Research on the relationship between e-commerce ratings and reviews based on Naive Bayes

Zeqing Qin 1*, Zesong Wang 2

1 School of Computer Science, Hubei University of Technology, Wuhan, Hubei, 430064, China
2 School of Computer Science and Information Engineering, Hubei University, Wuhan, Hubei, 430062, China
*Corresponding author's e-mail: zeqing@hbut.edu.cn

Abstract. With the popularization of the Internet, online shopping has gradually replaced traditional shopping methods. Product ratings and comments from buyers have become important indicators for users when purchasing products. Because of this, the behavior of employing someone to rate high, known as click farming, has gradually risen. To deal with this phenomenon, users pay more attention to the content of the review itself rather than the score to make judgments. Then why are the stores still keen on employing click farm? We think that the high scores seem to play a guiding role in the user's comments. To explore whether this phenomenon exists, we crawled sales information from Amazon and built a mathematical model. A Naive Bayes Classifier is employed to screen out the high-frequency vocabulary from the comments. And we defined the important concept of Irrational Discrimination base on Multiple Linear Regression. After analyzing the data for the whole year, we found that discrimination does exist. The specific vocabulary appearing in the review has a large star rating relevance shows that not only the time guiding role exists between star ratings and reviews, but also a strong interaction.

1. Introduction
With the rise of the e-commerce industry, shopping has become more convenient, but unlike offline shopping, since the products cannot be accessed, the ratings and reviews of online products are an important measure of the quality of users. The growing significance of this approval has unleashed a rather unscrupulous business activity, click farm. Click farm[1] is an organized group of low-paid workers employed to click on particular parts of web pages, especially approval buttons in social media as a way of making businesses seem popular. However, this phenomenon destroys the level playing field. We have noticed that most of these click farming behaviors are based on high scores, but the actual comments are monotonous and have no reference value. Since consumers are paying more and more attention to detailed reviews of products, why are the stores still keen on employing click farm? Does the specific star rating of the product have a certain guiding effect on later reviews? To explore whether there is a specific interaction between star rating and reviews, we chose Amazon's products as the research object and established a mathematical analysis model.
2. Data analysis and processing

2.1. Data analysis
We have observed that small household electrical appliances have been in high demand in recent years. We crawled 11471 pieces of data on hair dryer products from Amazon. Each piece of data contains information such as customer ID, product number, product name, user rating, user review, and the number of support votes and total votes. Among them, the rating is a specific score in the interval [1,5], but the reviews are diverse. To judge the guiding effect of the rating on the review, we first need to quantify the review. In this work, the quantification of textual information is a very important step.

2.2. Data processing
The amount of original data is so large that we should first screen the data based on the completeness and usefulness of the information.

- Remove products with less than ten reviews
  We do this for three reasons:
  1. Their star ranking has a strong chance and lack of universality.
  2. We cannot extract universal words in it.
  3. The data set is doped with some irrelevant products. Taking the hairdryer as an example, the product named Badger Balm Lavender Sunscreen Cream- SPF 30-2.9 oz is convenient for hairdryers, we have counted the reviews of these products and found them (All of them are less than ten), this step we screen out irrelevant variables.

- Delete specific reviews
  We found that there are malicious comments and slips in the data. The same user id reviews the same product multiple times a day. We have filtered the data. If the user id, purchase product name, and review date of multiple reviews are the same, we suspect that it is a problem review. We filtered out these reviews and excluded them.

3. The principle of our model
To quantify user reviews, we first use the natural language processing technology and a trained Naive Bayes Classifier[2] to get the positive sentiment ratio[3-4] of each review. After the evaluation of reviews, we propose the important concept of irrational discrimination and design an experiment to determine whether the phenomenon of irrational discrimination exists. The principle is to calculate the proportion of positive sentiment words in user reviews, which is divided into 3 steps:

- Acquire the positive and negative emotions of hair dryers
  To evaluate the sentiment of user reviews, we must first filter out the positive and negative words of these reviews, which are used as the training set of the classifier. Here we use the word segmentation analysis technology in natural language processing.

