User Evaluation Sentiment Analysis Model Based on Machine Learning

Zhixiang Lu*
School of Mathematics and Science,
Leshan Normal University
Department of Applied Statistics
Leshan, China
luzhixiang1998@gmail.com

Yanhong Chen
School of Educational Science,
Leshan Normal University
Department of psychology
Leshan, China
Zoe199810@foxmail.com

***Abstract*—In order to obtain a correlation between the online evaluation of three online sales products provided by Amazon and their corresponding star rating, we use the tf-idf algorithm for text mining on three online product evaluations, extract each product based on the AFINN sentiment dictionary of sentiment words with tf-idf values included in consumer reviews, and add their intersections into a thesaurus. We also extract the brand information implied in the information and observe a certain difference in the consumer evaluation of each brand. Based on our established thesaurus, we conduct a sentiment analysis of consumer text reviews, assign corresponding sentiment scores, and obtain the corresponding review score variables. Then, we find a significant correlation between the consumer sentiment score and its corresponding star rating. Based on the star rating, verified purchase, and helpful votes variables in the data, we use machine learning models to learn the sample characteristics of each star rating. The final three-product decision tree model has an accuracy of 78.33%. The accuracy rate of the 5-star evaluation is 99.41%, which can better distinguish the extreme evaluations of consumers to a certain extent. An evaluation model based on product brand, consumer star rating, and market share of product brand is established to predict the potential success or failure of the product to a certain extent.**

Keywords-component; text mining; sentiment analysis; machine learning; correlation analysis

I. INTRODUCTION

With the vigorous development of e-commerce, online consumer reviews have become a feedback mechanism that can help producers and businesses improve product quality, improve their own services, and thereby increase market competitiveness. For consumers, through the integration of positive and negative evaluations of product features, consumers who have purchased the product can intuitively obtain the degree of satisfaction with all aspects of the product, which helps them quickly make correct purchase decisions. However, the rapid increase in the number of online reviews and the complexity of the content have made it difficult to obtain information in user reviews. Therefore, this project uses a machine learning model to automatically extract the information that can be used in product evaluations to provide consumers and businesses with more efficient scientific information evaluation tool.

We analyzed the existing *star_rating* and *review_date* of the three products from the data publicly obtained from the Amazon e-commerce platform, and found the change rule of the consumer star rating of the corresponding product on the time axis, and the overall situation of different products receiving star ratings. The subsequent text mining was based on *review_headline*, *review_body*, and *product_title* in the data, using the *jieba* package in R language to extract the consumers' text reviews and product attribute characteristics. Based on *tf-idf*, a feature extraction was performed, and a custom AFINN sentiment lexicon was established to fit the reviews by most consumers for each product. Based on the thesaurus, we calculated sentimental scores for all consumer text reviews to obtain the *review_score* variable, and combined it with *help_vote, verified_purchase*, and machine learning to obtain consumer star-rating prediction models for each product. The model is based on 1-star and 5-star rating (extreme evaluation) with a high prediction accuracy (average 84.3%). We also extracted the product characteristics of 1-star and 5-star rated products at the same time, and found the reasons behind consumers rating them as 5 star and 1 star.

Based on our model, companies can explore the production direction of products and the characteristics of successful product launches, along with predicting product satisfaction in online sales. This can increase the possibility of gaining a better word-of-mouth for their products and promoting the demand for online sales.

II. DATA PREPROCESSING

A. Missing and Outlier Processin

Data preprocessing can reduce the errors in data analysis. First, the corresponding missing values of the three products are processed. Because there are fewer missing values, we choose to remove all of them (where microwave has no missing values, hair dryer has 9 missing values, and pacifier has 93 missing values). At the same time, there are many unreasonable outliers in the pacifier dataset. For example, there are less than 0 and greater than 5 in the star_rating variable. Because there are fewer outliers, we choose to remove all of them.

B. Text Cleaning

Text preprocessing is a crucial step before word segmentation to improve the performance of subsequent algorithms. Consumers' comments are often mixed with insignificant spaces, punctuations, and prepositions, which interfere with the effectiveness of the algorithm. Therefore, data cleaning, word segmentation, and stop word databases need to be set in the data preprocessing stage.

