Academic Journal of Humanities & Social Sciences
ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

Text mining and decision-making analysis of E-commerce Review based on R language

Xinyue Zhang, Hong Guo

Institute of economics Hebei University, Baoding City Heibei Province 071000

ABSTRACT. *We constructed a set of programming language system for text comment data mining and quantifying score, and put forward a text comment weighted synthetic score decision model —— Fussy and comprehensive evaluation model based on analytic hierarchy process(AHP). It can help companies maximize profits and give consumers a better shopping experience.* ***First****, we cleaned and processed the data, which was a huge and critical step. We got a comprehensive score for each review text by word segmentation, keyword filtering, and generic word analysis. In this process, we built a TF-IDF model and the output of the model can easily determine whether a word is a keyword or not. Based on this, a product star keyword database was built, and word clouds were generated to link up the following generic word analysis and output the product text review score results.* ***Then****, we proposed a model of fussy and comprehensive evaluation based on analytic hierarchy process. Since the boundary between good and bad of three products is not clear, it is difficult to classify them into a certain category. So we first used analytic hierarchy process to calculate the weight of evaluation, evaluation star, product star and then made a comprehensive evaluation of all the factors. According to the subjection degree theory of the mold mathematics, the comprehensive evaluation method transforms qualitative evaluation into quantitative evaluation, gives three products the evaluation and judges which star level they belong to.* ***Next****, we analyzed the sensitivity of the model and tried to find the direction of future improvement and development.* ***Finally****, according to our analysis, baby pacifiers and hair dryers are rated five stars on a comprehensive review. The manager can adjust the sales ratio of each product appropriately and improve the shortcomings of each product.*

KEYWORDS: *E-commerce review, Text mining, R language, Comment quantitative evaluation model*

1. Introduction

1.1 Statement of the problem

With the development of Internet e-commerce, data mining which is a high-tech

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

information product has become an inevitable trend. Under the background of "big data" era, data mining has rich practical value in enterprise marketing management. Powerful databases and data mining have also been a mainstay of e-commerce giants' personalizing marketing and precision services.

Social Consumer groups for the purchase of goods are also more inclined to the Internet channels, convenient, fast, labor-saving shopping experience is what online shopping consumers expect. Amazon, one of the e-commerce giants, has designed a rating system for consumers in its online shopping mall.Users can express their shopping experience and satisfaction according to the "stars" and comments. This will gather a lot of data. Therefore, how to deal with the data and mining the information in the data to judge different consumer preferences, explore the potential consumption are the key points in this article.

Sunshine Company plans to sell three new products online: microwave ovens, baby pacifiers and hair dryers. TF-IDF statistical analysis method is used for text data processing. And a model system is designed for analyzing and processing user rating and evaluation, aiming at enhancing product satisfaction, enhancing user experience, and bringing higher profit return to the company.

1.2Assumptions

- Assume that the only factors that affect the overall product rating are reviews, product stars, and review stars.
- In the process of data calculation, if the error is within a reasonable range, the effect on the result of data can be ignored.
- Suppose the impact on the overall product score: Review >Product Star> Review Star.
- In establishing the time series, it is assumed that the reviews within a few years of the initial few reviews have no reference value to the product score and can be omitted.
- Assume helpful reviews / total reviews:

Table 1: Relationship between proportion and star rating

Proportion	Star rating
proportion≤0.2	1-star
0.2<proportion≤0.4	2-star
0.4<proportion≤0.6	3-star
0.6<proportion≤0.8	4-star
0.8<proportion≤1	5-star

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

1.3 Notations

Table 2: Model inputs and symbols

Symbols	Definition
$tf_{i,j}$	number of occurrences of i in j
df_i	number of documents containing i
N	total number of documents
U	factor set
V	evaluation set
r_{ij}	the i element in factor set U to the j element in evaluation set V
R_{m*n}	the fuzzy and comprehensive evaluation matrix
A	weight of judging factors
P	judgment matrix
$\lambda_{\max}$	the maximum characteristic root
CI	the consistency index of matrix
RI	randomness index
B	fuzzy evaluation conclusion set

2. Indicator analysis and data cleaning

Every rigorous statistical analysis of big data starts with the collection, cleaning, sifting, and integration of large amounts of data, and we do a lot of work with that data over time. The data provided includes:marketplace, customer_id, star_rating, vine, verified_purchase and other data sets, which cover a wide range of issues including unnatural variables, structured data, ratings, and emotive text data. Calibration is the premise of model fitting analysis.

