Academic Journal of Humanities & Social Sciences
ISSN 2616-5783 Vol.3, Issue 3: 21-25, DOI: 10.25236/AJHSS.2020.030303

Research on Big Data of Customer Product Reviews Based on Text Sentiment Extraction and Statistical Analysis

Sicheng Wan[1], Jiajin Tang[2] and Xinyue Yang[3]

[1] *College of plant protection, Southwest University, Chongqing, 400715*
[2] *Hanhong college, Southwest University, Chongqing, 400100*
[3] *School of Geosciences, Southwest University, Chongqing, 400715*

ABSTRACT. *With the rapid development of the Internet and smart phones, in recent years, major global e-commerce service industries have attracted a large number of online sellers. At the same time, most consumers have also transformed traditional mall shopping into a convenient and fashionable online shopping model. When consumers purchase products on the e-commerce platform, most consumers will be able to make purchases and make certain evaluations and star ratings based on their preferences. This not only helps other consumers to have a preliminary understanding of the product before making a purchase, and to make judgments. At the same time, online sellers and e-commerce platforms can conduct background big data analysis and processing based on the indicators evaluated by users, and make a future Forecast the sales volume of goods in the market over time, and upgrade and improve the goods.*

KEYWORDS: *text sentiment extraction, principal component analysis, regression analysis*

1. Introduction

With the rapid development of smart phones and smart networks, in recent years, major e-commerce service industries have gradually increased, attracting many Internet merchants to join [1]. At present, online shopping has largely replaced the traditional mall shopping model [2]. When we select products on the web page, due to the variety of brands and types of products we need, the "comment" and the review information of previous buyers will largely determine our purchase choices. And the evaluation of these products has important guiding significance for suppliers and e-commerce platforms in product improvement, update design, and formulation of future marketing models, and also helps further integration of e-commerce and the market [3].

ISSN 2616-5783 Vol.3, Issue 3: 21-25, DOI: 10.25236/AJHSS.2020.030303

2. Characterization Analysis Forecasting and Evaluation

It should be noted that the review_body and review_date of the purchaser of the product are composed of text data. The text data has a certain degree of likes and dislikes on behalf of the customer in evaluating the product. The expression of this opinion is generally expressed by the linguistic emotion and subjectivity of customer reviews. This kind of data cannot be directly quantified for analysis and use, so we explore a quantification algorithm for text sentiment analysis and subjective and objective expression.

2.1 Text Sentiment Analysis

For the customer review information obtained by Amazon in these three product sales, our team applied Python programming method based on *TextBlob* to extract and analyze customer review information, and normalized the resulting data.

Polarity indicates emotional polarity: its range is between (-1, 1), "-1" means completely negative, and "1" means completely positive.

Subjectivity means subjectivity: its range is between (0, 1), where "0" is completely objective, that is, the author is stating a description of a fact, and 1 is completely subjective, that is, expressing the author's personal opinion.

We can use the data obtained from the above two variables to determine the general judgment of the customer after the purchase, which is convenient for subsequent analysis and processing.

2.2 Principal Component Analysis

2.2.1 Exploratory data analysis

Table 1 Descriptive statistics results report

	Minimum value	Maximum value	Mean	Standard deviation
helpful_votes	0	814	1.5363	11.3272
total_votes	0	848	1.9056	12.2747
polarity	-1	1	0.2724	0.2782
subjectivity	0	1	0.5547	0.2119

We perform preliminary descriptive statistics on the data obtained, such as Table 1. By describing the data reflected in the descriptive statistics table, we can proceed to these four variables.

ISSN 2616-5783 Vol.3, Issue 3: 21-25, DOI: 10.25236/AJHSS.2020.030303

2.2.2 Standardize the data

Extremely normalize the analysis indicators.

$$X'_{ij} = \frac{X_{ij} - \min X_{ij}}{\max X_j - \min X_j}. \quad (1)$$

In the mathematical expression: The j-th index of X'_{ij} represents the normalized result of the i-th data, and X_{ij} is the i-th data of the j-th index; $\max X_j$ and $\min X_j$ are the maximum and the minimum value.