978-1-7281-6685-8/21/$31.00 ©2021 IEEE
DOI 10.1109/ICMSBDA52494.2021.00135

(1) We converted all comments to lowercase to avoid case effects on word segmentation. For example, "Good" and "good" will be defaulted to two different words by the algorithm.

(2) We removed the spaces and punctuation marks, and kept the numbers in the text. Considering that the numbers may have some reference values, we did not discard them.

(3) Stop words are words that have no practical meaning and have no effect on the meaning of the entire sentence. They are usually high-frequency words, word letters, numbers, and special symbols, such as "the," "is," "at," "on," and "etc." Stop words need to be deleted during the preprocessing stage to prevent them from impacting the subsequent results. We removed 40 stop words such as "and," "if," "is," and "us" (the stop words thesaurus is customized here).

C. Text Segmentation and Thesaurus

Before analyzing the review text, the most important step is segmenting the data accordingly. Whether the segmentation is accurate will affect the quality of the subsequent work. The deepening of text sentiment research has led to the development of advanced word segmentation software and tools. Among them, the Chinese analysis package "jieba" in R language performs word segmentation on review texts with satisfactory results and is widely used in the industry. The specific steps for word segmentation are as follows.

(1) First, use the jieba package of R language to segment all reviews (title, body) of each product.

(2) Filter out words with a frequency of more than 10 in each product.

(3) Calculate the tf-idf value of each product segmentation that satisfies condition (2).

The main idea behind tf-idf is that if a word appears frequently in an article with a high term frequency (TF) and rarely appears in other articles, it is considered to have a good class discrimination ability and suitable for classification. TF indicates how often a term (keyword) appears in the text[1].

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Where $n_{i,j}$ is the number of occurrences of the word in the d_j, and the parent is the sum of the occurrences of all words in the file d_j;

IDF is Inverse Document Frequency. Inverse Document Frequency (IDF): The IDF of a particular word can be calculated by dividing the total number of files by the number of files containing the word, and then taking the log of the quotient. If there are fewer documents containing the term t and the larger the IDF, it means that the term has a good ability to distinguish categories.

$$idf_i = \log \frac{|D|}{1 + \left|\left\{ j : t_i \in d_j \right\}\right|}$$

Where $|D|$ is the total number of words in the corpus and $\left|\left\{ j : t_i \in d_j \right\}\right|$ represents the number of files containing words t_i. To ensure that the denominator is not equal to zero, add one.

TF-IDF is actually: TF×IDF. The high word frequency in a particular file and the low file frequency of the word in the entire file set can generate a high weight TF-IDF. Therefore, TF-IDF tends to filter out common words and retain important words.

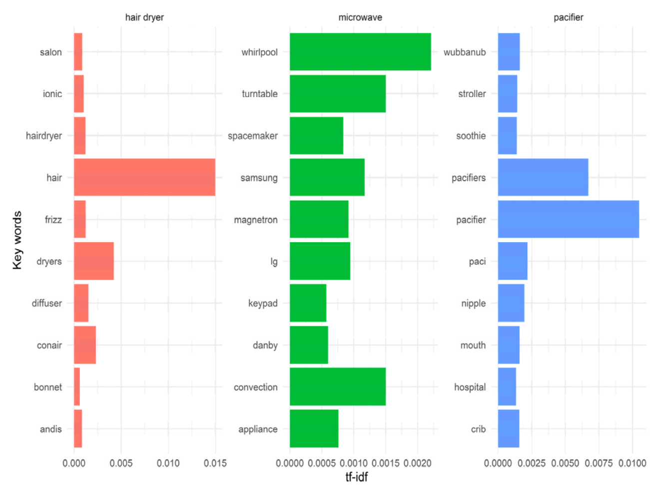

Figure 1. Top 10 words for TF-IDF values of three products

We select the top 10 words with the highest tf-idf value of each product in the consumer *review_headline*. We can see that it contains emotional words such as "loves." We kept the top 100 words with tf-idf values in the *review_headline* and *review_body* of each product, used the AFINN sentiment lexicon to filter out all sentiment words in it, and established a comprehensive sentiment lexicon containing both optimistic and negative sentiment attitudes. The comprehensive sentiment lexicon we finally built is the intersection of sentiment words with three products appearing more than 10 times and having a tf-idf value ranking within 100.