- Train the Classifier with large amounts of labeled data.
  The Text data used for training is labeled with positive and negative emotions. Through training on a large number of related data sets, the Naive Bayes Classifier can remember the main positive vocabulary of the item and identify them immediately when contact again.

- get the positive sentiment ratio of the review.
  The trained Naive Bayes Classifier recognizes and calculates the ratio of positive sentiment words in the text to get the positive sentiment ratio of text.

To make better use of the time factor[5], we have established a short-term monthly plan. The monthly plan can better see the short-term changes: The monthly model helps us to grasp the real-time dynamics and adjust in time. Because the data applicable to monthly analysis has a large number of reviews in a short period, and the details of the reviews of its products are more obvious, we have defined Formula 3 for reporting monthly reviews.

\[
MR_i = M \lor 1 \cdot MS_i + M \lor 2 \cdot MC_i
\]
Where $\overline{MS}_i$ represents the monthly average star value of product $i$, and $\overline{MC}_i$ represents the monthly average score of product $i$. $M\omega r1$ and $M\omega r2$ represent the weight of monthly report analysis. Considering the short-term sales, the discount can greatly affect the sales situation, and the monthly review situation will also be affected. Based on this, we define the Monthly Fluctuation Formula (2) for discount rate monthly review situation.

$$MF_i = M\omega r1 \cdot \overline{MS}_i + M\omega r2 \cdot \overline{MC}_i + M\omega r3 \cdot \overline{MV}_i$$

(2)

Where the monthly average discount rate is $\overline{MV}_i$. $M\omega r1$, $M\omega r2$ and $M\omega r3$ represent the weight of monthly fluctuation.

4. Quantification of reviews

First, we calculated the average star ratings of various products and sorted them from high to low. In various types of data, we took the top ten stars as the positive representatives and the last ten as the negative representatives. The word frequency of the group was counted.

After performing word segmentation, we conducted appropriate manual intervention on the text content. Words such as "I / He / She" that have nothing to do with emotional expression and product characteristics were removed. Figure 1 is a quantitative description of the corresponding vocabulary.

![Figure 1. Description of the vocabulary](image)

Then, we trained the classifier using the positive feature and negative feature mentioned above, and the positive sentiment ratio of the review was calculated.

5. Interaction Of Star Rating And Reviews

In this section, we explore whether a large number of specific star values in the early days could guide the later reviews. This is an irrational phenomenon. We define it as "discrimination" and establish a model of discrimination[6]. Based on time, the impact of early star ratings on later reviews is presented. We use Multiple Linear Regression[7] to explore discrimination in the comments. Because reviews are strongly influenced by star ratings, we use product sales data within one year and explore whether star ratings in the first half of the year will have a significant impact on reviews in the second half.

Because our discussion involves time effects, we use the above decision-making model to select 11 pacifier products suitable for the monthly plan as the research object. The ID of the eleven products is shown in Table1.

| Variable Name | ID       |
|---------------|----------|
| A             | 473538559|
| B             | 667171015|
5.1. Descriptive analysis

First, we used the monthly fluctuation model to evaluate the performance of these 11 merchants in the first half of the year and made Figure 2 based on the corresponding review scores in the second half of the year. The bar graph in the figure shows the average review score (Score Of Phase) of each product in the second half of the year, and the broken line is the combined score (CS) of the product in the first half of the year. The horizontal dashed line in Figure 10 represents the average of the comprehensive performance.

According to common sense, the performance of the first half should largely determine the review score in the second half. It can be seen from the trend in the figure that the comprehensive score of A in the first half is higher than K, but the review score of the second half is much lower than K. Therefore, we suspect that there is discrimination. That is to say, the current review score depends not only on past performance but also on other irrational factors.

5.2. Verification of discrimination

Because of discrimination, we mainly want to see whether there is a significant relationship between the praise rate and individual products. Will different stores be sought after or discriminated against. To quantitatively analyze the above problems, we constructed Formula 11 based on the idea of Multiple Linear Regression:

$$S_{OR_i} = \alpha + \sum \beta_n \times i_{id_n} + \lambda \times Controls_i + \epsilon_i$$

Where $S_{OR_i}$ represent the ith review’s score. To avoid the effects of multicollinearity, we introduce a series of product dummy variables $id_n (n=10)$. If the $i_{th}$ sample is from the $k_{th}$ product, $id_k = 1$, $id_{id_k} = 0$.

