Text quantization model based on TF-IDF and NLTK toolbox

Yidong Hu[1], Haoran Hua[1], Jinzhao Song[1], Huan Zhang[1], Chuheng Sun[2]

[1]School of Information and Communication Engineering, Harbin Engineering University, Harbin, Heilongjiang, 150001

[2]University of Technology Sydney Business Faculty, Sydney, Australia

Keywords: TF, DIF, NLTK, polynomial fitting

Abstract: With the development of Internet technology, more and more people choose online shopping, resulting in a large number of online product reviews. These online reviews convey a wealth of information. Our task is to help Sunshine analyze the online reviews and star ratings of customers of its products, and obtain different measurement methods and customer preferences for different products with different requirements. We performed correlation analysis on reviews, helpfulness ratings, and star ratings. We perform data cleaning based on the strength of the correlation between products, which reduces the sample size. Then, in order to quantify customer reviews, we used TF-IDF to mine high-frequency keywords for different product reviews. After manual screening and expansion, keywords are divided into different topics. Then we use the NLTK toolbox to assign different scores and weights to these topics. Based on these scores and weights, we get a quantitative score of customer online reviews. We also verified the reliability of our scoring criteria by analyzing the correlation between the ratings of online reviews and star ratings. Secondly, we fit the functional relationship between the time of different products and the quantitative score of the text. By further analyzing these functional relationships through MATLB, we have obtained criteria for measuring potential success and failure.

1. Introduction

With the rapid development of the Internet, more and more consumers have chosen online shopping as a fast and convenient way of shopping. Amazon provides customers with personal ratings and textual reviews (comments) to express personal information and opinions about the product, and other customers can submit ratings in these reviews. When consumers buy a new product, their decision in online shopping is paying more and more attention to user evaluation [1]. As a result, companies can use this data to understand the markets in which they participate, including when to participate and product design feature choices [2]. However, in the face of a large number of online reviews, how to filter and use the effective part of it has become a major problem. Based on this situation, we have established models to solve these problems [3].

2. TF, DIF and NLTK Model

When we extract the information in the comments, we need to do the corresponding data mining. Since all data for data mining must be in numeric form, the text exists as a string. Therefore, we need to digitize the string first. TF-IDF is one such method that can convert strings to numbers [4].

The main content of TF-IDF is: if a word or phrase appears frequently in one article with a high TF and rarely appears in other articles (indicating that the IDF is low), the word or phrase is considered to have a good classification Ability, which indicates suitability for classification.

The TF-IDF value of a word pair can be expressed as the following formula. This value indicates the importance of a word pair for classification. The larger the value, the better the classification is. The smaller the value, the worse the classification is.

$$tf\_idf_i = tf_i * idf_i \tag{1}$$

2.1 TF

In a given text, TF refers to how often a word appears in the text. This number is a normalization of the term count to prevent it from leaning towards documents that are too long. The same word may have a higher number of words in a long file than a short file, regardless of whether the word is important or not. For a vocabulary in a particular text, its importance can be expressed as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

2.2 DIF

IDF is a universal measure of vocabulary importance. The IDF of a specific word can be divided by the total number of files and the number of texts containing the word, and then the logarithm of the quotient obtained is:

$$tf_{idf_i} = tf_i * idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{3}$$

2.3 NLTK

Natural language processing is an important direction in the field of computer science and artificial intelligence. NLTK (Natural Language Toolkit) is a Python-based class library, which contains rich text processing and text mining APIs. It is also currently the most popular natural language programming and development tool. In the research and application of natural language processing, proper use of the functions provided in NLTK can greatly improve efficiency.

2.4 Digitizing text before analysis

Keep all vine online reviews first, and then exclude online reviews from people who haven't purchased the product. In order to deal with the correlation between star ratings, reviews, and help levels, the text of the three products is now digitized: replace Y with 1, and N with 0. Stars are divided into positive, negative, and intermediate ratings: 4, 5 stars are positive, 3 stars are intermediate, and 1, 2 stars are negative. Related research shows that extreme emotions have a greater impact on customers, so 1 is used for positive reviews and 0 is used for bad reviews. Nicky Somohadjo proposed in 'the effect of online reviews on the review attitude and purchase intention' that the length of online text reviews has a certain effect on the effectiveness of reviews. We use Excel to calculate the number of spaces in each comment, and then add one to the result to get the word count of each comment.

