Research on Online Review of Commodities

Xintian Cai[a, *], Wen Wen

International School Beijing University of Posts and Telecommunications, Beijing, 100089, China

[a]caixintian@bupt.edu.cn

*Corresponding author

Keywords: data mining, SPASS, comment text

Abstract: First, we performed data preprocessing including cleaning and filtering the data, performing independent analysis on the data, etc., removing redundant tags, extracting high-frequency words, classifying them using LDA natural language classification model, and finally identifying 7 topics. In order to analyze the relationship between the review text and the rating, we built a review model and qualitatively and quantitatively analyzed the characteristics of the text review. We analyze it through lexical network analysis, a topic model based on word2vec, and sentiment classification of unsupervised reviews. Then we built a comprehensive product scoring model to score products. The product comprehensive analysis model combines the analytic hierarchy process (AHP) and entropy weight method to improve accuracy; the model also introduces sales volume adjustment parameters to measure the sales popularity of a product among similar products.

1. Introduction

With the popularity of online shopping, online shopping platform, such as Amazon, has established a rating and comment function [1]. Through ratings and reviews, users can express their opinions and suggestions on products. The company can understand all aspects of products, accurately predict future trends and formulate appropriate strategies [2]. The analysis and modeling of information such as commentary and time are of great significance to the formulation of the company's strategy [3].

2. Data processing and visualization analysis

2.1 Data preprocessing

In order to ensure the accuracy of the conclusion and the robustness of the model, it is necessary to conduct a series of data preprocessing for outliers and missing values.

2.1.1 Data cleaning and filtering

Because of the inability to identify the impact of some products in the comments, it may contain some duplicated data, erroneous data, and random data in the analysis of user comments. Based on the Python platform, we made some adjustments to the data compilation method [4]. After completely removing the data bar containing the garbled data, we use SPSS to process data that contain missing values of some specific variables [5].

Based on data analysis of three types of products, combined with data statistics, we can describe the analysis.

(1) The quarterback data of help and sum are much larger than those of other variables. Based on the particularity of this article, the difference of discount variables should be reasonable.

(2) After considering the number, we get the effective data, including the purchase data of 18905 nipples, the purchase data of 1614 micro microwave ovens, and the purchase list of 11451 hair driers.

2.1.2 Independence analysis

There is a certain correlation between data. Such information is often duplicated. Based on this, we need to delete redundant variables and independent variables.

(1) Part of the information is invalid: for example, the purchase of odd numbers, commodity codes, and information coding data are often invalid. Therefore, we have eliminated them here.

(2) Since almost all MKP variables in this problem are filled with us, we need to delete this redundant variable directly.

(3) The user name does not have the actual meaning. Here we delete the data and delete the same variable code.

(4) We consider that the PID of variables is equivalent to the names of variables. After comparative analysis, we believe that the names of variables are more reasonable.

(5) We transform the date sequence and transform it into continuous natural day series, for example, transforming 8/31/2015 into "20150831", so that it can be conveniently sorted according to time, and is conducive to the prediction of time series in subsequent models and the aggregation of data in time dimension.

2.1.3 Data visualization analysis

(1) We made an average aggregate measure according to the time scale of the score, and took the Pacifier dataset with more samples as an example to draw the curve of the score changing with time.