2.1 Data sources and indicator analysis

Sunshine company provides us with data from amazon.com, which provides real time shopping records and ratings, reviews and descriptions of microwave ovens, baby pacifiers, and hair dryers in the official back office for a specified period of time.

We took a deep look at the Green Label Project and applied its ideas to the metrics analysis in this article. "Green Label" is Amazon official evaluation system which is a high authority. Comments come in a green color: Vine customer Revier of free Product, an independent opinion of a Vine voices member, not influenced, changed, or edited by the seller. Amazon will also not delete Vine green marked

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

reviews, which will provide a true reference for customers who want to browse or buy later.

2.2 Data cleaning and processing

2.2.1 Data screening

We sifted through the microwave, pacifier, and hair dryer data as follows: First, we dealt with verified_purchase. We filtered out the "Y" value to ensure that the reviews were the first-hand experience of the buyer. And we did a careful screening according to the requirements.

2.2.2 Data grouping

The data were sorted by comment star, and the selected comments were divided into 5 groups from 1 star to 5 stars. And then we imported them into .csv file.

2.2.3 Word segmentation

Use R3.6.3 to read the data, load the English stop dictionary, and then run the data using the code program, the flow chart is as follows:

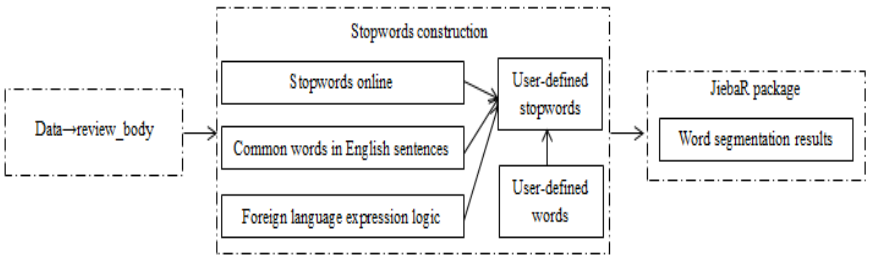

Figure 1: Word segmentation flowchart

2.2.4 TF-IDF keyword search

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$$TF - IDF = TF \times IDF$$

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

$$tf_{i,j} = \text{number of occurrences of i in j}$$

$$df_i = \text{number of documents containing i}$$

$$N = \text{total number of documents}$$

According to the formula, the more times a word appears in a document, the greater its TF value, and the fewer documents that contain a word in the whole documents, the greater the IDF value. So the higher the TF-IDF value of a word, the greater the probability that the word is the keyword.

The disadvantage of TF-IDF Keyword Extraction Algorithm is that in order to extract the keywords in a document accurately, a whole package is needed to support it. In order to solve this problem and make the construction of keyword database more perfect, the IDF value of all words is calculated in advance in JiebaR package, which has certain effect for document keyword extraction.

Figure 2: Microwave word cloud of Star1 and Star5

2.2.5 Word cloud results

On the left side of the Figure2 is the word cloud from one-star review of microwave oven. The words “buy”, “service” and so on appear frequently, which shows that consumers are less satisfied with the product price or after-sales service. The five-star evaluation of the key words are “easy”, “perfect”, and so on.That means consumers are more satisfied with the ease of use and size of the product and at the same price this product fits the consumer satisfaction.

