Research on English Text Analysis Based on TF-IDF

Zhijie Qu[1], Yu Shu[1], Fangyi Yang[2] and Yuhang Li[1]

[1]College of Mathematics and Information Science, Hebei University, Baoding, Hebei, 071000, China

[2]College of Management, Hebei University, Baoding, Hebei, 071000, China

Keywords: Online Reviews, Usefulness, TF-IDF

Abstract: The emergence and explosion of online reviews is both an opportunity and a challenge for merchants. In this paper, we described the relationship between star ratings, reviews, and help ratings, and explored trends in product reputation over time. First, we explored text processing methods in the English environment. Based on the TF-IDF idea, we filtered out the stop words in the text, calculated the word frequency, inverse document frequency and TF-IDF value, and extracted the 22 higher-ranked words as keywords. Then, we generated the set of opinion words according to the four opinion word mining rules. Finally, based on the correlation between star ratings, reviews, and help scores, we identified the factors that affect the usefulness of online reviews as Review Len, Total Votes, Different Rating, and Emotion Analysis, and verified the validity of the polynomial regression model through the correlation analysis of the influencing factors.

1. Introduction

With the rapid development of e-commerce, the traditional information dissemination methods have undergone tremendous changes. When people make purchase decisions, the way to seek help has gradually shifted from the offline friends' comments to the online comments on products. Online ratings and reviews provide other customers with a way to learn about the quality of goods and services. At the same time, it also provides a favorable way for the company to better understand the market conditions and increase the number of goods sold.

2. Preprocess Data

If a word is a keyword, it should appear multiple times in the comments. Therefore, we perform "Term Frequency" statistics. Generally speaking, the most frequently used words are "I", "is" and "at". They are called "stop words," meaning words that are not helpful in finding results and must be filtered out.

When we filtered out stop words, we found that the three words "Awesome", "buy" and "well" appeared as many times. This does not mean that all keywords are equally important.

Therefore, determine an importance adjustment coefficient to measure whether a word is a common word. If a word is relatively rare, but it appears many times in the reviews, then it is likely to reflect the characteristics of the product, which is the keyword we need.

Step 1: Count term frequency

$$TF = \frac{word\_count_i}{\sum_{i=i}^{N} word\_count_i}$$

Where $word\_count_i$ indicates the number of times a word has been mentioned in a comment.

Step 2: Calculate inverse document frequency

The most common words are given the smallest weight, the more common words are given a smaller weight, and the less common words are given a larger weight. This weight is called "Inverse Document Frequency" and its size is inversely proportional to how common a word is.

At this time, a corpus needs to be established to simulate the use environment of the language.

$$IDF = log\left(\frac{\sum_{i=1}^{N_1} document\_count_i}{1+document\_count_i}\right)$$

Where $document\_count_i$ is the number of documents in the corpus that contain the word.

Step 3: Calculate TF-IDF

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

It can be seen that TF-IDF is directly proportional to the number of times a word appears in a comment, and inversely proportional to the number of times that word appears in the corpus.

Calculate the TF-IDF value of each word in the comment, and then arrange them in descending order, and extract the first few words as keywords.

Opinion words are words with user opinions in user reviews, which can represent the user's emotional tendencies. The idea of this mining rule is as follows:

$$if\ \exists\left(M, NP = f\right) \to po = M$$

$$if\ \exists\left(S = f, P, O\right) \to po = O$$

$$if\ \exists\left(S, P, O = f\right) \to po = P$$

$$f\ \exists\left(S = f, P, O\right) \to po = P$$

Where po is the opinion, M is a modifier, NP is a noun phrase, S is a subject, P is a predicate, O is an object, and f is a combination of product attributes and nearby words.

The meaning of these four rules [1] is: if a modifier exists in front of a noun phrase, then the modifier is considered a opinion word. If the subject is the main predicate-object structure, the object is considered to be an opinion word. If f is the main predicate-object structure, the predicate is considered to be an opinion word. If the subject is the main predicate-object structure, the predicate is considered to be an opinion word.

3. Star Rating and Helpfulness Ratings

Star rating is a variable used to describe the intensity of emotional tendency of consumers when evaluating products. Consumers can score digitally (1-5) on the products they are reviewing. The higher the score, the stronger the tendency for positive evaluation. The lower the score, the stronger the tendency for negative evaluation. Otherwise, the evaluation is more neutral.

In the case of an overloaded number of online reviews, perceived usefulness is the degree to which customers' subjective perceptions generated by viewing reviews from other consumers are useful.

