neural network and VAR models

Yimeng Chen[1,a], Yue Wang[2,b*], Xingyin Duan[3], Junzhang Li[4]
[1] College of Economics, Sichuan Agricultural University, Chengdu, China
[2] College of Economics, Sichuan Agricultural University, Chengdu, China
[3] College of Resources, Sichuan Agricultural University, Chengdu, China
[4] College of Economics, Sichuan Agricultural University, Chengdu, China

Abstract. Based on the data of reviews and scores of pacifiers sold in Amazon online market from February 2011 to August 2015, this paper extracts the text emotion words and the deviation degree of text content from the theme through the LDA theme model, and then combines the text length, based on VAR model to analyze the impact of the overall star level volatility of the market by comment length, text emotional words and topic deviation. Further, this study compare the prediction of star level by VAR model and BP neural network model, and finally put forward a more stable prediction model.

1 Introduction

Consumers can select commodities based on the information displayed by merchants, and merchants can adjust their sales strategies at different times based on consumer comments on commodities [1]. During the e-commerce transaction process, all transaction behaviors can be retained in the form of data. The text of e-commerce reviews is subjective and often difficult to use directly by sellers. Therefore, finding the link between the review text and the star rating is of some help to the seller in analyzing the product market.

Regarding the extraction and analysis of e-commerce review information, many scholars have conducted analysis and research. The LDA model is an important model for extracting text information. By the LDA theme model, [2] and [3] provide intuitive information for users of e-commerce reviews based on the analysis from multiple dimension. A large number of researchers believe that emotional orientation has a significant contribution to the overall evaluation of products. [4] and [5] believe that the emotional tags of text content have an important impact on people's rapid grasp of product characteristics.

2 Data and methods

2.1 Data source and processing

The data used in this study represents ratings and comments from customers of pacifiers ovens sold in the Amazon market from June 2004 to August 2015.

2.1.1 Data cleaning

We have deleted the items where "vine" and "verified purchase" are either N or Y, that is, the merchant does not participate in the Amazon vine review program, and the customer does not purchase the product in the Amazon market, or the merchant participates in the Amazon vine review program and the customer purchases the product in the Amazon market.

As the evaluation and comment of e-commerce platform are often influenced by some network water forces, based on the criteria of "network water forces" [6], we propose the criteria combining with the actual data:

a) The same user commented on the same product for many times;
b) There is a rating, and the comment title is inconsistent with the comment content;
c) There is too much difference between single comment and other comments;

Based on the above criteria, we manually filter and delete comments with the above characteristics.

2.1.2 Establish index model

Table1. prediction model index system

Index	Number	Unit
Topic	M1	%
Positive words	M2	Unit
Negative words	M3	Uint
Measure words	M4	-

[a]E-mail:736239481@qq.com [b*]Corresponding author E-mail: wangyue@sicau.edu.cn

Star rating	M5	-

2.1.3 Topic deviation and emotional vocabulary

In this paper, LDA topic model is used to predict star level of products based on text classification. LDA model is essentially a Bayesian network with clear logical hierarchy. It is mainly divided into three layers: words, themes and documents. LDA model is used to estimate the topic distribution of documents. It gives the topic of each document in the document set in the form of probability distribution.

LDA model can also be used to calculate the degree of deviation between each comment and the topic. The greater the degree of deviation, the higher the possibility that it is a cyber Navy. There are two parts to calculate the deviation degree between the comment and the topic.

In this paper, the standard topic is generated by the method of average value. For each product, the probability of all comments is taken as the average, and a document topic model of standard comments is generated.

For each comment, we calculate the degree of deviation from the standard comment. In this paper, we use cosine similarity calculation method to get M1, the formula is as follows.

$$\cos\theta = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}} \quad (1)$$

In addition, after analyzing some documents and extracting the topic distribution, we can cluster or classify the text according to the emotional tendency in the comments, and get M2 and M3.