2.5 Correlation analysis

Using SPSS software to perform correlation analysis on the quantized text data, the correlation results are shown in the figure:

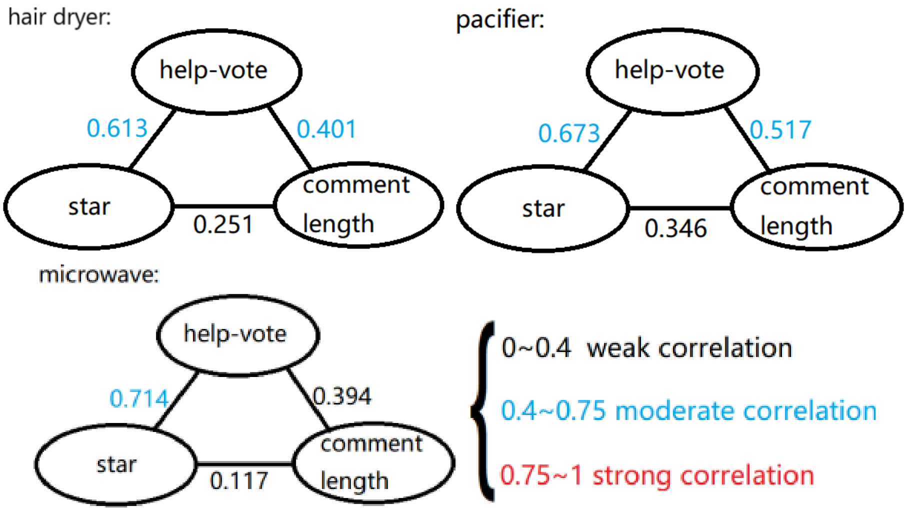


Figure 1. Correlation Analysis

According to the correlation value in the figure, among these three products, the helpful vote is moderately correlated with the star rating and tends to be strongly correlated. That is, the helpful vote can largely reflect the star level. Comment length has a weak correlation with star rating and helpful vote, so the effect of comment length can be ignored. Relevant research shows that comments that are liked as useful comments can better reflect their reliability than stars. Therefore, the data can be further filtered based on whether the comments are liked or not. That is, to exclude online reviews of non-vine customers who did not have a helpful vote, so as to get a smaller sample.

3. Correlation model between online comment score and customer rating

3.1 Topic words

High-frequency words according to their characteristics, deleted some high-frequency words that did not help the analysis problem, and formed several corresponding topics. We have correspondingly expanded the different topics and added some commonly used vocabularies. The obtained topic words and expanded vocabularies are shown below.

Hair-dryer: 1time= [day, week, fortnight, month, year]; 2air= [heat, temperature, puff, blow, speed, velocity]; 3price= [cost]; 4product= [cord, string, rope, quality, charact- er, trait, texture, property, performance, way, us]; 5power= [efficiency].

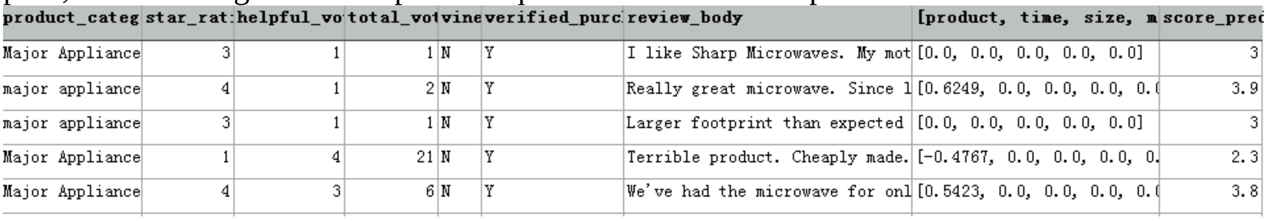
Pacifier: 1time= [day, week, month, year]; 2price= [cost, spend, money, expense]; 3size= [space, measure]; 4mouth= []; 5quality= [character, trait, texture, property, performance].

Microwave: 1space= [size, measure]; 2price= [money, dollar, cost, spend]; 3product= [quality, character, trait, texture, property, performance]; 4food= [heat, daintiness, taste, delicacy]; 5model= [type, category, kind, way, use].

These keywords can largely reflect the customer's focus on the product, which attributes of the product they prefer. Merchants can focus on certain aspects of the product based on the corresponding keywords to cater to consumer preferences and get better sales.

