

Summary

With the rapid development of Internet business, online shopping has become the main shopping method for more and more customers. A large amount of data is also generated in this process, including customer ratings and comments on the goods. These data may contain a wealth of information that can help merchants evaluate current products and provide a reference for further business decisions.

The purpose of our model is to mine valuable information for the evaluation of the product from the data set of customer scoring and evaluation, use this information to analyze the reputation, development and prospects of the product and establish the correlation between scoring and evaluation. In order to make full use of the rich text information in the data set, we propose a method that integrates word frequency statistics, semantic analysis, vector operation and frequent item set mining to extract emotions and keywords from the text. In order to make full use of various forms of data to evaluate the product, we have developed a comprehensive star rating, emotion rating and the credibility of comments. Finally, we use these important indicators and features to combine the temporal model, correlation analysis, visualization technology, and machine learning algorithm to analyze the actual data. In part 1 of the model, we quantify the credibility of comments based on the helpfulness rating of comments, the type of customer, and receipt. We used word frequency statistics technology and TF IDF method to extract keywords with different frequencies from user comments, and selected emotion keywords and property keywords. Through the semantic analysis and vectorization calculation of emotion keywords, the emotional feature vector and emotion score of comments are formulated.

Combine the star rating, emotion score and credibility to give a comprehensive score for the product. In addition, we use the fp growth algorithm to extract key phrases that imply the properties or features of the product. The above indicators highly summarize the content of the data set and have important information value. In part 2 of the model, we first point out the two important indicators of comprehensive score and key phrase. Whether the trend in composite score over time reflects changes in product reputation is judged by linear regression significance analysis. By comparing a comprehensive score with other products and combining the time series to judge the potential success or failure, the advantages or disadvantages of the product extracted from the keyword set also have important reference value. Using the method of differential analysis, we found that. A specific star rating does not have much impact on subsequent reviews.

Ultimately, we confirmed using error analysis and density distribution analysis that positive or negative affect affect the star rating. The results of decision tree classification using emotion feature vectors provide more evidence for the relationship between emotion and star rating. An analysis of key phrases shows that customers who praise or complain about the nature of the product can also indicate a star rating.

Keywords: TFIDF,online comment,ML,FP tree.

Contents

1	Introduction	2
2	Assumptions and Justification	2
3	Notations	2
4	Model	2
4.1	Model buliding	2
4.1.1	Overview	2
4.1.2	Pre-process	4
4.1.3	Data analysis method	7
4.2	How it be applied in practical problems	8
4.2.1	Restatement of the Problem	8
4.2.2	Find important features based on ratings and reviews	9
4.2.3	Time-based measures and patterns	10
4.2.4	Evaluation method of products	11
4.2.5	Relation between past ratings and current reviews	12
4.2.6	Relation between rating and review in one record	13
4.3	Results	15
5	Further discussion	16
5.1	Contribution	16
5.2	Conclusion	16
	Appendices	19

1 Introduction

Recently online comments have developed into a well-liked and effective strategy for sellers to gain consumer feedback and raise the caliber of their services. Online comments are particularly appropriate for online purchases.[2]Companies have strong willing to get important information from online comments and other features.

However,some problems are not simple,for example,to find relation between past ratings and current reviews or to predict a product will be successful or not can be challenging.Particularly, some data are in text form which is need more different and complex method to analyse than numeral data.

Data digging and text analyze technologies are powerful when being applied in these problems,current paper has tried using a technology called TF-IDF to process text data[1],moreover, some have used machine learning algorithm like decision tree[3] and other algorithms to classify[4]. This paper analyses problems and illustrate thinks and ideas about solving problems,after that it gives assumptions used and demonstrates justification of them,following assumption, notation part draws a table to show all notations of important variables and concepts.Before discussion,model part build a general model using varies technologies and using it to solve problems,next it gives result.In discussion,details strength and weakness of model are shown, following are contribution and conclusion,in the end overall results and code are lying on the appendix.

this paper build a model using several machine learning algorithm and some widely using text analyze technology TF-IDF,it aims to answer what feature of data is the most informative and solve practical problems.Finally, both case studies and experiments are conducted to demonstrate that the proposed methods are effective.

2 Assumptions and Justification

1. The number of comments and the stars given by users in a certain month are closely related to the number of comments and the distribution of the stars, and mainly only related to the number of comments and the distribution of the last month,which means month $i-1$ affects month i and it can be detected in month $i+1$.Suppose

$$\begin{cases} S_m = \sum a \\ b_{mn} = a_{mn}/S_m \end{cases}$$

2. Assuming that users mainly read the comments of the past month and the number of comments and positive reviews for the next month were closely related to the number of positive or bad reviews of the last month.

3 Notations

Notations are in table 1.

4 Model

4.1 Model buliding

4.1.1 Overview

Model was shown in fig.1. This Model was built to analyse data of customer-supplied ratings

Definition	Meaning
N	the total number of months
SR	star rating
TES	text emotional score
R	reliance
CSI	Comprehensive satisfaction indicator
$tf(x, t_i)$	number of occurrences of x in text t_i
$tf-idf(x, t_i)$	TF-IDF word
t_i	the i th text
$df(x)$	number of occurrences of x in all texts
S^K	important word set
G^K	group consists of important words
$S_H(t_i)$	high-frequency word set
$S_L(t_i)$	TF-IDF word set of t_i
S_i	different type word sets,i represents type of word
$V_E(t_i)$	emotional feature vector
a_{mn}	number of comments with n stars in the month m
b_{mn}	proportion of n star comments in the total number of comments in the month m
S_m	Total number of comments at month m
e	gap between star rating and emotional score

Table 1: Symbol Table

and reviews associated with products, aims of the model were to solve problems including to find the most informative indicators, relations between past ratings and current reviews, rating and review in one record, additionally, to find a method to evaluate each products to be whether successful or not. The Model consisted pre-process, data analysis method parts that converted data into more appropriate forms and analyzed data respectively.

Review also plays a important role in data analyse, providing large information to be reference, however, several reviews may be too long and contain many irrelevant statements to the products, it is therefore necessary to extract the features from the text of reviews.

Review may contain two types of important features, the emotional attitude of customers towards the product and the evaluation of certain characteristics of the product. The emotional attitudes of customers towards the product may be partly related to their rating, while the evaluation of some characteristics of the product may reflect the advantages or disadvantages of the product. Therefore, these two types of information are of great value for the commercial analysis of the products. It is important to obtain them from the text.