2.1.4 Measure words

In addition, to quantify the amount of information in each comment, we added M4 to measure the number of words. We divide the amount of information in each comment into ten levels. The higher the level, the more information the comment contains.

2.1.5 Other indicators

The rest comes from Amazon.

2.2 BP neural network

BP neural network structure includes input layer, hidden layer and output layer, and each layer is connected by nodes. The model adjusts the weight and threshold of the network continuously through error back propagation to minimize the sum of squares of the network error. After training, the neural network model can store a large number of input-output mode mapping relationships, so as to obtain the prediction results.

The neural network architecture adopted in this paper has one input layer, one output layer and two hidden layers with 10 nodes whose activation function is Tansig and purelin respectively. Then, gradient descent algorithm of Momentum Back Propagation and dynamic adaptive learning rate traingdx are selected for optimization. We divided 53 samples into 47 for training and 6 for test. A star prediction model based on BP neural network is established.

2.3 VAR model

By taking every endogenous variable in the system as a function of the lag value of all endogenous variables in the system, the model is constructed, avoiding the requirements of structural model. VAR model is an effective prediction model for the interconnected time series variable system. At the same time, vector autoregression model is frequently used to analyze the dynamic influence of different types of random errors on system variables.

VAR model describes that n variables (endogenous variables) in the same sample period can be used as linear functions of their past values. A VAR (p) model can be written as:

$$y_t = C + A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + e_t \quad (2)$$

C is n × 1 constant vector, each Ai (A1、A2…) is n × n matrix. Et is n × 1 error vector.

In this study, the VAR star prediction model will be established based on the theme deviation degree and comment sentiment of microwave oven review, and compared with neural network model to find out the best star prediction model for e-commerce market.

3 Model validation

3.1 BP neural network

3.1.1 Prediction accuracy of the model

This paper uses the percentage of prediction error to measure the prediction accuracy of the model. The calculation formula is as follows:

$$p_e = (a_i - e_i)/e_i \quad (3)$$

Among them, "p_e" is the percentage of prediction error, "a_i" is the predicted value of the test sample, and "e_i" is the actual value of the test sample.

3.1.2 Fitting degree of the model

This paper uses AIC information criterion to measure the good of statistical model. The calculation formula is as follows:

$$AIC = \ln(RSS/n) + k/n \quad (4)$$

Where n is the number of training samples and K is the number of explanatory variables.

3.2 VAR model test

3.2.1 Unit root inspection

Unit root test is usually used to check whether there is a unit root in a time series. If there exists unit root, it means that this is a non-stationary time series.

3.2.2 Characteristic root test

The eigenvalue is used to judge whether the VAR model is stable or not. If all the roots of the model fall in the unit circle, the VAR model has reference value.

4 Results and analysis

4.1 BP neural network

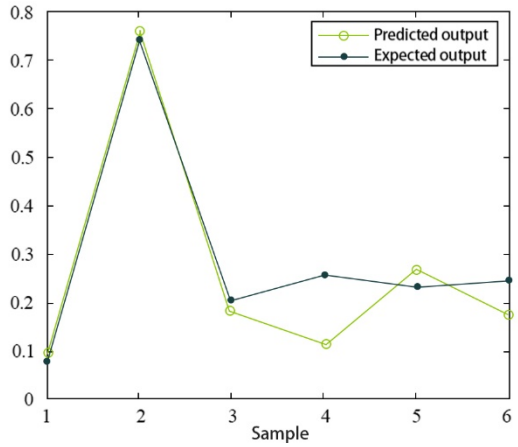

Fig1. Prediction results of BP neural network

Through the MATLAB operation model, the model has a good prediction ability for star series, and the prediction error is between - 0.05% - 1.1%. The RSS of the model is 0.0248 and AIC is -8.6618.