2.2.3 Principal component analysis

KMO and Bartlett Test: There is a certain correlation between KMO values equal to 0.5, which is suitable for principal component analysis.

Variance Explained: It can be found that the two components with eigenvalues greater than one explained 86.886% of the original index, so the above index can use two main components to express the information reflected by the original index.

We have consulted the relevant literature: Generally speaking, indexes larger than 0.7 in the rotation component matrix are extracted as constituent indexes corresponding to the principal component indexes. According to the index values in the above matrix, we classify helpful_votes and total_votes as component one and name them (C_1). Subjectivity and polarity, these two data are obtained from the same data source-text comments, have a strong correlation, classified as component two, and named (C_2).

Component Score Coefficient: Here we extracted the coefficients in the component score coefficient matrix, combined with the relevant information of the rotated component matrix, and calculated the principal component scores: C_1, C_2.

$$C_k = \sum_{m=1}^{n} a_j X_j. \quad (2)$$

In this equation, C_k is the score of the k-th principal component. The principal component is composed of (m can take values 1, 2 .... j .... n), and a_j is the j-th index corresponding to Score, X_j is the data corresponding to the j-th index.

2.3 Construction of comprehensive indicator functions

We standardized the data of the star-rating in the data file so that the standardized star_rating is between (-1,1), where a negative value represents a negative review given by the user to the product, and 0 represents a user given the product Medium reviews. Positive values represent positive reviews given by users. The trend of each value to the positive and negative critical values represents the

ISSN 2616-5783 Vol.3, Issue 3: 21-25, DOI: 10.25236/AJHSS.2020.030303

trueness of the customer's preference for the product being reviewed. We combine C_1, C_2 and rating to give the following functional relationship:

$$F\left(C_1, C_2, s\right) = s * C_1 + C_2. \tag{3}$$

We divided the time sequence in the file by month, and averaged the reputation value in the unit month with the number of votes to get the remaining average value in the month and normalized it.

2.4 Time Series Auto regressive Analysis

Whether a series of specific star ratings will have a certain impact on subsequent evaluations

We integrate the data by month into the monthly node data to evaluate the star rating by month. If the average star rating within a month is relatively low \ high, it shows that a specific series of positive or negative reviews appeared this month. In a period of time, thewill definitely not be all positive or negative reviews. There will always be negative reviews, and vice versa. And averaging can better reflect the star rating of this month.

In this regard, we performed a time series analysis on the review data C_2 and the star data s. It can be seen that if the star rating trend declines or rises within a certain period of time, it will be more obvious in the trend of reviews. The time-varying curve of the two matches our conjecture: after a series of specific star ratings, the comment curve will fluctuate correspondingly with the changing trend of star ratings. Therefore, we believe that a series of specific star ratings will have a certain impact on subsequent evaluations.

Then we guessed that the comment content has strong expressive power and detailed expression ability. We guess whether the reviews with stronger emotions or more extreme subjectivity are related to customer evaluation.

3. Conclusion

The helpful_votes and tota_votes among the data have a strong correlation with the relevant scores after our customer sentiment analysis is extracted. So the company should focus on the relationship between the two variables when performing statistical analysis on the big data of product sales. Among many reviews, the Helpful_votes is filtered to check a larger number of reviews, and update and improve products reasonably according to the relevant content.

ISSN 2616-5783 Vol.3, Issue 3: 21-25, DOI: 10.25236/AJHSS.2020.030303

References

[1] Li Shuyan. Research on the Impact Mechanism of Mobile Shopping Comprehensive Consumption Experience on Re-purchase Intention [D]. Huaqiao University, 2016.

[2] Li Jinhai. Research on Online Shopping Hybrid Recommendation Model and Strategy Based on Online Review Mining [D]. Jiangsu University, 2016.

[3] Lynette Shi, Liang Xun, Sun Xiaolei. Study on Appraiser Utility Mechanism Based on Online Rating and Comments [J]. China Management Science, 2016, 24 (05): 149-157.