Take the hair dryer as an example to analyze the internal relationship and interaction between star rating and help level. See the appendix for examples of microwave ovens and baby pacifiers.

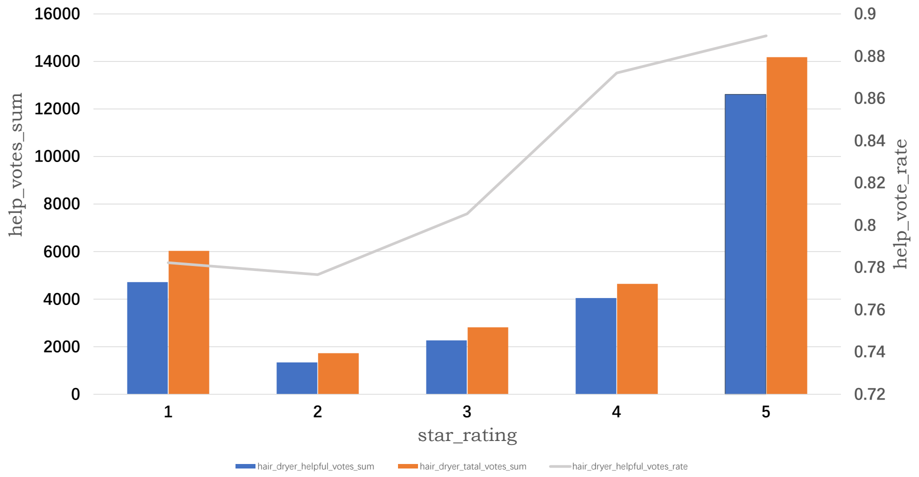

Figure 1. Hair dryer's star rating and helpfulness rating

Based on the analysis in Figure 1, there are more records with a rating of one star and five stars than other stars, which means that consumers are more likely to express stars with strong emotional colors that can cause wide communication Post-purchase experience. And extremely high star ratings can easily resonate with other consumers. Therefore, companies need to pay more attention to the content of extreme evaluation, collect useful information that can promote product promotion, and use it as the basis for developing sales strategies.

Next, from the perspective of usefulness, potential consumers make helpful evaluations that are far greater than unhelpful evaluations, and the helpful evaluation ratio is roughly positively correlated with star ratings. This shows that potential consumers don't pay too much attention to useless reviews, and only submit helpful ratings in those reviews when the perceived usefulness reaches a certain threshold.

4. Star Rating and Reviews

Generally speaking, a higher star rating is more positive, a lower star rating is more negative, and a middle star rating is relatively neutral. Based on the idea of cosine similarity, let's explore the similarity of reviews with different star ratings.

First we perform text vectorization. We segment the comment content, list the words contained in each comment, calculate the frequency of each word, and determine the word frequency vector.

We use cosine similarity to quantify the degree of similarity of reviews of the same star. The closer the cosine value is to 1, the closer the angle of the word frequency vector to 0 degrees, the more similar the two vectors are. This is called “cosine similarity [2]”.

The expression of the cosine similarity between the word frequencies vectors obtained from the pre-processed text vectorization is:

$$cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

It can be seen in Table 1 that the three products have a large difference in the content of one-star and five-star reviews, and the content of neutral-star reviews is similar. This indicates that extreme assessments will contain a wealth of information, which will help merchants gather customer needs. And based on the highly similar comments, we can extract the functional words of the products with higher attention.

Table.1. Similarity of different star reviews

Star_Rating	Pacifier similarity	microwave similarity	hair_dryer similarity
One star	1.4901	1.5319	1.5514
Two stars	1.5632	1.565	1.5588
There stars	1.559	1.5648	1.5694
Four stars	1.5692	1.5659	1.5703
Five stars	1.5661	1.5587	1.549

5. Factors affecting the helpfulness of online reviews

Online reviews are users' positive, objective, dissatisfied and negative evaluations of product attributes and services in the form of text, pictures, etc. with the help of the Internet. The impact of online reviews on the proliferation of new products is intricate.

This article studies the helpfulness of reviews from the length of the review content, the proportion of product attributes in the review content, the total number of people who voted on the review, the average of the difference between the review score and the average product score, and the emotional color in the review content.

(1) Review_Len

In general, the more words in a comment, the more product information it contains. Through comments, businesses can get more information about product functions and services. It can encourage businesses to improve their products, increase market attraction and promote the promotion of new products. However, the length of online comments is not the longer the better, and the marginal utility of its impact is decreasing.

(2) Attributes

After using the jieba library in Python software to perform word segmentation and word frequency statistics on each group of comments, we obtained the word frequency statistics tables of product attributes for microwave ovens, pacifiers, and hair dryers. Table 2 is a statistical table of some attributes of the three new products. For details, please refer to the appendix.