In pre-process part, the four indicators helpful votes, total votes, vine, verified purchase were converted into one indicator called reliance. After that review headline and review body were used to obtain high-frequent word bags by a technology named term frequency-inverse document frequency (TFIDF) which is a common weighting technology for information retrieval and data mining to evaluate the importance of a word to a document set or one of the documents in a corpus, then review date was converted into mouths from the original format.

In data analysis method part, several machine learning methods were tested on a example. This part provided that a method how to process text data. After building model, some practical and valuable problems were solved by the model, including To find important features, and Time-based measures and patterns, relations between past ratings and current reviews, rating and review in one record and how to judge a successful product. In the end results of these solutions

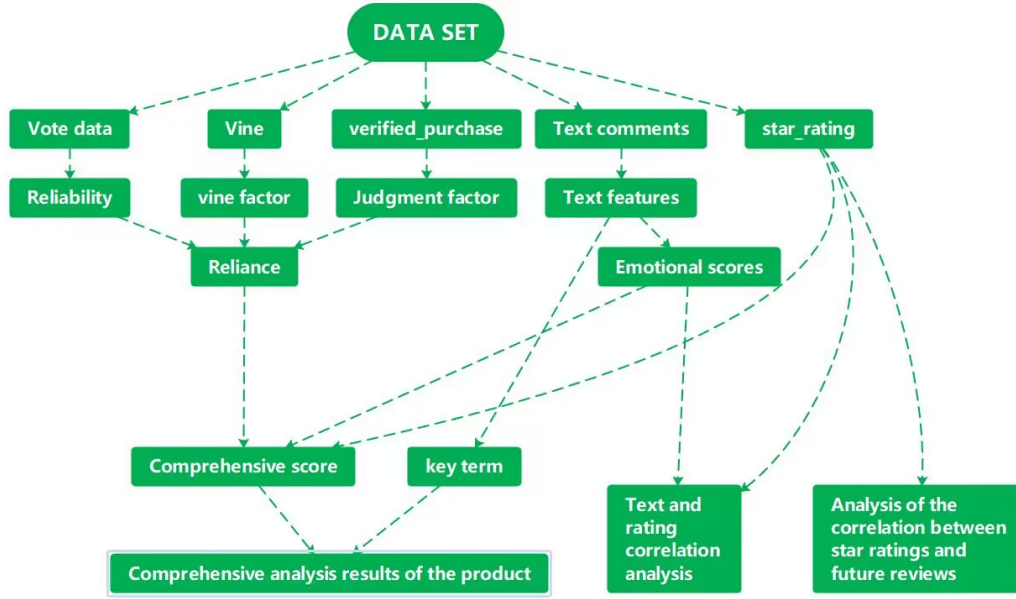


Figure 1: Model

and explanations were given.

4.1.2 Pre-process

The four indicators, helpful vote, total vote, vine and verified purchase, actually are one kind of indicators that reflect whether the respective record is reliable or not so that they can be converted into one indicator, named reliance in this paper. The higher the reliance, the higher the importance of the respective comment.

Reliance: This indicator was created since there are some crucial problems that needed to be resolved in order to evaluate and improve product quality based on online reviews automatically, including how to use less reliable online reviews, how to identify quality flaws from online reviews, and how to suggest more practical or affordable evaluation indices to improve service quality based on online reviews. Reliance was calculated as:

$$R = r \times v \times j$$

r is calculated as:

$$r = 0.5 \tanh [0.15 (2H - T)] + 0.5$$

Where H represents the number of helpful votes for this record while T represents the total number of votes for this comment. $2H - T$ is essentially the number of votes for helpful minus the number of votes for unhelpful. In addition, the larger the $2H - T$, the larger the r is, when $2H - T$ tends to positive infinity, r tends to 1, means to be trusted, when $2H - T$ tends to negative infinity, r tends to 0, means not to believe. When $2H - T$ is zero, that is, people think it helpful as many as unhelpful, $r = 0.5$. 0.15 is the modulation factor determined in advance. Under the calculation of r , when the vote number of helpful is 10 votes higher than that of unhelpful, r can reach 0.95, which is close to 1, which means very worthy of trust.

v is obtained from the data set directly, using rule as follows:

$$v = \begin{cases} \frac{N}{n_v} \times \frac{1}{5}, & \text{Vine} = Y \\ \frac{N}{N - n_v} \times \frac{4}{5}, & \text{Vine} = N \end{cases}$$

where n_v is number of Vine people. According to statistics, vine users account for about 1-2 percents of the total users. Therefore it was proper that allocating 20 percents of the total weight was given to the evaluation of vine users while the remaining 80 percents was given to ordinary users.

j is also obtained from the data set directly:

$$j = \begin{cases} \frac{N}{n_{pur}} \times \frac{9}{10}, & \text{Purchase} = Y \\ \frac{N}{N-n_{pur}} \times \frac{1}{10}, & \text{Purchase} = N \end{cases}$$

Since the comments before receiving the goods are irresponsible and not worthy of reference, so the relevant information of the comments should be given a small judgment factor, that is, to reduce its reference ability, and in order to ensure that the total reference ability remains unchanged, the other users need a judgment factor slightly greater than 1 to maintain. The corresponding emotional score, $E(t_i)$, describes the emotional type of the text.

Text emotional score: This paper adopted a keyword-based model to summarize the emotion of the comments and found some features from them. Text process was shown in fig.2.

The model used ordinary word frequency statistics and TF-IDF word frequency statistics, the TF-IDF method calculated the relative number of occurrence of the resulting word x in t_i , to obtain high frequency keywords and low frequency keywords, respectively. high frequency keywords were chosen to contain valuable information as much as possible, while low-frequency keywords reflected special attributes and the appearance of these words can indicate the characteristics of a certain product.