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

Figure 3: Baby pacifier word cloud of Star1 and Star4

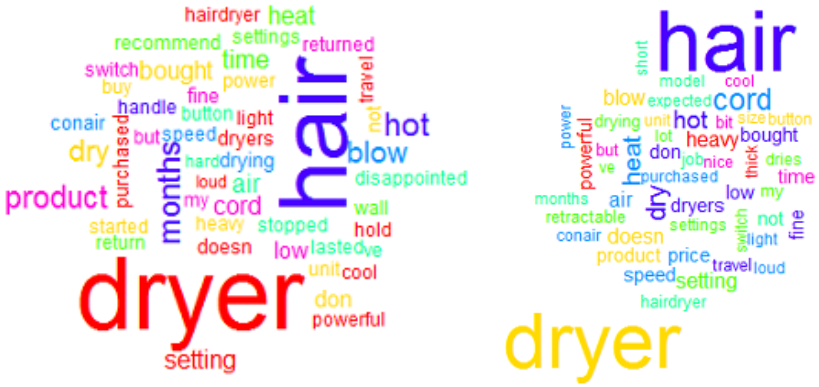

Figure 4: Hair dryer word cloud of Star2 and Star3

As can be seen from the above three sets of pictures, the results of the data processing are directly reflected in the word clouds. The more frequently the words appear, the larger the area they occupy. These words are more representative of the popularity of the comments corresponding to the star.

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

3. Model construction

3.1 Innovation point

Online shoppers have different definitions of what is good and what is bad. Most consumers will give a lower star rating unless they are particularly dissatisfied with a product, but in most cases they will give a favorable rating. Online commodity trading is an electronic currency transaction based on the trust between buyers and sellers. Once a merchant's comments quality is low due to non-human reasons, it will bring economic losses or poor consumer experience to consumers and will bring many difficulties to the company's development and the later decision-making implementation. Therefore, we set up the R language text comment common word screening and comparison program for Sunshine, and built the model with the combination of fuzzy comprehensive evaluation and AHP.

3.2 Text quantitative analysis

To a certain extent, the analysis of customer evaluation emotion can be understood as the classification of customer evaluation information. The other two products can be treated in the same way and the results are as follows.

First, through text pre-processing, 1615 comments of microwave oven products are sorted by star rank, and the 50 words with the highest frequency under the star rank are used as the key word library of the star rank. Then five new microwave star rank key word files are generated, from one to five stars.

The original microwave oven product comment content is imported into CSV file to get the key word for the comment.

```
 [1] "purchased"    "item"           "reasons"     "Union"
 [5] "Made"         "assembled"      "TN"          "American"
 [9] "Ive"          "convection"     "oven"        "double"
[13] "oven"         "styling"        "unit"        "nice"
[17] "sleek"        "overcomplicated" "enhances"   "kitchen"
[21] "cooking"      "simple"         "built"       "functions"
[25] "thawing"      "frozen"         "items"       "popping"
[29] "popcorn"      "You"            "punch"       "button"
[33] "steaks"       "defrosting"     "popcorn"     "popping"
```

Figure 5: Microwave oven comment keywords sample diagram

Then the results were compared with five new star keyword files, and the analysis of common terms was carried out. The results showed which of the five star keyword files had a better fit between a comment of the product and the Star Keyword file, and the more common terms there were, a quantifiable score for the content of the review.

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

Table 3: Sample table of quantitative scoring of commentary text

star / comment	star 1	star 2	star 3	star 4	star 5
1477	7	8	4	5	7
1480	7	8	11	10	7
1492	6	5	8	5	6

Using similarity text analysis, we can largely restore the critics' emotional attitude at that time, and form a good quantitative score for the satisfaction of goods' distribution, appearance, quality and use experience. As shown in the table above, there are 7,8,11,10,7 common words respectively in the 1480 comments of microwave products and the 1-5 star keyword library. Therefore, it can be judged that the text of this article is the most common words of three-star comments. According to the above ideas we can have more intuitive reflection of the authenticity of the review compared with the star analysis of the situation. It can be helpful for the business in the development of sales plans.Finally, emotion analysis was used to synthesize the results of star rating.