III. Star Rating and Sentiment Analysis

A. Star Rating Analysis of Three Products

First, the star ratings in the data are extracted, and a star-rating analysis is performed on three different products so that the merchant can intuitively see the consumer evaluation to develop choices and strategies for online sales.

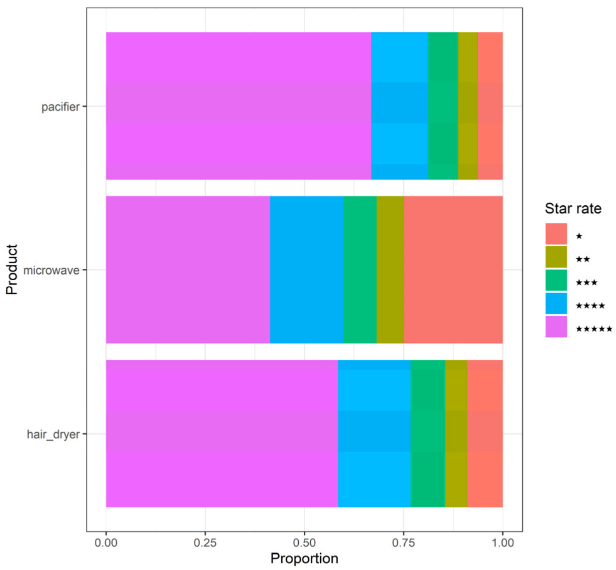

Figure 2. Star-rating ratio chart of each product

- As can be seen from the above chart, pacifier sales are better than the other two products, and microwave sales are the worst. This may be related to the nature and price of the product. Pacifiers are a necessity and consumable. Its usage rate is higher than the other two products. Therefore, its purchase rate is also higher than the other two products. Microwave oven is more expensive. It is not a daily necessity, and therefore, the sales volume is low.
- If 4 stars and 5 stars are used as a standard for praise, the best rate of hair dryer is 86.7%, followed by the pacifier with the rate of praise being 81%, followed by the lowest rate of the microwave oven, which is only 59.8%. Note that the 1-star ratings of microwave ovens reach 24.89%, indicating they are far from the buyer's expectations upon arrival. Such a bad review may also affect the online sales of microwave ovens.

B. Consumer Sentiment Analysis of Each Product Categories

In data preprocessing, emotional information in the comments on the three products has been mined to establish an emotional thesaurus. Based on this, different words are assigned different weights, and a sentiment analysis is performed on the text reviews of the three different products.

If the consumer reviews contain positive sentiment words, one point will be added. If they contain negative sentiment words, one point will be subtracted. The final score is the total of positive and negative scores. The initial score is 0. If the final score is greater than 0, it is recognized as a positive review, while scores less than 0 are considered as negative reviews, and scores equal to 0 are considered as neutral reviews[3].

After an extensive analysis, we found that most users in the *review_headline* have already indicated their satisfaction with the product; therefore, we prefer to use the review title for the emotional evaluation analysis.

TABLE I. EMOTION THESAURUS EMPOWERMENT

Product	Negative	Neutral	Positive
microwave	149(9.2%)	890(55.1%)	576(35.7%)
hair_dryer	284(2.5%)	5866(51.2%)	5312(46.3%)
pacifier	315(1.7%)	9730(51.6%)	8816(46.7%)

It can be found from the chart that based on the *review_headline* consumer sentiment analysis, the results are Positive and Negative are basically the same as the previous analysis of consumer star ratings. From the 5-star and 1-star ratings as a reference, microwave also has the least praise and the most negative reviews, the pacifier's praise in sentiment analysis is slightly higher than that of hair dryer, which is basically consistent with the star rating analysis (the star rating is based on 4 stars and 5 stars), and the reason for the slight deviation may be the emotional thesaurus is not perfect, because the previously established thesaurus is based on the intersection of the top 100 words of the tf-idf value of each product. This is for the sake of robustness and convenience[2], but sacrificing pertinence and accuracy, but from the emotional praise rate judging from the situation, the thesaurus we have established can be very suitable for identifying most emotional words in the evaluation of various products. Based on the comprehensive star rating and the emotional review of the text review, we can get the following information:

1) The star rating and text sentiment evaluation results of the products are basically the same. We can combine the two to make a comprehensive evaluation of the products.