We used each product as an independent variable. Here, the significance of the regression is to explore whether the coefficients for each store are jointly significant. If yes, it means that there is a difference in the relationship between different stores and the positivity of reviews. In this work, we use OLS to estimate the coefficients and test the significance of $\beta_0 = \beta_1 = \beta_2 = ... = \beta_n$. We get $F(10, 3756)$...
= 5.38 and Prob>F = 0.0000, which shows the existence of discrimination. The solid line in Figure 3 represents the standardized regression coefficient of each product, where the point marked with * indicates that the store is significantly different from 0. To make the difference in review positivity more intuitive, we used dotted lines to indicate the difference between the review positivity of each store compared to store A. We can see that the review positivity has a great relationship with the store itself. It is rising from left to right.

Figure 3: Verification of discrimination

5.3. Follow the trend

After confirming the existence of discrimination, we further judge whether the existence of this phenomenon is determined by a specific star rating. Based on the above analysis, we divide existing products into three categories based on review scores:

1. Discrimination: A, B, C
2. Normal class: E, F, G
3. Sought after: H, I, J, K

\[
SOR_i^{(1)} = \alpha^{(1)} + \sum \beta_n^{(1)} \times Class_n^{(1)} + \lambda^{(1)} \times Controls_i^{(1)} + \varepsilon_i^{(1)}
\]

\[
SOR_i^{(2)} = \alpha^{(2)} + \sum \beta_n^{(2)} \times Class_n^{(2)} + \lambda^{(2)} \times Controls_i^{(2)} + \varepsilon_i^{(2)}
\]

Then, we divided the above-mentioned stores into two categories according to the negative star rating (1, 2 stars) in the first half of the year. Formula 12 is for negative stars and Formula 13 is used for positive stars. If the \(i_{th}\) sample is from the \(k_{th}\) Class, \(Class_k = 1, Class_{i(k\neq k)} = 0\). Finally, we calculated the absolute values of the standardized regression coefficients of the two regressions, both of which are significant, to show the terms more intuitively. As can be seen from Figure 4, the significance of low-star data in high-discrimination stores is obvious, which indicates that the product is not well commented and it is related to its early acquisition of a large number of low-stars. In the same way, the high surname data is obvious in the overly sought-after products, which shows that the early star rating has a great influence on the later reviews.

Figure 4: the significance of star data
6. Conclusion and future work
After analyzing the data of products for the whole year, we found that discrimination does exist. Through further refinement, we concluded that the specific star rate in the first half of the year will indeed lead the review in the second half of the year. We found that the specific vocabulary appearing in the review has a large star rating. Relevance shows that not only the time guiding role exists between star ratings and reviews, but also a strong interaction.

In the future, our model will be optimized by getting more comprehensive data and applied to deal with some practical engineering problems.

References
[1] N. Li, S. Du, H. Zheng, M. Xue and H. Zhu. (2018) Fake reviews tell no tales? dissecting click farming in content-generated social networks. China Communications, vol. 15, no. 4, pp. 98-109.
[2] Peng, F., Schuurmans, D. & Wang, S. (2004) Augmenting Naive Bayes Classifiers with Statistical Language Models. Information Retrieval 7, 317–345.
[3] Wawre, S. V., & Deshmukh, S. N. (2016) Sentiment classification using machine learning techniques. International Journal of Science and Research (IJSR), 5(4), 819-821.
[4] Pang, B., Lee, L., & Vaithyanathan, S. (2002) Empirical methods in natural language processing. In: Proceedings of the ACL-02 conference. 10: 79–86.
[5] Zhang Yanfeng. (2018) Research on Time Series Correlation Rules of Online User Reviews Behavior. Jilin University.
[6] Pope, D. G., & Sydnor, J. R. (2011) What’s in a Picture? Evidence of Discrimination from Prosper.com. Journal of Human resources, 46(1), 53-92.
[7] Gülden Kaya Uyanık, Neşe Güler. (2013) A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences Volume 106, Pages 234-240.