3.3 Model Components

3.3.1 Fuzzy and comprehensive evaluation method

Factor set U={comment,rating star,product star} $U = \{u_1, u_2, u_3\}$

Evaluation set V={star1,star2,star3,star4,star5} $V=\{v_1, v_2, v_3, v_4, v_5\}$

If the membership degree of the first element in factor set U to the first element in evaluation set V is r_{i1}, the result of a single-factor evaluation of element i is $R_i = (r_{i1}, r_{i2}, ..., r_{in})$ .The fuzzy and comprehensive evaluation matrix is R_{m*n} .

$$R_{3*5} = \begin{Bmatrix} r_{11}, r_{12}, r_{13}, r_{14}, r_{15} \\ r_{21}, r_{22}, r_{23}, r_{24}, r_{25} \\ r_{31}, r_{32}, r_{33}, r_{34}, r_{35} \end{Bmatrix}$$

The Matrix of the three types of products is as follows:

$$\text{Microwave oven:} \begin{Bmatrix} 0.2579 & 0.1476 & 0.2103 & 0.2943 & 0.0978 \\ 0.1876 & 0.0327 & 0.1131 & 0.1458 & 0.5298 \\ 0.2489 & 0.0693 & 0.0830 & 0.1858 & 0.4130 \end{Bmatrix}$$

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

Baby pacifier: $\begin{Bmatrix} 0.1327 & 0.0493 & 0.1545 & 0.2953 & 0.3682 \\ 0.2866 & 0.0305 & 0.1319 & 0.0961 & 0.4549 \\ 0.0547 & 0.0482 & 0.0760 & 0.1439 & 0.6772 \end{Bmatrix}$

Hair dryer: $\begin{Bmatrix} 0.1346 & 0.1091 & 0.2819 & 0.2966 & 0.1783 \\ 0.1608 & 0.0178 & 0.0890 & 0.1113 & 0.6210 \\ 0.0866 & 0.0528 & 0.0862 & 0.1817 & 0.5927 \end{Bmatrix}$

3.3.2 Analytic hierarchy process

Weight of judging factors A={ a_1, a_2, a_3 }

Table 4: The weight of evaluation factors was obtained by AHP

Scale	Implications
1	Indicates that the two elements are of equal importance
3	Indicates that the former is slightly more important than the latter
5	Indicates that the former is significantly more important than the latter
2,4	Represents the intermediate value of the above adjacency judgment
reciprocal	If the ratio of importance of elements i and j is a_{ij}, then the ratio of importance of elements j and i is $1/a_{ij}$.

We consider that the evaluation is obviously more important than the star rating, and slightly more important than the product star rating and then we get a judgment matrix: $P = \begin{Bmatrix} 1 & 5 & 3 \\ 0.200 & 1 & 0.333 \\ 0.333 & 3 & 1 \end{Bmatrix}$

Find the maximum characteristic root $\lambda_{\max}$ and eigenvectors.

Step one: Column vector normalization

Step two: Summation by line $w = \begin{Bmatrix} 1.900 \\ 0.318 \\ 0.781 \end{Bmatrix}$

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

Step three: Normalization $\overline{w} = \begin{Bmatrix} 0.630 \\ 0.106 \\ 0.264 \end{Bmatrix}$

Step four: P $\overline{w} = \begin{Bmatrix} 1.952 \\ 0.320 \\ 0.792 \end{Bmatrix}$

Step five: $\lambda_{\max} = \sum_{i=1}^{3} \frac{\left(P\overline{w}\right)_i}{n\overline{w}_i}$ =3.039

Step six:CI= $\frac{\lambda_{\max} - n}{n-1}$,n=3,CI=0.02. CI is the consistency index of matrix, the random consistency index RI is introduced which can be obtained by looking up the following table.

Table 5: The random consistency index RI

n	1	2	3	4	5	6	7
RI	0	0	0.58	0.90	1.12	1.24	1.32

Step seven: $CR = \frac{CI}{RI} = \frac{0.02}{0.58} = 0.034$,CR<0.1. Therefore, the judgment matrix has satisfactory consistency. Weight A={0.630,0.106,0.264}.

Step eight:B=A*R= $\{b_{11}, b_{12}, b_{13}\}$,The elements in B represent the degree of membership for each rating level.Final evaluation X= $\max\{b_{11}, b_{12}, b_{13}, b_{14}, b_{15}\}$.