Table.2. Summary of word frequency of attributes of three new products

Commodity	Glossary describing product attributes
Microwave	space, looks, sharp, price, power, size, counter space, repair
Baby Pacifier	size, nipple, quality, clean, price, soft, easy put, looks, service
Hair Dryer	time, price, power, speed, light weight, air flow, flat iron

Based on the results of mining high-frequency attribute words from online review information, Sunshine can implement "user-centric" product design.

(3) Total_Votes

In the review mechanism, potential consumers can evaluate whether the content in the review is helpful. The more reviews a review receives, the more attention the consumer receives, which affects the usefulness of the review.

(4) Differ_Rating

Evaluation polarity is a variable used to describe the intensity of emotional tendency of consumers when evaluating products. Ghose and Ipeirotis [3] found that extreme reviews were more useful than neutral reviews in their research. However, Mudambi and Schuff [4] found that in experiential goods, neutral evaluation was more welcomed by consumers than extreme evaluation. This article gives the following mathematical calculation expressions to evaluate polarity:

$$Differ\_Rating = Star\_Rating - Star\_Rating\_average$$

According to Figure 2, it can be concluded that the percentage of Amazon Vine Voices' helpful votes is 76%, and that of ordinary users is 85.13%. This shows that after being awarded the title of

Amazon Vine Voices by reviewers, their reviews will affect the credibility of the product for free trial.

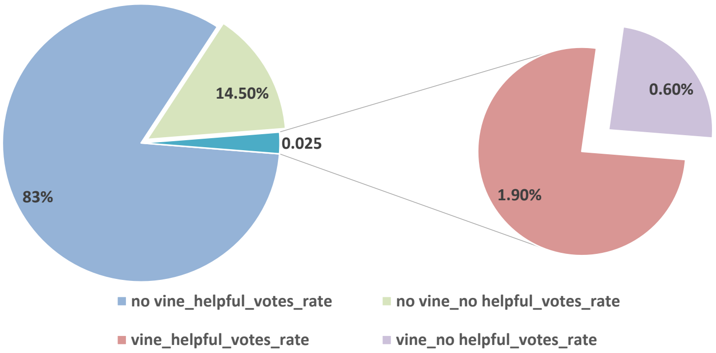

Figure 2. Helpful ratio of Amazon Vine Voices to regular reviews

(5) Emotion_Analysis

Emotional analysis, also known as opinion mining, is the process of analyzing, processing, inducing, and inferring subjective text with emotion. According to the processing of the review data of the above-mentioned text nature, first of all, the effective information in the review sentence is extracted, the praise attribute is defined, and then the judgment is made based on the extracted words. Get the sentiment attribute value expression for the review:

$$Emotion\_Analysis = 2\alpha + \beta$$

Where α indicates the number of negative comments in a single review, and β indicates the number of positive comments in a single review.

6. Conclusion

In this paper, we first process the text reviews of the three products. Through word segmentation, word frequency statistics, and DF-IDF algorithm to vectorize the text, so that the qualitative text can be expressed quantitatively, it seems to get the emotional tendency of customers. Next, after preprocessing of text-based data and numerical data, we identified 5 factors that affect the usefulness of online reviews, namely Review_Len, Total_Votes, Different_Rating, Emotion_Analysis, Attributes, and verified the validity of the polynomial regression model through the correlation analysis of influencing factors. It can mine highly useful reviews for merchants, and can timely adjust product sales strategies through reviews.

References

[1] Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews [M]. Natural Language Processing and Text Mining. Springer London, 2007: 9-28.

[2] Ruan Yifeng. Application of TF-IDF and cosine similarity [EB / OL]. [2013-3-21]. https://www.ruanyifeng.com/blog/2013/03/cosine_similarity.html.

[3] Ghose A, Ipeirotis P G. Designing Novel Review Ranking Systems: Predicting the Useful-ness and Impact of Reviews [C]. Proceedings of the ninth international conference on electronic commerce, New York, 2007: 3030-310.

[4] Mudambi S, Schuff D. What makes a helpful online review? A study of customer reviews on Amazon.com [J]. MIS Quarterly. 2010, 34 (1): 185-200.

[5] Chatterjice P.Online review: Do consumer Use them [J]. Advances in consumer research, 2001 (28): 133-139.

[6] Sang A, Ismail R, Boyd C. A survey of trust and reputation systems for online service provision [J]. Decision Support Systems, 2007, 43 (2): 618-644.