The total process was shown in following figure: To get high-frequency keywords, First, determining the range of text for word frequency statistics. Under the differences in the size of the given data set, it was difficult to obtain high-frequency words from the full text of all reviews in the larger data set and the title of the review actually well summarized the content of the reviews. Therefore, for the hair dryer data set and the pacifier data set (more than 10,000 data volume),

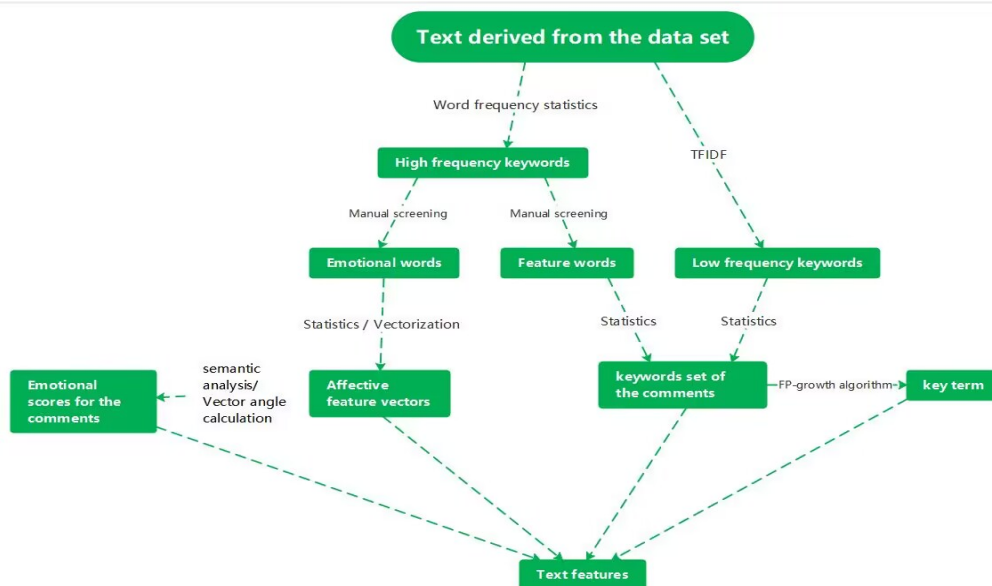


Figure 2: text process

only the word frequency in the review title were counted. For the microwave data set, the full review were counted

Since some of the most frequent words in all texts were often articles, pronouns, AIDS and other words without practical meaning, these words needed to be removed and two different ranges of word frequency statistics could be used to obtain the ideal collection. For the first time, model selected the words with a high frequency of words and most of the words were meaningless words, from which selected the valuable words as a part of the high-frequency keywords. In the lower boundary of the first word frequency in the second statistics, the 100 words with the highest word frequency not above the upper boundary were selected. They were combined with the keywords obtained in the first statistics to obtain the high frequency keyword set S^H . The specific screening methods were described as follows:

$$x \in S^H \text{ if } \begin{cases} df(x) > T \text{ and } x \text{ is valuable} \\ df(x) < T \text{ and } df(x) \text{ is large enough} \end{cases}$$

where T was threshold, as determined by the text range, the threshold value was 100 in the hair dryer data set and pacifier data set and 150 in the microwave data set. The CountVectorizer class provided in the Sicit-learn library of Python could find all the words whose word frequency were within a specific range.

Due to the limited number of high-frequency keywords (no more than 150), many valuable words were missed. These words did not appear frequently, but could reflect some characteristics of the product. TF-IDF method was used to obtain these low-frequency keywords which was a method to calculate the word frequency of each text by combining the word frequency and particularity of words. TF represents the word frequency in the text, and IDF referred to the inverse document frequency. The calculation formula is as follows:

$$tfidf(x, t_i) = tf(x, t_i) \times \left(\ln \frac{1 + |D|}{1 + |D_x|} + 1 \right)$$

The coefficient at the end of the right formula represented the inverse document frequency, $|D|$ represented the total number of text, $|D_x|$ represented the number of texts containing words x , the common word inverse document frequency was low in many texts, while the words appearing in only a small number of documents had a high inverse document frequency. Therefore, the TF-IDF method could screen out those words that rarely appeared in all texts but may contain special information, namely, low-frequency keywords. Using the TF-IDF method (for the three datasets, the scope of text is the title and full text of the comment), generate low-frequency keyword collection S^L ,

$$x \in S^L \text{ if } tfidf(x, t_i) \geq 5 \text{ and } df(x) \in [5, 100]$$

The TfidfVectorizer class provided in the Sici t-learn library of Python can calculate the TF-IDF values for all words with word frequency in a specific range.

TES was calculated based on the emotional feature vector angle, different texts contain different but close-sense emotional words. The corresponding similar words in the text were merged following classification $S_{positive}, S_{negative}, S_{passive}, S_{adversative}$ to gain the emotional feature vector of the text V_E ,

$$\begin{aligned} n_i &= \sum_{x \in S_i} tf(x, t_i) \\ i &\in \{\text{passive, negative, adversative, positive}\} \\ V_E(t_i) &= [n_{positive}, n_{passive}, n_{adversative}, n_{negative}] \end{aligned}$$

since the complete comment content may contain irrelevant information and interfere with the evaluation of emotion. Consider only the comment title on the data set with sufficient comment title content.

Moreover, because the comment titles were short and few complex semantic changes occur, the effects of positive and negative words could be added through the analysis of semantics to reduce S_E to two dim vector $v = [v_1, v_2]$, v_1 measured positive emotions, was $n_{positive}$ initially, v_2 measured negative emotions, was $n_{negative}$ initially.

Some useful rules were created as follows:

1. When at least 3 categories of words appeared simultaneously in the text, it was considered impossible to obtain accurate emotions from the semantics, and the text is neutral, set both v_1 and v_2 to 1
2. When the text had no more than 2 categories of words and had negative words, the text was considered to have a small amount of negative emotions, added v_2 plus 0.5
3. When no more than 2 types of words appeared in the text and the turning words appear, the smaller on between v_1 and v_2 plus 1
4. Vector $[1,0]$ could represent a fully positive text,

The φ was transformed and transformed into 1 to 5, where 1 represented completely negative text and 5 represents completely positive text. This value was called the emotion score E of the text. The calculation and transformation formula of φ and E were as follows:

$$\varphi = \begin{cases} \frac{\pi}{4}, v_1 = v_2 = 0 \\ \cos^{-1} \left(\frac{v_1}{\sqrt{v_1^2 + v_2^2}} \right), E = 1 + \frac{4\pi - 8\varphi}{\pi} \end{cases}$$

CSI: CSI took reliance and TES and rating into consideration, it was a all-round indicator for a record. CSI incorporated information from all reviews under a product, including the credibility of each comment and importance, user ratings and emotional scores of each text comment evaluated after our model analysis.