4.2 Vector autoregressive model

4.2.1 Stationary test of variables

The results of ADF test are as follows: all six variables reject the original hypothesis at 5% significance level, that is, all variables are stationary series.

Table2. Unit root test of variables

Variable sequence	ADF test	P value	Conclusion
M1	-7.522694	0.0000	stable
M2	-3.738347	0.0063	stable
M3	-8.49889	0.0000	stable
M5	-4.048987	0.0025	stable
M4	-6.448361	0.0000	stable

4.2.2 Model results and tests

In econometrics, information criteria such as SC, AIC and HQ can help to determine the optimal lag number of the model. As shown in the table below, the results of SC and HQ support the selection of 1-Phase lag. Therefore, this paper establishes VAR (1) model.

Table3. Validation of VAR model's later determination value

Lag	LogL	LR	FPE	AIC	SC	HQ
0	108.277	NA	9.31E-09	-4.303	-4.108	-4.230
1	179.978	125.476	1.34E-09	-6.249	-5.080*	-5.807*
2	207.338	42.181	1.26E-09	-6.347	-4.203	-5.537
3	231.116	31.704	1.45E-09	-6.297	-3.178	-5.118
4	267.061	40.438*	1.10E-09	-6.753	-2.659	-5.206

* Indicates significant at the 5% level

4.2.3 Stability test of VAR model

As shown in the figure, all eigenvalues are in the unit circle, so the model is stable.

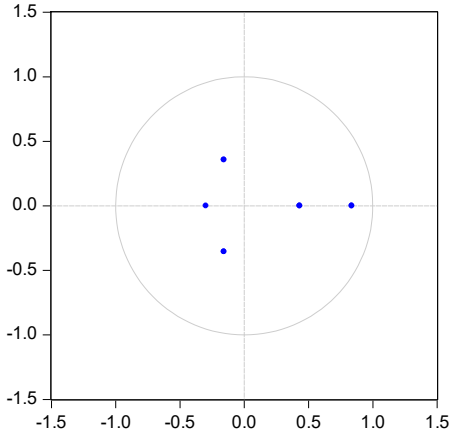

Fig2. Discriminant diagram of VAR system stability

4.2.4 VAR results and analysis

a) Equation estimation results

$$M5 = 0.1755 * M1(-1) + 0.0087 * M2(-1) + 0.0346 * M3(-1) - 0.0017 * M4(-1) + 0.7401 * M5(-1) + 1.2449 \quad (5)$$

b) Impulse response

Figure 4 shows the response of as to the change of one standard deviation unit of each variable. The impact of *M5* on itself is 0.0267 in the current period, and then gradually decreases to a stable level; for the impact of *M2, M3 and M4*, the current fluctuation of *M5* is 0, reaching the peak in the second period, and then decreases with the increase of the number of periods and finally tends to be stable.

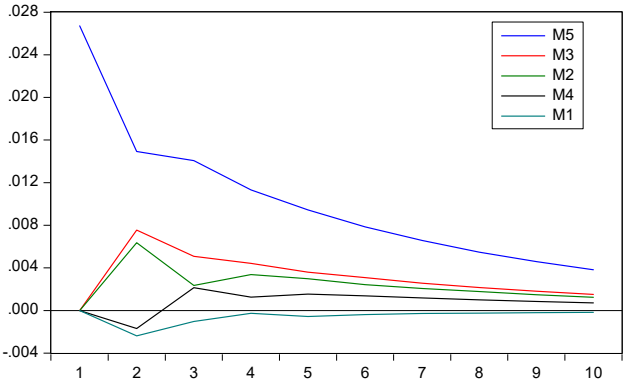

Fig3. Impulse response of M5 to impact of various variables

c)Variance decomposition

It can be seen from table IV that *M5* has the largest contribution to itself, which is stable at about 86% after the sixth period; the *M1* and *M2* is less than 1%. The

contribution of the former decreases with the increase of the number of periods, and the contribution of the latter increases with the increase of the number of periods. It is worth noting that *M2* and *M3* have more and more influence on *M5* with the increase of the number of periods, and the contribution rate of *M3* is twice that of *M2*.

