Data Wealth Mining Based on Text and Time

Jie Liu

Department of Power Engineering, North China Electric Power University, Baoding 071000, China
ljel@ncepu.edu.cn

Abstract. In this paper, online sales strategies for microwave ovens, baby pacifiers and hairdryers in Amazon are studied based on sentiment analysis and opinion mining. Firstly, we analyze and quantify star ratings, reviews, and helpfulness ratings based on the corresponding factors, and give one mathematical model for the evaluation of product. Secondly, we identify and discuss time-based measures and patterns within each data set to analyze whether the reputation of the product is rising or declining. Finally, we predict the goods that may be potentially successful or failing taking all indicators into consideration, and put forward suggestions on important design features that would enhance product desirability.

Keywords: Sentiment Analysis and Opinion Mining, Long Short-Term Memory, LSTM, Natural Language Processing, NLP, Reviews and Ratings

1. Introduction

In the online marketplace it created, Amazon offers customers an opportunity to rate and review purchases, such as star rating, review and helpfulness rating. Then, companies use these data to gain insights into the markets in which they participate, the timing of that participation, and the potential success of product design feature choices.

Now Sunshine Company is planning to introduce and sell three new products in the online marketplace: a microwave oven, a baby pacifier, and a hairdryer. They want to analyze these to inform their online sales strategy and identify potentially important design features that would enhance product desirability.

2. Assumptions

To simplify the considered problems, we make the following basic assumptions, which are properly justified.

- Reviews with high rating valence are perceived as more helpful than ones with a low rating valence.
- Reviews with a large number of votes are perceived to be more helpful than ones with fewer votes.
- The source of the data is reliable and the data is complete.
- Assuming that the customer's personal identity, cognitive level and wealth level are all the same, then the reviewer's own situation is not taken into account.
- Suppose the sales volume of the goods is proportional to the number of re-views.
• It is assumed that the more words in the reviews, the richer and more detailed the content of the comments.
• Assuming that customers with verified purchase “N” are not purchased on Amazon platform, or at a huge discount, their comments are less credible.
• Assuming users whose vine is “Y” have a higher reputation and a higher credibility of their reviews.

3. Methodology

3.1. Quantitative Data Analysis
The hairdryer, microwave and pacifier datasets provided had missing and partially filled in data that would have been challenging to effectively utilize, so we firstly process redundant data and missing value.

3.1.1. Star Rating [1]. We describe the star rating of the three products qualitatively and quantitatively. We counted the proportion of 1 star, 2 stars, 3 stars, 4 stars and 5 stars of hairdryer, microwave and pacifier in all reviews, as shown in figure 1. The average star ratings of the three products are 4.12, 3.44 and 4.30 respectively.

![Figure 1. Star Rating Percent](image)

3.1.2. Review Valence. The relevant research indicates that the description and analysis of reviews can start with the reviewers and the reviews themselves. Since we are unable to mine customer information from customer ID, we convey the potential characteristics of comments through review depth review sentiment analysis [2], review readability, reveal reputation and number of reviews.

(1) Review Depth [3]
Review depth is usually described by the number of words in the text of the reviews, and the study on review depth can improve the diagnostics of information. In general, longer reviews are considered to be shorter ones that contain more information and are more persuasive. So we will analyze the reviews of the three products as corpus one by one.

(2) Review Sentiment Analysis
We take the reviews of three products as corpus, carry out text sentiment analysis based on Long-Short Term Memory RNN [4] in KERAS, and use tanh activation function during training. Taking the review with customer ID 50274025 in the hairdryer data file as an example, the implemented model framework is shown in figure 2.
According to the analysis of the datasets, the reviews with medium length and strong subjectivity are the most readable and the most helpful; the reviews that are too short or too long or contain too much product professional information are not readable.

(4) Review Reputation

In this paper, we define that the review reputation is mainly affected by vine, verified purchase and review depth:

\[ RRE = \sqrt{0.287V^2 + 0.018VP + 0.695RD} \]

(5) Number of Reviews

To a certain extent, the number of reviews can reflect the customer's attention to the product market, namely, sales volume, so we analyze the demand of this product market by analyzing the number of reviews.