CSI was calculated as

$$CSI = \sqrt{rating^2 + TES^2} \ln R$$

where rating was just rating from data set, the higher the reliance, the bigger the CSI, however, the quantity of CSI was based on TES and rating. Companies can judge the current market performance of the product, and they can track how the composite score changes over time. CSI comprehensively reflected the satisfaction of the product in the hearts of customers and generally shows whether a product was successful.

4.1.3 Data analysis method

By analyzing the comment data given in the table, evaluated whether the star rating given by the user will affect the number of comments or the praise rate. It was appropriate to choose one month to be a period.

Constructs the keyword set of comments After screening out the S_E from the S_h , the remaining words still had research value, including adjectives, verbs and nouns related to the product. These high-frequency words may reflect the important nature of the product and these properties were valued by customers. These words form a collection of S_F , and the keywords in the comments are likely to come from S_F and $S_L(t_i)$. Based on the above analysis, the keyword set of the text could be constructed as follows:

$$S^K(t_i) = S_F \cap t_i + S_L(t_i)$$

The $S^K(t_i)$ in the formula contains words in t_i with critical information or important properties, equivalent to a concentration of the comments.

Key phrases were constructed by using the FP growth algorithm The keyword sets of each review constituted the shopping basket transaction. FP growth algorithm was used to extract the frequent item sets from the keyword sets of all the reviews of the product. The phrases represented by these frequent item sets appear many times in the comments.

The FP growth algorithm was an efficient association analysis algorithm that used a compact data structure called the FP tree to organize the data and extracts the set of frequent terms from this structure. The FP tree was compressed by reading the transaction one by one and mapping it to a path of the FP tree and iterated in the runs.

The entire database was first scanned once, and the number of occurrences of each element was recorded and arranged in descending order by frequency for an word x in data set to obtain the sorted data. The data set was then scanned once and then filtered again, and elements smaller than minimum confidence, which was a threshold set in advance, were removed to generate the item header table. According to the descending order of times.

The FP tree was built based on the database after sorting the item header tables. At the beginning, the FP tree had no data, establish a root node named NULL, then read into the sorted data set, inserted into the FP tree, inserted into the FP tree according to the sorted order, after that the top node was the ancestral node, while the bottom was the descendant node. If there was a shared ancestor, the corresponding common ancestor node was counted by 1. After insertion, if a new node appears, the node corresponding to the item header table would link to the new node through the node link table. FP tree establishment was complete until all data have been inserted into the FP tree. Find the conditional pattern base from the bottom item of the header table. Frequent set. of the header item from conditional pattern basis recursive mining. A conditional pattern base was a collection of prefix paths appearing in conjunction with the suffix pattern in FP-Tree. That is, the set of ancestral paths for all nodes of the same frequent term in the FP tree. This algorithm extracted the set of frequent items from a set of comments as the key phrases G^K , these key phrases could represent important features of the product and had the potential to demonstrate their associations.

4.2 How it be applied in practical problems

4.2.1 Restatement of the Problem

There were tens of thousands of comments so as to draw product conclusions and operational strategies useful for merchants, including:

- Identify a specific data measurement method based on reviews and ratings to help merchants track product sales in a timely and effective manner. To solve this problem the text techniques were used to evaluate the importance of a word to a document set or a corpus

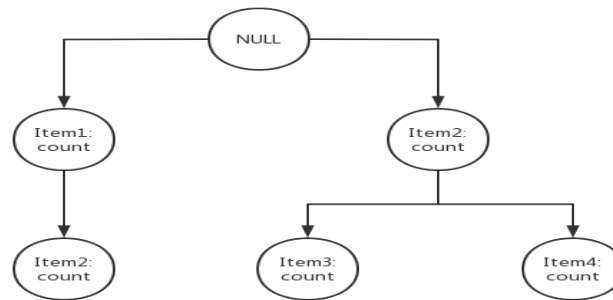


Figure 3: FP tree

and use a decision tree for data analysis to evaluate specific product sales from ratings and text reviews.

- Determined a algorithm to evaluate the reputation of a product so that it can clearly reflect the reputation of the product with various factors. This paper integrated the information of each comment, each comment voting, user identity information specific quantified as a comprehensive index, It reflects the importance of the comment in the analysis, covering the credibility of the comment, whether the user was important, etc. It acted as a weighting factor on review ratings and text comment analysis, which calculated the monthly reputation score. Through the reputation of the distribution in time it can be observed the product reputation based on time.
- Determined the evaluation information of the product in the table based on text comments and scoring in order to specifically reflect the business status of the product. CSI was created as the overall evaluation indicator for this product. Several information types use the special algorithm to calculate the final total evaluation score CSI, which can fully and accurately reflect the situation of the product in the vast majority of customers' thoughts.
- Due to assumption of reading habit of customers, positive rating rate and bad rating rate of each month in the table were extracted and its correlation with the number of comments of the next month was illustrated. The correlation analysis used a specific quantitative calculation index, which can reflect the positive/negative correlation or no correlation between the two kinds of data.

4.2.2 Find important features based on ratings and reviews

According to the model, there were two types of data metrics of concern. One was the rating classification index, which includes the star rating, emotion rating and comprehensive rating of the comments; the other was the text class index, namely the key phrases extracted from some user comments. By tracking these metrics, companies could understand the product's reputation on the market and the strengths or weaknesses it may have.

1. Star rating was an intuitive reflection of the popularity of a product. Through the analysis of the number of high scores or the number of low scores of the product, the reputation of a product. The comprehensive score combined the first two scores and considered the

reliability of the comments. Therefore, the comprehensive score, especially the comprehensive score, is what the company needs to focus on.

2. Companies could gain the reputation of the product in the market from CSI, but specific content was needed to develop countermeasures. Using the extraction method of the key phrase GK proposed in the model, companies could obtain the key phrases in a certain product or some specific review. These phrases often contained the important and special nature of the product. By tracking these text indicators, the company could timely find the advantages or disadvantages of the product, analyzed the reasons for the success or failure of the product, and made the next step.