2) Consumers are more willing to buy necessities and consumables online. The daily consumption of pacifiers is 11.68 times that of microwave ovens and 1.65 times that of hair dryers. The reviews of pacifier products also contain more positive emotions than other products, making it easier to obtain praise.

3) The price of the product also may affect the sales of the product to a certain extent. Although the market price and profit price of each product have not been obtained, in general, the microwave oven is the most expensive and the pacifier is the cheapest. In terms of sales volume alone, the cheapest pacifiers generally have the highest sales volume, and the most expensive microwave ovens have the lowest sales volume.

C. Star Rating Prediction Model

TABLE II. VARIABLES DESCRIPTION

Variables	Description
rating_level	*Star_rating* in the original data
title_score	Final score for sentiment analysis evaluation of *review_title* based on thesaurus
body_score	Final score of *review_body* sentiment analysis based on thesaurus
review_score	Sum of sample *title_score* and *body_score*
verified_purchase	Original data (Y:1 N:0)
helpful_votes	Original data

Decision tree is a non-parametric supervised learning method, which is mainly used for classification and regression. The purpose of a decision tree is to construct a model that can predict the value of a target variable by learning a simple decision rule, the IF THEN rule, from the characteristic attributes of the sample data. We choose a conditional inference tree to learn the characteristics of star ratings given by consumers. The conditional inference tree is similar to traditional decision trees, but the selection of variables and segmentation is based on significance testing, not pure progress or the same. We use *rating_level* as the dependent variable and *title_score*, *verified_purchase*, and *helpful_votes* as predictors to build a conditional decision tree model.

In conditional inference tree, regression relations are estimated using a binary recursive partition in a conditional reasoning framework.

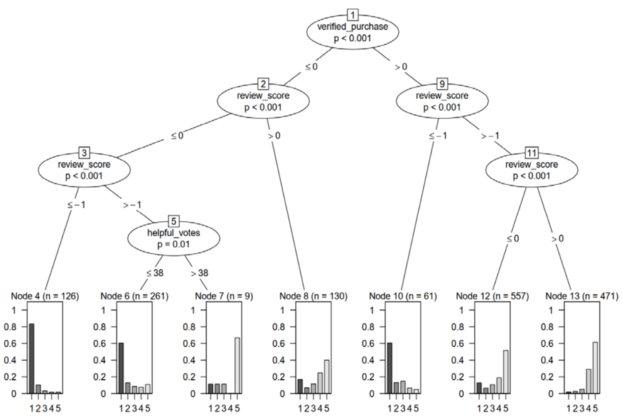

Figure 3. *Conditional inference tree model result graph*

Here, a microwave oven is taken as an example. Through the results of the conditional decision tree model, it can be found that the nodes in the output tree of the decision tree are the corresponding dependent variables and corresponding p values obtained after the significance test, and each p value passes the significant test[4]. The splitting conditions are shown on the left and right branches, and a total of 13 nodes and 7 endpoints are finally divided. Each endpoint contains the number n of samples of different categories, and the probability that the sample belongs to 1 to 5 stars. As can be seen from the figure, when the verified purchase is 0 (that is, N) and the review score is less than -1, the probability that the consumer gives a 1-star bad review is as high as 83.4%; when the verified purchase is 1 (that is, Y When) and the review score is less than -1, the probability that the consumer gives a 1-star negative rating is as high as 73.8% at the same time, which shows that the review score does have great reference value. According to the previously established emotional lexicon, it can indeed be better. Relatively accurate sentiment analysis scores on consumer text evaluations. It is interesting that when the verified purchase is 0 (that is, N) and the review score is greater than -1, if *help_votes* is less than or equal to 38, most (60.8%) consumers will give a 1-star differential rating, and when *help_votes* When it is greater than 38, 67.3% of consumers give 5-star praise instead, indicating that *help_votes* is also a key variable.

The model has good predictions for 1-star and 5-stars, and the discrimination accuracy is 73.13% and 94.90%, respectively. The independent variable here does not include *body_score* and *review_score* which is the result of text sentiment analysis on *body_review*. The conditional decision tree prediction model can better predict the potential of a product's potential success or failure[5]. From this, it can be obtained that based on the combination of text review metrics, star rating metrics, verified purchase, and help votes, consumers have higher prediction accuracy for star ratings that are extreme.