3.1.3. Helpfulness rating. We describe and analyze helpfulness rating according to review reputation, review rating and reveal depth.

- Reveal reputation: Same as defined in reveal reputation above
- Reveal rating: We quantitatively describe the comprehensive score of reveal rating based on SR and RSA.
- Reveal depth: Review depth is described by the number of words in the text of the reviews.

3.2. Product’s Success Evaluation

The quantitative process of the five indexes is as above, and the results are as follows:

- SR: 1-5 stars correspond to 1-5 points respectively.
- RSA: The sentiment polarity obtained by LSTM [7] can quantitatively describe RSA.
- RD: Sentence length size can quantify RD.
- RRE: \[ RRE = \sqrt{0.287V^2 + 0.018VP + 0.695RD} \]
- NR: Number of reviews.

Then, the correlation of these five indexes is analyzed by SPSS, and it is found that the correlation between them is relatively high. Therefore, we determine the most informative data measures based on principal component analysis [6]. We normalize the data of the five indexes, and then select
ratings(RA), reviews(RE) and sales volume(SV) as the principal components. And, from the above analysis, we can know:

\[
\begin{align*}
RA &= \omega_1 \times SR + \omega_2 \times RSA \\
RE &= RSA \times \sqrt{0.287 \times V^2 + 0.018 \times VP + 0.695 \times RD} \\
SV &= NR
\end{align*}
\]

3.3. Product’s Reputation Based on Time Measures

Based on the evaluation model of the product’s success, and considering the time measurement, we apply graphs to describe and predict the reputation trend of the three products. So we select the ARIMA [8] (Autoregressive Integrated Moving Average Model) model. Afterward, we can get graphs of the reputation of the three products over time.

4. Results

4.1. Product’s Success Evaluation

(1) Hairdryer

| Table 1. Hairdryer’s Success Evaluation |
|----------------------------------------|
| Contribution Ratio | Cumulative Contribution Ratio |
| Ratings(RA) | 68.6564% | 68.6564% |
| Review(RE) | 22.2538% | 90.9102% |
| Sales Volume(SV) | 9.0898% | 100% |

A comprehensive evaluation model(The most informative data measures) of hairdryer products based on ratings and reviews:

\[
PSE = 0.6866RA + 0.2225RE
\]

(2) Microwave

| Table 2. Microwave’s Success Evaluation |
|----------------------------------------|
| Contribution Ratio | Cumulative Contribution Ratio |
| Ratings(RA) | 68.1464% | 68.1464% |
| Review(RE) | 25.2496% | 93.396% |
| Sales Volume(SV) | 6.604% | 100% |

A comprehensive evaluation model(The most informative data measures) of microwave products based on ratings and reviews:

\[
PSE = 0.6815RA + 0.2525RE
\]

(3) Pacifier

| Table 3. Pacifier’s Success Evaluation |
|----------------------------------------|
| Contribution Ratio | Cumulative Contribution Ratio |
| Ratings(RA) | 64.2416% | 64.2416% |
| Review(RE) | 26.9018% | 91.1434% |
| Sales Volume(SV) | 8.8566% | 100% |

A comprehensive evaluation model(The most informative data measures) of pacifier products based on ratings and reviews:

\[
PSE = 0.6424RA + 0.2690RE
\]
4.2. The Trend of Products’ Reputation

(1) Hairdryer

By calculating the autocorrelation function and partial correlation function, it is determined to take \( d = 0 \). And then, we utilize AIC and BIC [9] criteria to determine the order, take ARIMA \((1, 1, 1)\). The residual analysis of hairdryer’s model is shown in figure 3, and the reputation trend of hairdryer (within the red line is 95% confidence interval, the black line demonstrates predictive value) is shown in figure 4.