4.2.3 Time-based measures and patterns

This paper chose month as the least time period to analyze and tried to find the relationship between time and CSI. Two products, which were 'remington ac2015 t|studio salon collection pearl ceramic hair dryer, deep purple' and 'conair 1875 watt tourmaline ceramic hair dryer', were chosen to be analyzed, the result was shown in fig.4, fig.5: It could be seen that product1 kept the same level while product2 kept decreasing which means customers were not

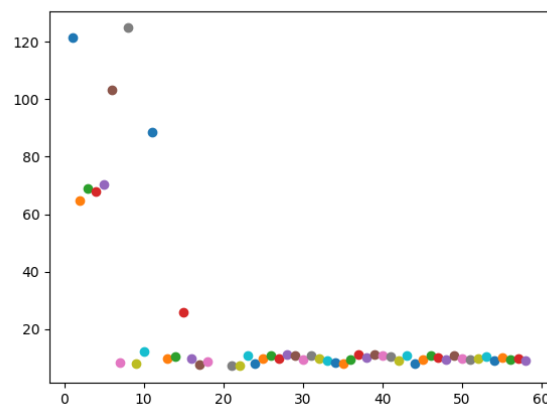


Figure 4: Tendency of product1

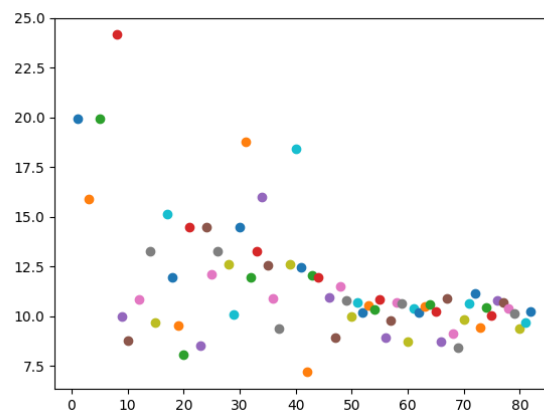


Figure 5: Tendency of product2

satisfied with it.

4.2.4 Evaluation method of products

To analyze the potential success or failure of the product according to the two indicators formulated in 5.2.1, the development of the product could be analyzed from the following three aspects: the comparison of comprehensive scores between different products, the change trend of comprehensive scores over time, and the potential advantages or disadvantages of the product. The following was the general process of the analysis process:

1. In business, competition in similar products was widespread. Products that often got high reviews and had positive customer reviews may gain more market, while products that often got low customer reviews and negative customer reviews may lose the market. The comprehensive score measures the reputation of the product in the market. If the comprehensive score of a certain model is significantly higher than that of other products over a period of time, the product may be considered to succeed. for example, the following are the average combined scores of four models of microwave ovens over the same period: The microwave oven composite product score of 215953885 was significantly lower than the other products, so it could be concluded that the product will be a potential failure. The microwave oven composite product score of 215953885 was significantly lower than the other products, so it can be concluded that the product will be a potential failure.
2. In 5.2.2, the trend of the comprehensive score of the product was indicated according to the chronological order. In fact, the success or failure of the product could be predicted by analyzing it in a certain period of time. The stability test based on time series or linear regression could reflect the trend of the product. If the composite score of a product continues to rise over time or maintain a high score, the product may remain high in the following time, and the product could be considered a potential success; otherwise, if the composite score of a product declines over time or maintains a low score, the product may remain low in the following time, and the product could be considered as a potential failure. The conclusions in 5.2.2 could be used as an important basis for this issue.
3. The comprehensive score was only a rough evaluation of the quality of the product, so as not to judge the success or failure of the product. At this point, it was more important to explore the potential advantages or disadvantages of the product, because these factors directly affect the performance of the product in the market. analyzing all the positive or negative reviews of a product in the early stages of the market and extract the key phrases. Considering the way GK is generated, these key phrases appear many times in the comments, reflecting certain properties of the product. If these key phrases may reflect the good nature or characteristics of the product, the product may fail. For example, some of the key phrases extracted from the initial high score reviews of the model "conair 1875 watt tourmaline ceramic hair dryer" were as follows: From these phrases could summarize many advantages of the blower, such as light, quiet, strong wind and

Product id	423421857	215953885	109226352	771401205
Average CSI	11.41	3.31	11.54	13.88

Table 2: The average CSI of the microwave oven

powerful	light
would	light
lightweight	
powerful	
quality	
quickly	
quiet	

Table 3: Part of the key phrases of a hair drier

so on. Therefore, it could be considered that this hair dryer was a potentially successful product, and it was proved that this product does develop well in the later stage. If other types of hair dryers reflect such a nature, it could also be considered as a potentially successful product. Some of the key phrases that could be extracted from the initial low score reviews of the microwave oven with the model "samsung smh1816s 1.8 cu. ft. stainless steel over-the-range microwave" were listed below: From these phrases, it could be summarized that the microwave oven often had problems, and the after-sales warranty service was not satisfactory. The microwave could therefore be considered a potential failure to actually score badly in the later stages. If other types of microwave ovens reflect such a nature, it could also be considered a potential failure product.

4. different ML classification:

4.2.5 Relation between past ratings and current reviews

lack of illustration To solve this problem, standard of high and low score class was determined that the stars into 1-3 stars and 3-5 stars, representing high and low stars level respectively, then chose to analyze the relation between the indicators b_{Lm}, b_{Hm} which represent proportion of high or low score comments in the month m , the total number of comments S_{m+1} in the next month and the number of low stars $a_{L(M+1)}$ and the number of high star comments $a_{H(M+1)}$. These parameters are given following

$$\begin{cases} a_{Lm} = a_{1m} + a_{2m} + a_{3m} \\ a_{Hm} = a_{3m} + a_{4m} + a_{5m} \\ b_{Lm} = b_{1m} + b_{2m} + b_{3m} \\ b_{Hm} = b_{3m} + b_{4m} + b_{5m} \end{cases}$$

Determine two difference operators A, B

$$\begin{aligned} A(x_m, y_{m+1}) &= (x_m - x_{m-1}) \cdot (y_{m+1} - y_m) \\ B(x_m, y_{m+1}) &= \sum_{m=1}^{N-1} (x_m - x_{m-1}) \cdot (y_{m+1} - y_m) \end{aligned}$$

A is the product of the increase of an indicator this month over the last month and the increase

problem	repair
problem	warranty
problem	error
customer	survice
repair	survice

Table 4: Part of the key phrases of a microwave

of another indicator next month over this month. The meaning is assumed that the increase of comments next month over this month and the change of high grade / low grade comments proportion this month has a strong relationship, the multiplication represents the contribution value of the change of high grade / low grade comments proportion respectively to the change of the next month, for the positive equivalent to the incentive effect and negative is equivalent to the retroaction.