IV. Conclusions

This research analyzes and mines tens of thousands of comments on Amazon e-commerce platform three types of products. Based on the TF-IDF algorithm and the AFINN emotional vocabulary, this project extracts various product-specific emotional words for quantification, and judges reviews based on machine learning. The emotional tendency of the company provides data support for merchants to understand the advantages and disadvantages of products and users' purchase decisions in a timely manner. This has important practical value for consumers, merchants and e-commerce platforms.

A. Weakness

1) Data limitation: When the factors affecting the satisfaction of individual products of microwave ovens, pacifiers, and hair dryers are affected, the variables are limited to the given data. The impact of the same product price on satisfaction (such as cost-effectiveness) has not been further explored. Interference factors such as product endorsements, these factors, such as advertising, seasonal or unexpected events, will also make our forecast data sway, but if the impact is small, the overall trend will not change.

2) Without considering logistics services: In the comments, there will be low star level reviews (poor reviews) caused by delivery service attitude, package damage and logistics delays, which we did not discuss.

3) More powerful thesaurus: Based on the consideration of robustness and efficiency, our thesaurus is based on the intersection of the emotional vocabulary in the top 100 of the tf-idf value of each product as content. If we take the union or select the emotional vocabulary

B. Research Results

1) Consumer goods dominate: Judging from the analysis of the consumer satisfaction of the three products, the sales of pacifiers as daily consumables are the highest (their sales volume is 1167.86% for microwave ovens and 164.55% for hair dryers). If 4 stars and above are used as favorable standards, pacifiers The praise rate of pacifiers is also the highest of the three products, reaching 81.11% (among them, the microwave oven is only 59.88% and the hair dryer is 76.72%).

2) Brand effect: In the product_title of each product, we have found popular brands of each product. For example, the three popular brands danby, sharp and whirlpool among the microwave oven brands are higher than other brands with an average of 3.86, 3.88, and 3.63 stars. And 29.4% of the sales volume of microwave ovens came from *danby*. At the same time, popular brands of hair dryers are *andis*, *babyliss*, *panasonic*, and popular brands of pacifiers are *wubbanub*, *philips*, and *mam*, and 50% of sales are from *Wubbanub*.

Due to the rapid development of information technology, the Internet has brought tremendous changes to the business processes of businesses, and it has also had a profound impact on consumers' decision-making models. Our research has found that nearly half of consumers will refer to product descriptions and other consumers' evaluation information about the product before making a purchase decision. Therefore, Internet product evaluation plays an important role for both merchants and consumers. The comment sentiment analysis technology used in this project, for e-commerce platforms, can analyze user preferences and provide more personalized information through the collection of user age, gender and other information, combined with product features extracted from product reviews. For merchants, it is possible to detect and optimize products in time and accurately in terms of performance, quality, appearance, service, etc., and effectively improve user experience; for consumers, it is possible to avoid spending a lot of time sifting reviews and make more efficient and objective consumer decision-making.

ACKNOWLEDGMENT

This research was financially supported by 2020 Leshan Normal University National College Student Innovation and Entrepreneurship Training Program Fund Project (Project No. 202010649017).

REFERENCES

[1] Peng Xin. (2014). A brief discussion on the analysis and application of book sales data of publishing houses. Publishing Science, 22 (1), 43.

[2] Wang Jian, & Zhang Junni. (2017). Application of Statistical Model in Chinese text Mining. Mathematical Statistics and Management, 36 (4), 609-619

[3] Zhang Jinyan & Guo Pengjiang. (2009). Application of deterministic time Series Model and ARIMA Model (Doctoral dissertation)

[4] Sun Caiyun & Yang Xiaojing. (2008). Establishment and application of product ARIMA model. Journal of North China University of Science and Technology, 5 (2), 85-89

[5] Peng, G., & Yuefen, W. (2016). Identifying optimal topic numbers from Sci-Tech information with DA model. Data Analysis and Knowledge Discovery, 32(9), 42-50.

[6] Kostoff, R. N., Briggs, M. B., Rushenberg, R. L., & Bowles, C. A. (2007). An overview of China's and India's science and technology literature. Science Focus, 2(4), 1-06.

[7] Chang, Y. P., & Wu, C. T. (2019). The Study of Resilience for Educators: A Bibliometric Citation Meta-Analysis. Jiaoyu Yanjiu Yuekan= Journal of Education Research, (300), 79-101.