Table 6: Membership of different products

	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}
Microwave	0.2471	0.1147	0.1664	0.2499	0.2268
Baby pacifier	0.1284	0.0470	0.1314	0.2342	0.4590
Hair dryer	0.1247	0.0846	0.2098	0.2466	0.3346

As can be seen from the table above, the total score of the microwave oven on the 4-star membership degree is the highest, so the total evaluation can be seen on 4-star. Baby pacifier on the 5-star membership degree is the highest, so the total evaluation can be seen on the 5-star.Hair dryer on the 5-star membership degree is the highest, so its overall evaluation can be seen as about 5 stars.

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

4. Model Analysis

4.1 Sensitivity analysis

In order to verify the rationality of the model, we analyze the sensitivity of fuzzy and comprehensive evaluation model and the analytic hierarchy process model. For the mold and composite evaluation model, we can appropriately change the factor set of the weight of the three variables. We change A={0.630,0.106,0.264} to A=

{0.580,0.106,0.314}.This means increasing the star weight by 5% , reducing the weight of the comments by 5% , and then recalculating the integrated comments set B=A*R= $\{b_{11}, b_{12}, b_{13}\}$. Next, normalize. Last, evaluate X= $\max\{b_{11}, b_{12}, b_{13}, b_{14}, b_{15}\}$

Table 7: Membership of different products

	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}
Microwave	0.2447	0.1108	0.1600	0.2465	0.2426
Baby pacifier	0.1245	0.0470	0.1275	0.2266	0.4744
Hair dryer	0.1223	0.0817	0.2000	0.2409	0.3553

As can be seen from the table above, although the membership for each star has changed slightly, the final result has not changed, such as the total rating of the microwave oven can still be regarded as 4 stars. This indicates that the sensitivity of the model is low.

For the AHP model, we can do a simple sensitivity analysis for the microwave oven. First, calculate the average total value of the microwave oven with the original weight.

$$s = 0.630 * u_1 + 0.105 * u_2 + 0.264 * u_3 \quad \text{s=3.4281}$$

Then the total value of the microwave oven is recalculated by changing the good weight.

$$s = 0.580 * u_1 + 0.105 * u_2 + 0.314 * u_3 \quad \text{s=3.5375}$$

The total evaluation value has a certain change, but the change is small, which shows that the sensitivity of the model is not high.

4.2 Potential improvements

First of all, the treatment of text comment is not careful enough. Using TF-IDF model for keyword selection, word bag, co-word analysis of the emotional score is not very accurate. LDA can be used for sentiment analysis to improve the accuracy of the final result.

ISSN 2616-5783 Vol.3, Issue 3: 52-63, DOI: 10.25236/AJHSS.2020.030307

Next, for the time series model, we only used the simple model to do the fitting between the total score and the time. And the English sample data is too big, therefore, the time unit which chooses is also quite big, this will affect the model result. We can apply more advanced models stochastic point process (especially hawkes process).We can treat rating as stochastic event for better results.

5. Conclusion

Looking at the full text, the application of big data mining in the process of enterprise development and operation management has very important value.

Companies should correctly understand the status of text processing and comprehensive evaluation in the future. In order to improve the ability to collect and analyze data, make the decisions of the company more accurate and efficient, and meet the needs of the company in production, operation, management and other needs. Make the development of the enterprise more suitable with the modern social and economic environment, thereby strengthening the market competitiveness of the enterprise and creating more efficient economic benefits for the companies. I hope that my team's program design and model analysis can provide some help for Sunshine Company.

References

[1] Chao Song, Xiao-Kang Wang, Peng-fei Cheng, Jian-qiang Wang, Lin Li. SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis [J]. Knowledge-Based Systems, 2020.

[2] Bing Liu. Text sentiment analysis based on CBOW model and deep learning in big data environment [J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(2).

[3] Shailendra Kumar Singh,Manoj Kumar Sachan. SentiVerb system: classification of social media text using sentiment analysis [J]. Multimedia Tools and Applications, 2019, 78(22).

[4] Kashfia Sailunaz,Reda Alhajj. Emotion and sentiment analysis from Twitter text [J]. Journal of Computational Science, 2019, 36.