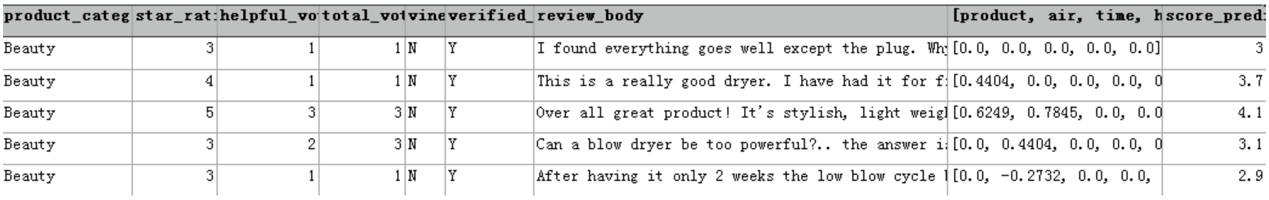
3.2 Establishment of scoring standards

Using the filtered topic words and their expanded words above, we use Python-based NLTK to perform double conjunction processing on specific words. Common collocations include adjectives plus nouns. We use NLTK to find corresponding adjectives corresponding to the topic words and their expanded vocabulary, and use NLTK to assign corresponding values to the topic words and their expanded vocabulary. The corresponding weight vectors are automatically set, and then NLTK uses these words and their corresponding weights to score each comment accordingly. The score range after our calculation is [-1, 1]. In order to better correspond to the star rating, we perform a one-to-one linear range transformation on it, and get the interval as [1, 5] rating interval. The higher the rating of the online review, the higher the customer satisfaction with the product. Due to limited space, the following is an excerpt of the quantified data of each product.



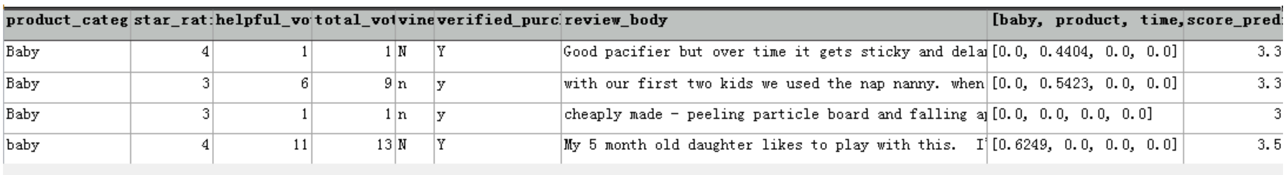
product_categ	star_rat	helpful_vo	total_vo	vine	verified_purc	review_body	[product, time, size, m	score_pred
Major Appliance	3	1	1	N	Y	I like Sharp Microwaves. My mot	[0.0, 0.0, 0.0, 0.0, 0.0]	3
major appliance	4	1	2	N	Y	Really great microwave. Since 1	[0.6249, 0.0, 0.0, 0.0, 0.	3.9
major appliance	3	1	1	N	Y	Larger footprint than expected	[0.0, 0.0, 0.0, 0.0, 0.0]	3
Major Appliance	1	4	21	N	Y	Terrible product. Cheaply made.	[-0.4767, 0.0, 0.0, 0.0, 0.	2.3
Major Appliance	4	3	6	N	Y	We've had the microwave for onl	[0.5423, 0.0, 0.0, 0.0, 0.	3.8

Figure 2. Microwave



product_categ	star_rat	helpful_vo	total_vo	vine	verified_	review_body	[product, air, time, h	score_pred
Beauty	3	1	1	N	Y	I found everything goes well except the plug. Wh	[0.0, 0.0, 0.0, 0.0, 0.0]	3
Beauty	4	1	1	N	Y	This is a really good dryer. I have had it for f	[0.4404, 0.0, 0.0, 0.0, 0	3.7
Beauty	5	3	3	N	Y	Over all great product! It's stylish, light weig	[0.6249, 0.7845, 0.0, 0.0	4.1
Beauty	3	2	3	N	Y	Can a blow dryer be too powerful?.. the answer i	[0.0, 0.4404, 0.0, 0.0, 0	3.1
Beauty	3	1	1	N	Y	After having it only 2 weeks the low blow cycle	[0.0, -0.2732, 0.0, 0.0,	2.9