For example, the proportion of high or low grade comments in this month increased or decreased compared with the last month, and the comparison index in the next month also increased or decreased compared with this month, since that we believe that this index has an incentive effect on the change of the comparison index, and this formula is positive at this time. For another example, the proportion of high-grade or low-grade comments increased compared to last month, but the comparison index in the next month decreased compared with this month, base on the fact, we think that this index has an inhibitory effect on the change of the comparison index in the next month, then result of A is negative. This result has the property of positive when an index has an incentive effect on the other index, negative when an index has an inhibitory effect on the other index, and tending to zero when they have no relationship. By plotting the value of result of A for each month, it can be obtained from the data distribution whether there is a significant positive or negative correlation with the data $(x_m - x_{m-1}) \cdot (y_{m+1} - y_m)$

In order to quantify the results more convincing, this paper used operator B to sum results of A value of each month to analyze the comprehensive results of all months according to the calculated values. the result of B was a large positive number could show that the change of the proportion of high / low grade comments obviously played a positive role on the change of the compared index. Oppositely, it can be shown that the change of the proportion of high / low grade comments obviously played a negative role on the change of the compared index. Due to the characteristics of result of A, if there was almost no relationship between the two, result of B would tend to zero.

For instance, The impact of the proportion of high-grade comments and its change on the total number of comments in the next month of product "remington ac2015 t|studio salon collection pearl ceramic hair dryer, deep purple" was calculated by

$$\begin{aligned} &A(b_{Hm}, S_{m+1}) \\ &B(b_{Hm}, y_{m+1}) \end{aligned}$$

furthermore, as following figure shows:

4.2.6 Relation between rating and review in one record

In common sense, there was a strong correlation between star ratings and the emotion of the comments. It could be confirmed the above view by comparing the emotional score and the emotional score with the star rating: comments with significant positive emotions were likely to mean higher star ratings, in which customers used words "good", "perfect", "love" to express satisfaction with the product; additionally comments with significantly negative emotions were likely to mean lower star rating, in which customers often use "bad" and "disappointed" to express dissatisfaction with the product.

The gap between star rating and emotional score e could be defined as:

$$e(t_i) = |S(t_i) - E(t_i)|$$

Where S indicated the star rating of comment and E indicated the emotional rating of comment. The gap between the emotion of a certain comment and the star level could be evaluated according to e. The e values of all comments in the three data set were calculated and the results

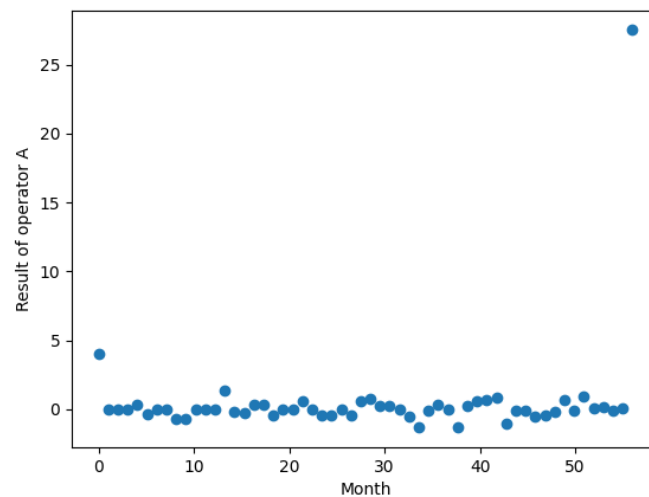


Figure 6: Relation between Result of operator A and month of "remington ac2015..."

are shown in the following figure: The difference between star rating and emotion rating was less than or equal to 1 is about 70 percents of the total and the emotion of these reviews was basically the same, while the comments with the difference greater than 2 were few, indicating that there was not much difference between emotion rating and star rating for most reviews. In addition, the distribution of star rating and emotion score was plotted, as shown in the figure below: As could be observed from the figure, the star rating and emotion scores had similar peak shapes in the high and lower segments. Since the more reliable the emotion score was, the stronger the emotion (which is clearly found in the comments), the result also showed that the comments with obvious positive or negative feelings also tend to be biased towards both ends. The emotion trait vector $V_E(t_i)$ was a more accurate description of the comment emotion than the star rating. Mining the potential relationship between the emotional feature vector and the star rating. this paper took the emotional feature vector of each comment as the attribute and the star score as the category (only 3 categories, 4 and 5 high, 3 medium, 1 and 2 low) and constructed a decision tree classifier to classify the emotional feature vectors and got the results: The classifier had high accuracy in the three data sets, indicating that the emotional feature vector could be used as an important indicator to evaluate the star rating, the positive words (such