![Figure 3. The residual analysis of hairdryer’s model](image)

![Figure 4. The reputation trend of hairdryer](image)

As can be seen from the graph, the prediction curve of hairdryer's reputation will be stable over time, which indicates that hairdryer’s reputation is in a stable state [10].
(2) Microwave

By calculating the autocorrelation function and partial correlation function, it is determined to take $d = 0$. And then, we utilize AIC and BIC criteria to determine the order, take ARIMA(2, 0, 2). The residual analysis of microwave’s model is shown in figure 5, and the reputation trend of microwave is shown in figure 6.

As can be seen from the graph, the prediction curve of microwave's reputation will be rising with small fluctuations over time, which indicates that microwave's reputation is in a rising state.
(3) Pacifier

By calculating the autocorrelation function and partial correlation function, it is determined to $d = 0$. And then, we utilize AIC and BIC criteria to determine the order, take ARIMA($3, 0, 3$). The residual analysis of pacifier’s model is shown in figure 7, and the reputation trend of pacifier is shown in figure 8.

As can be seen from the graph, the prediction curve of pacifier's reputation will be dropping with small fluctuations [11] over time, which indicates that pacifier's reputation is in a dropping state.
5. Conclusions
From the analysis of the three product data files, we can find that the pacifier market is gradually saturated, and customers generally do not like the existing pacifier products and satisfaction will decline. The reputation of the existing pacifier products is positive and will be in a state of decline (although accompanied by certain fluctuations). Reversely, there is a large market for microwave, and the customers’ demand for microwave is increasing exponentially year by year, and the satisfaction is the same. The reputation of the existing microwave products is positive and will be in a rising state (there is a small fluctuation in the early stage and a steady rise in the later stage). In the hair dryer market, although the market competition is fierce, customer satisfaction is increasing year by year. And the product reputation is and will be in a stable state. Therefore, we suggest that Sunshine Company can increase investment in the microwave market, maintain the original hair dryer sales strategy, reduce the input products in the pacifier market and speed up the improvement of pacifier products.

References
[1] Chen, Yuanlin, et al. "Analysis of Review Helpfulness Based on Consumer Perspective." *Tsinghua Science and Technology* 020.003 (2015): 293-305.
[2] P. Lak and O. Turetken, "Star Ratings versus Sentiment Analysis -- A Comparison of Explicit and Implicit Measures of Opinions," in *2014 47th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa, HI, USA, 2014 pp. 796-805.
[3] Chua, Alton Y. K., and S. Banerjee. "Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth." *Journal of the Association for Information Science and Technology* 66.2 (2015): 354-362.
[4] Liu, Zhiwei & Park, Sangwon. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*. 47. 140–151. 10.1016/j.tourman.2014.09.020.
[5] Hu, Ya Han, and K. Chen. "Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings." *International Journal of Information Management* 36.6 (2016): 929-944.
[6] Huang, Albert H., et al. "A study of factors that contribute to online review helpfulness." *Computers in Human Behavior* 48.jul. (2015): 17-27.
[7] Korfiatis, Nikolaos, Elena García-Bariocanal, and Salvador Sánchez-Alonso. "Evaluating content quality and helpfulness of online product reviews: The inter-play of review helpfulness vs. review content." *Electronic Commerce Research and Applications* 11.3 (2012): 205--217.
[8] Zhou, Shasha, and B. Guo. "The Interactive Effect of Review Rating and Text Sentiment on Review Helpfulness." *International Conference on Electronic Com-merce and Web Technologies* Springer International Publishing, 2015.
[9] Guo, Yue, S. J. Barnes , and Q. Jia . "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation." *Tourism management* 59.APR. (2017): 467-483.
[10] Gan, Qiwei, and Y. Yu. "Restaurant rating: Industrial standard and word-of-mouth - A text mining and multi-dimensional sentiment analysis." 2015.
[11] Alves, André Luiz Firmino, et al. "A Comparison of SVM Versus Naïve- Bayes Techniques for Sentiment Analysis in Tweets: A Case Study with the 2013 FIFA Confederations Cup." 2014.