Figure 3. Hair dryer



product_categ	star_rat	helpful_vo	total_vo	vine	verified_purc	review_body	[baby, product, time,	score_pred
Baby	4	1	1	N	Y	Good pacifier but over time it gets sticky and dela	[0.0, 0.4404, 0.0, 0.0]	3.3
Baby	3	6	9	n	y	with our first two kids we used the nap nanny. when	[0.0, 0.5423, 0.0, 0.0]	3.3
Baby	3	1	1	n	y	cheaply made - peeling particle board and falling a	[0.0, 0.0, 0.0, 0.0]	3
baby	4	11	13	N	Y	My 5 month old daughter likes to play with this. I	[0.6249, 0.0, 0.0, 0.0]	3.5

Figure 4. Pacifier

According to the data in the figure, for these three types of products, the scores of online reviews have a strong correlation with customer ratings. Our scoring standard for customer online review content can largely reflect the customer's own star rating, which verifies the rationality and scientificity of our scoring standard for customer review content.

4. Online scoring model based on polynomial fitting

Since it has been verified that our online review content's rating criteria can well reflect the star rating situation, we use the online content rating to represent customer satisfaction with this product. Since the amount of data for small samples is still very large for function fitting, in order to reduce the amount of data, different scores of different days in the same month are treated as a weighted average. The score obtained represents the rating of the online review content for this month. Based on this, the fitting function of the time of the three types of products and the corresponding score is constructed, as shown in the figure below.

X is a positive integer with a value range of [1. 41], and the corresponding month is 2010.3-2015.8. X is a positive integer with a value range of [1, 43], and the corresponding month is 2012.1-2015.8. x is a positive integer with a value range of [1,27], and the corresponding month is 2013.10—2015.8.