Table4. Variance decomposition results of M5

Period	S.E.	M1	M2	M3	M4	M5
1	0.027	0.000	0	0	0	100
2	0.032	0.514	3.883	5.465	0.2940	89.844
3	0.036	0.511	3.600	6.482	0.5883	88.819
4	0.038	0.460	3.993	7.116	0.6316	87.800
5	0.040	0.450	4.273	7.427	0.7357	87.114
6	0.040	0.439	4.429	7.656	0.8144	86.661
7	0.041	0.431	4.54	7.792	0.8693	86.368
8	0.042	0.426	4.618	7.883	0.9073	86.165
9	0.042	0.423	4.67	7.945	0.9341	86.028
10	0.042	0.420	4.706	7.987	0.9525	85.934

d)Prediction effect

The VAR model is established to predict as in 2015m03-m08 for six months, with a prediction error of 0.07% - 0.27%, which has a good prediction ability.

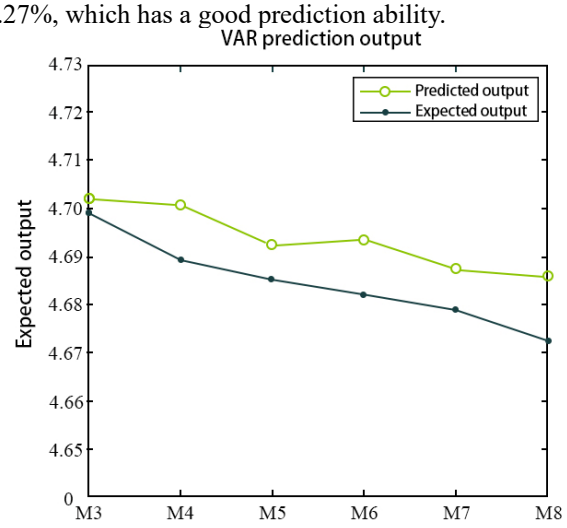

Fig4. Prediction results and errors of VAR model

5 Results and analysis

We use LDA topic model to extract the content of e-commerce text reviews, and analyze the potential influencing factors of pacifier review stars on Amazon from three aspects: emotional vocabulary, number of review words, and topic deviation. By analyzing the results of VAR model, we find that the whole market stars, positive emotion words and negative emotion words in the previous period have a significant impact on the current market stars. It is worth noting that in the variance decomposition, the contribution of negative emotion words to star level fluctuation is nearly twice as much as that of positive emotion words, which indicates that in the pacifier market, consumers are more likely to be affected by negative emotion words when commenting on goods. At last, we use BP neural network model and VAR model to predict the data out of sample. The prediction results show that the two models have excellent prediction ability.

Acknowledgment

The research was financially supported by the Sichuan Provincial Innovation Training Program for College Students (Item No.: 1921996527).

References

1. Feng Jiao, Yao Zhong. Research on the influence of online comment information on purchase decision based on social learning theory [J]. China management science, 2016,24 (09): 106-114.
2. Zhang min. emotional analysis of e-commerce reviews based on text mining [J]. Industry and Technology Forum, 2020,19 (02): 63-64.
3. Zhang Fan. Construction of mixed model of fine-grained theme emotion in e-commerce review [J]. Business economy research, 2017,24:55-57.
4. Mu Shengjiao. "Internet plus" environment, e-commerce review big data analysis and Application Research [J]. technology and information, 2018,08:71.
5. Li Chenlong, Tao Wan. Acquisition and research of e-commerce review based on Hadoop [J]. Journal of Jiujiang University (NATURAL SCIENCE EDITION), 2019,03:64-68.
6. Wang Junbo. Online naval identification based on e-commerce review [D]. Beijing Jiaotong University, 2016.