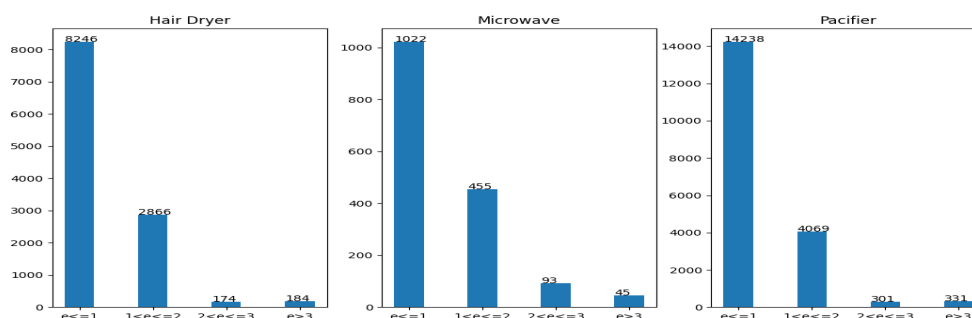


Figure 7: e distributon

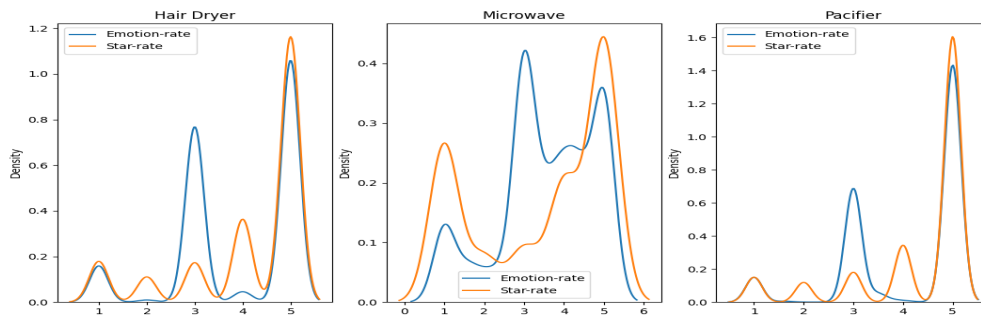


Figure 8: Distribution of star rating and emotion scores

parameter	Hair dryer	Microwave	Pacifier
Test sets proportion	0.3	0.3	0.3
Accuracy	80 percents	72.8 percents	83.5 percents
Features importance	[0.46 0.16 0.08 0.30]	[0.62 0 0 0.38]	[0.47 0.11 0.08 0.34]

Table 5: Results of emotional classification

as "good", "perfect", "love", etc.) and negative words (such as "don", "didn", "doesn", etc.) were important features, and the comments with high positive word content could be found; the comments with high negative word content could easily be classified into low scores. Reviews containing these words had larger associations with star ratings.

By analyzing the key phrases of a product model, it could be concluded that "compliment" reviews describing the excellent functions of the product were likely to have a high star rating, while "complaint" reviews with complaints about problems and defects in the product were likely to have a low star rating. give an example, Key phrases extracted from all five-star reviews of the hairdryer with the model "conair 1875 watt tourmaline ceramic hair dryer". Many of the phrases reflect a customer description of the advantages of the product, Such as "light", "weight", "quickly", "quiet", "light", "powerful" et al. Similar results also appeared in the high-score reviews of other products; when extracted key phrases from all one-star reviews of the model microwave, "samsung smh1816s 1.8 cu. ft. stainless steel over-the-range microwave", The phrases that reflect the customer's problems and complaints about the after-sales service, such as "problem", "error", "repair", "survice", "repair", "cost", "problem", "warranty" et al. These reviews reflect the multiple shortcomings of this type of hair dryer, Similar results also appeared in low-score reviews from other products.

4.3 Results

As to whether a specific user rating affected the number of comments and favorable ratings in the future, analysis shows that there was no obvious correlation between the two. In these indicators, the indicators of each month were randomly distributed around zero, positive and negative, according to the indicator definition, which means that they sometimes had small positive correlation, sometimes had small negative correlation, however, overall tend to be around zero, that is, there was no obvious relationship between them.

Since user reviews were closely related to user ratings, they were closely related according to analysis. This paper used the text analysis of the text reflects the emotional assessment to 1-5 points, which was corresponded to the user rating. According to statistics, three kinds of

products in the difference between the text score and the user rating were less than or equal to 1 points of comments were 71 percents, 75 percents, 81 percents, and three kinds of products in the difference between the text score and user rating were less than or equal to 2 points were 91 percents, 97 percents, 97 percents respectively. It could be shown that the majority of the comments of text and the user rating was basically consistent, only individual comments association is weak.

5 Further discussion

5.1 Contribution

In terms of text information in the data set, this paper first conducted the word frequency statistics on the text to find the high-frequency keywords. In addition, TFIDF method was used to extract the low frequency keywords in the text. High frequency keywords screening into emotional words and feature words were classified. After statistics and vectorization of emotional words, emotional feature vector were given and further calculated to emotional score of the text. Emotional scores reflected the emotion shown in a text. After the statistical processing the set of keywords in the comments was gained. Using the FP-growth algorithm, it could get the key phrases needed, that is, the key information in the text information. So that the text features were extracted. Text features could reflect the overall emotional tendency of the user as well as the most common advantages or disadvantages of the product.

Under consideration of voting information, user identity (vine) and the information whether purchase, this paper designed a special algorithm as reliance, representing the importance of the comments in information analysis. In order to ensure that the final analysis results objective and fair, more comprehensive and accurate.

According to the extracted text features and other data, the CSI was created of each product based on various indicators. In addition, according to the number of good comments and bad comments this paper extracted for each month (according to the star rating) and conducted a correlation analysis on the number of comment this month and the next month, so as to quantitatively determine whether the existing reviews would affect the number of comments and the praise rate of users in the future. two correlation indexes were used to analyze the correlation of the two, which could see the correlation from the data distribution, and also quantified the correlation size from the calculated correlation indexes.

In the text feature extraction, this paper re-evaluated the closer rating of the emotion at 1-5 points based on the emotion displayed by the text words. Analyse of the difference between the emotion score and the rating given by the user comment. According to the distribution of the difference, it could be vividly shown whether the user comment is closely related to the rating given by the user.

For commodity review data, the paper designed a series of models, algorithms and indicators, extracts the useful key information from the comments, assesses the general attitude of the user from a large amount of data, at the same time analyzes the user comment attitude and the correlation of multiple factors, analysis result were of great significance to business in the future.

5.2 Conclusion

By analyzing the following conclusions from user ratings, user written comments, and vote numbers:

1. Hair dryer products: The overall rating of hair dryer products was high. According to extraction of text vocabulary, many users that the hair dryer had good heating function, strong wind power and many users liked its light and lovely appearance, relatively small noise. But at the same time, some people thought that the hair dryer was too hot to unbearable.
2. Microwave oven products: From the distribution of user satisfaction index of each model of microwave oven, only model samsung smh1816s 1.8 cu.ft. is significantly lower than that of other models. Based on the text, numerous reviews showed that it was of poor quality and had poor aftermarket warranty service. However, other microwave products were generally rated well.
3. In terms of baby pacifier products, the vast majority of people gave them high praise, only a few people gave them few opinions. In our comprehensive evaluation index, users' evaluation of this kind of products is quite high, and users basically had no general counter-interview. Generally speaking, users are very satisfied with this kind of products.
4. Analysis showed that there is no obvious correlation between a specific user rating and the number of comments and favorable ratings in the future. In correlation indicators, the indicators tend to be around zero, which mean that there is no obvious pattern between them.
5. According to the results of analysis, user reviews were closely related to user ratings. According to statistics, three kinds of products in the difference between the text score and the user rating was less than or equal to 1 points of comments is 71 percents, 75 percents, 81 percents, respectively. It could be seen that the majority of the comments of text and the user rating is basically consistent.