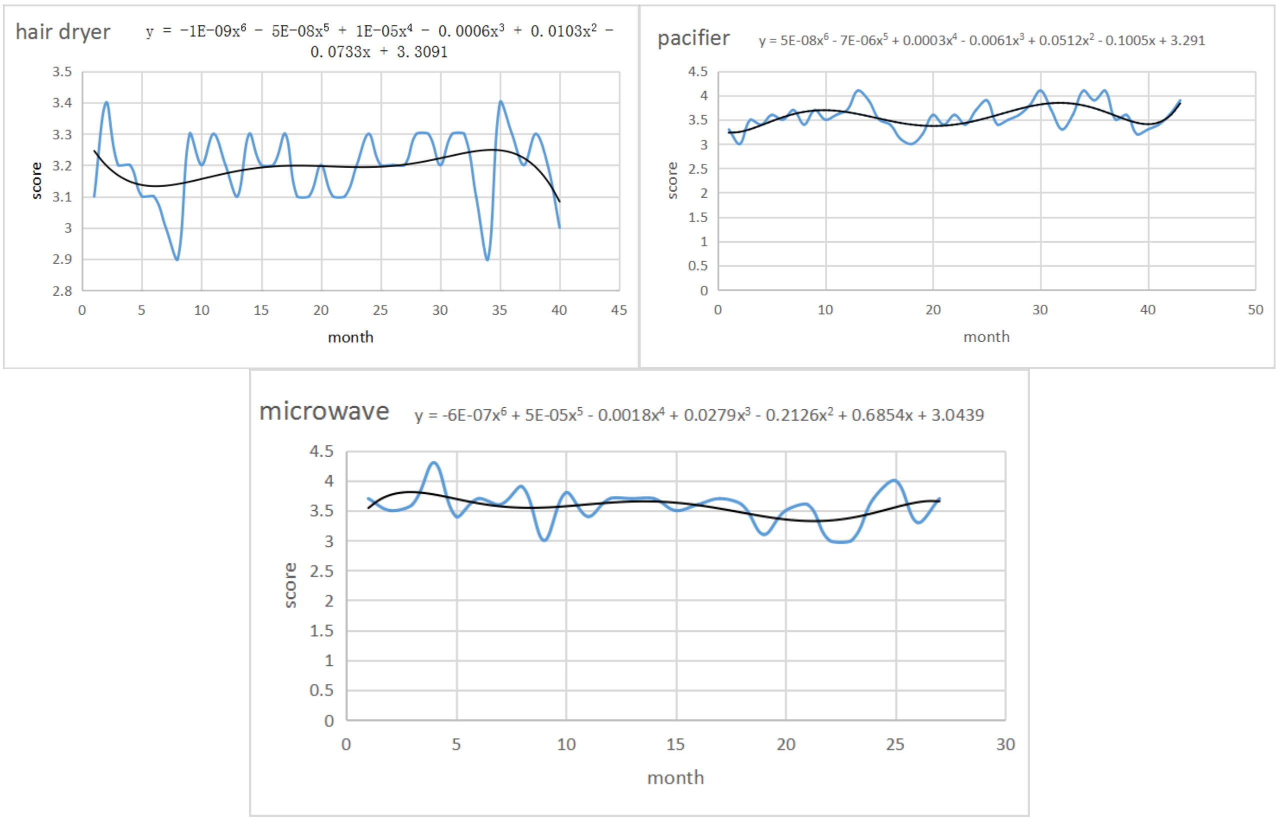


Figure 5. Fitting function of three kinds of products

The advantage of using the polynomial fitting method is that it removes the influence of extreme value fluctuations and can describe its changing trend more intuitively. The disadvantage is that due to the limitation of the polynomial degree, the fitting is not accurate enough.

5. A measure of failure or success.

In order to determine the combination of text metrics and star-based metrics that would best indicate potential success or failure, we analyzed the function of the fitting time and text metric scores above. We find the corresponding month when they fit the maximum and minimum slope of the function. The larger the slope, the greater the potential for success, and the smaller the slope, the greater the potential for failure. By analyzing the scores and star ratings of the textual metrics for the dates contained in this month, we find the combination of metrics that best indicates potential success or failure. As the previous analysis has confirmed that there is a strong correlation between our text metric rating standard and the customer's star rating, it indicates that there should be an approximate functional correspondence between them. This shows that a function expression can be obtained by fitting. If the data from the first few days of a month satisfies this expression of star rating and rating, we can consider that the sales of the products in this month or later are potentially successful or failed. We call this expression a measure of failure or success.

(1) Hair dryer

The fitting function is:

$$y = -1E-09x^6 - 5E-08x^5 + 1E-08x^5 + 1E-05x^4 - 0.0006x^3 + 0.0103x^2 - 0.0733x + 3.3091$$

We use MATLAB to find that the slope of this function is the largest when x = 10, and its corresponding time is 2012.9. We fit the text score and star rating for this month. We use MATLAB to find that when x=41, the slope is the smallest and the corresponding time is 2015.8.

(2) Microwave

The fitting function is:

$$y = -6E-07x^6 + 5E-05x^5 - 0.0018x^4 + 0.0279x^3 - 0.2126x^2 + 0.6854x + 3.0439$$

We use MATLAB to find that when x = 1, the slope of the function is the largest, and its corresponding time is 2013.10. We use MATLAB to find that when x = 5, the slope of the function is the smallest, and its corresponding time is 2014.11.

(3) Pacifier

The fitting function is:

$$y = 5E - 08x^6 - 7E - 06x^5 + 0.0003x^4 - 0.0061x^3 + 0.0512x^2 - 0.1005x + 3.291$$

We calculated with MATLAB, when x = 5, the slope of the fitting function is the largest, and the corresponding date is 2012.6. We use MATLAB to find that when x = 34, the slope of the fitting function is the smallest, and the corresponding date is 2014.11.

In summary, as long as some of the days in a month satisfy the corresponding success or failure metric function, you can determine whether it has the potential for success or failure.

6. Conclusion

In order to solve these problems, we first analyze the correlation between star rating, helpful vote, verified purchase, and number of reviews, and perform preliminary data elimination based on the degree of correlation. We then used text clustering TF-IDF to quantify customer reviews in a targeted manner. Based on the quantitative results, a corresponding review scoring standard was built, and its rationality was verified. In the next section we make several assumptions that play a large role in our analysis and processing of data and models. At the same time, we also did the following: (1) Analyzed which attributes of the product customers value more. Companies can design their own products based on this. (2)We use the established quantitative scoring criteria to give a reasonable score to the online evaluation. Based on this, the relationship between time and score is established to reflect its change with time. (3) According to the change relationship between time and score, find out the potential success or potential failure products.

References

[1] Yang Jiankun. Research on the Impact Mechanism of Online Reviews on the Diffusion of New Products. 016.

[2] Li Yujie. Social Behavior and Personality: International Journal. Volume 47, Number 9, 2019.

[3] Nicky Somohadjo 'the effect of online reviews on the review attitude and purchase intention' May 20.

[4] https://blog.csdn.net/jiangzhenkang/article/details/86749717.