References

- [1] F. Yamout and R. Lakkis, "Improved TFIDF weighting techniques in document Retrieval," 2018 Thirteenth International Conference on Digital Information Management (ICDIM), 2018, pp. 69-73, doi: 10.1109/ICDIM.2018.8847156.
- [2] Zhu Liye and Yuan Deng-hua, "When are the buyers willing to post comments on an on-line feedback system? The moderating effects of online interactivity between buyers and sellers," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), 2016, pp. 1-5, doi: 10.1109/ICSSSM.2016.7538485.
- [3] M. Aufar, R. Andreswari and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1-7, doi: 10.1109/ICoDSA50139.2020.9213078.
- [4] A. Razdan and S. S, "Hate Speech Detection using ML algorithms," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9670987.

Team's analysis and results

Dear Marketing Director of Sunshine Company:

Hello, we are glad to write for you. First of all, we sincerely thank you for choosing us to serve you. We are very pleased to tell you that the data analysis of your three categories of goods based on user ratings and reviews has been completed. Now let us tell you more detailed.

The first is the hair dryer products. The vast majority of the comments of the hair dryer products are five-star reviews, but according to our analysis of the text comments, in fact, nearly half of the people revealed their dissatisfaction in the text comments. The comments of these people actually only correspond to the medium rating level. From our extraction of text vocabulary, we found that there are many users think that the hair dryer heating function is good, strong wind, which shows that your hair dryer performance is very good. And there are a lot of people like its light and lovely appearance, small noise and other characteristics. But at the same time, many people mentioned that the wind is too hot, and we suggest that it can be optimized in this respect. Among hair dryer products, the most popular model is remington ac2015 t | studio salon collection pearl ceramic hair dryer, in our evaluation indicators, users of it gave very high evaluation

About microwave oven products, only half of the people to the five-star praise, there are still nearly thirty percent of people even gives a bad review. But in fact, according to our analysis of the text comments, overall most people are not that disgusted with your microwave oven, the vast majority of people's actual attitude on microwave oven is that it still can be improved, corresponding to about 3-5 stars. From the analysis of the text comments, we found that many people have mentioned that the microwave oven has a fault, which needs to be repaired. It seems that this problem needs to cause your attention, and it is suggested to solve similar hidden dangers as soon as possible. Microwave oven products model "samsung smh1816s 1.8 cu.ft.stainless steel over-The-range microwave" is the most problematic, a number of reviews show that your microwave is of poor quality, often bad and has poor aftermarket warranty service, and we think this is a failed product.

The distribution of baby pacifier products is similar to the evaluation distribution of hair dryers. The vast majority of people gave five-star praise and praise, and only a few people put forward some small opinions. In our comprehensive evaluation system, most of the product users are quite high evaluation of it, we hope you will continue to make persistent efforts, the business is booming.

In addition, according to our data analysis, in fact, the user rating of the product will not have a significant impact on the number of comments or the praise rate in the future, so there is no need to worry too much about the low rating in the past.

The above are some opinions and product conditions summarized according to the comment data. For details, you can also check our analysis report. Thank you again for choosing us. I wish you a prosperous business.

Sincerely yours,

Your team

Appendices

Core python code

```

for i in range(len(date)):
    if date[i] not in dic_date.keys():
        dic_date[date[i]]=0
    dic_date[date[i]]=dic_date[date[i]]+1
    if date[i] not in dic_rating.keys():
        dic_rating[date[i]]=0
    dic_rating[date[i]]=dic_rating[date[i]]+rating[i]*reliance[i]/dic_reliance[date[i]]
    if date[i] not in dic_emo.keys():
        dic_emo[date[i]]=0
    dic_emo[date[i]]=dic_emo[date[i]]+emo[i]*reliance[i]/dic_reliance[date[i]]

rating = np.array(rating)
emo = np.array(emo)
a = np.multiply(rating, emo)
R = (np.mean(a) - np.mean(rating) * np.mean(emo)) / (np.sqrt(np.var(emo) * np.var(ra
for i in range(len(pur)):
    if pur[i] == 'N':
        pur[i] = 0
    else:
        pur[i] = 1
    if vine[i] == 'N' or vine[i] == 'n':
        vine[i] = 0
    else:
        vine[i] = 1
print(sum(pur))

weight_vine1 = 1 / (sum(vine) / len(vine)) * (1 / 5)
weight_vine0 = 1 / ((len(vine) - sum(vine)) / len(vine)) * (4 / 5)
weight_pur1 = 1 / (sum(pur) / len(pur)) * (9 / 10)
weight_pur0 = 1 / ((len(pur) - sum(pur)) / len(pur)) * (1 / 10)
for i in range(len(date)):
    if date[i] not in dic_date.keys():
        dic_date[date[i]] = 0
    dic_date[date[i]] = dic_date[date[i]] + 1
    if rate[i] >= 3:
        if date[i] not in dic_rate_high.keys():
            dic_rate_high[date[i]] = 0
        dic_rate_high[date[i]] = dic_rate_high[date[i]] + 1
    if rate[i] <= 3:
        if date[i] not in dic_rate_low.keys():
            dic_rate_low[date[i]] = 0
        dic_rate_low[date[i]] = dic_rate_low[date[i]] + 1

# The mining of crucial word groups

path1 = "path of data set"
Data = pd.read_excel(path1, usecols=[4, 5])
Data.set_index('Star_rating', inplace=True)

```

```
Kwords = Data.loc[1].dropna().values.T.tolist()[0]

Cf = CountVectorizer(min_df=7)
Kf = Cf.fit_transform(Kwords)
Klist = Cf.get_feature_names_out()

# FP-growth
Kw_list = []
for s in Kwords:
    Kw_list.append(re.sub('[' + string.punctuation + ']', '', s).split())
fp = FP(7)
wg_HF = pd.DataFrame(fp.fit(Kw_list, Klist))
```