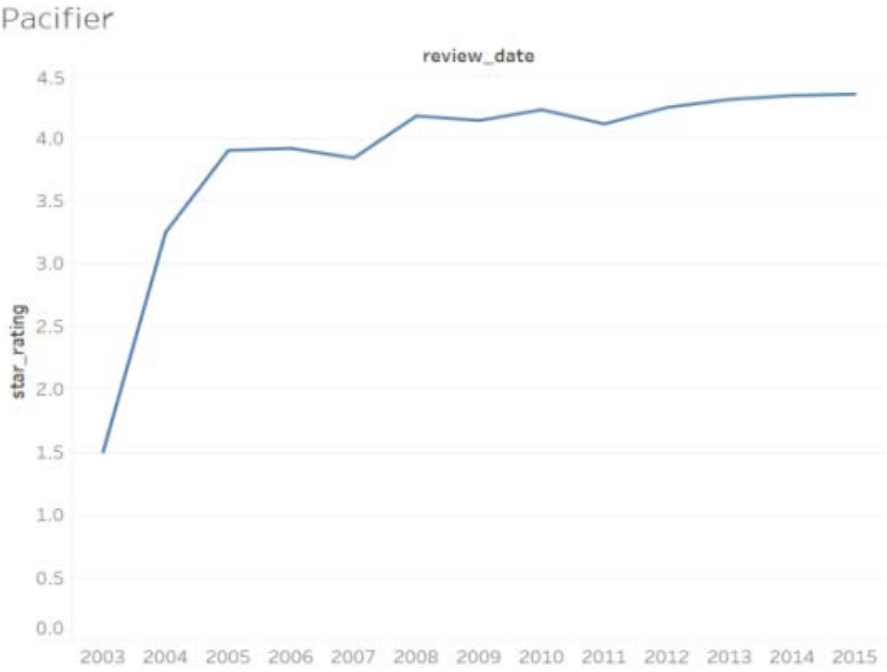

Figure 1. The curve of grading with time

As can be seen from the above figure, the score increases with the increase of time, but tends to be gentle, and basically no longer changes after 2010, indicating that the products have a good trend over a longer time span[6].

(2) We analyzed the relationship between the sales rate and the sales volume, and found that the two were correlated. Sales volume rose rapidly when the favorable rate increased.

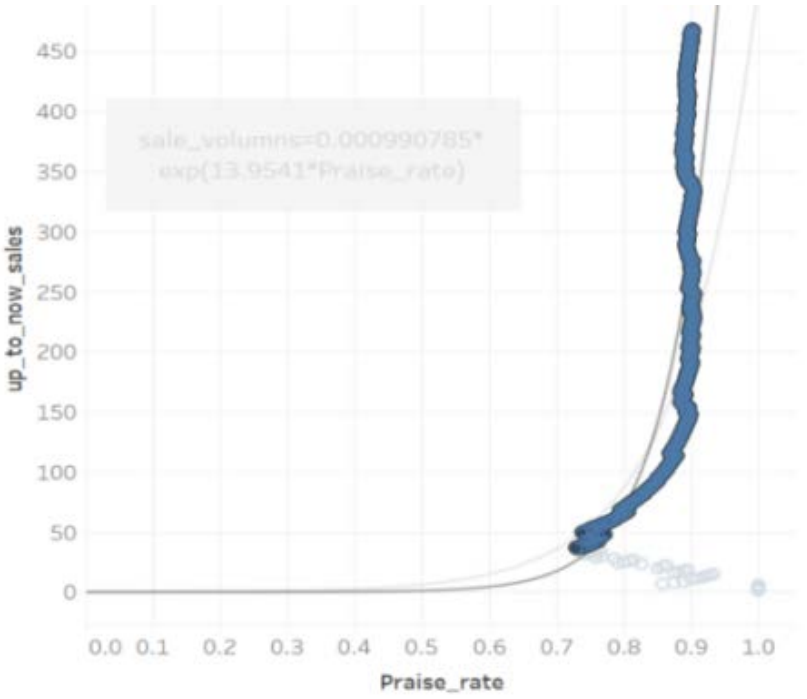

Figure 2. Relationship between favorable ratings and sales volume

The exponential function can be used for regression analysis. The results show that the sales rate increases rapidly with the increase of the favorable rate.

We have roughly processed the evaluation text in hair_dryer, microwave and pacifier tables. We extracted the words and removed the words which are not related to the subjective emotional tendency, such as "a" and "the". We have counted the number of the higher frequency words, and we can conclude that the vocabulary in the evaluation contains a lot of "like" and "love"[7]. In order to further classify and analyze these words, we use pacifier as an example to classify LDA words by using LDA natural language classification model. Finally, we identify 7 themes, and we can give an example of one of them.

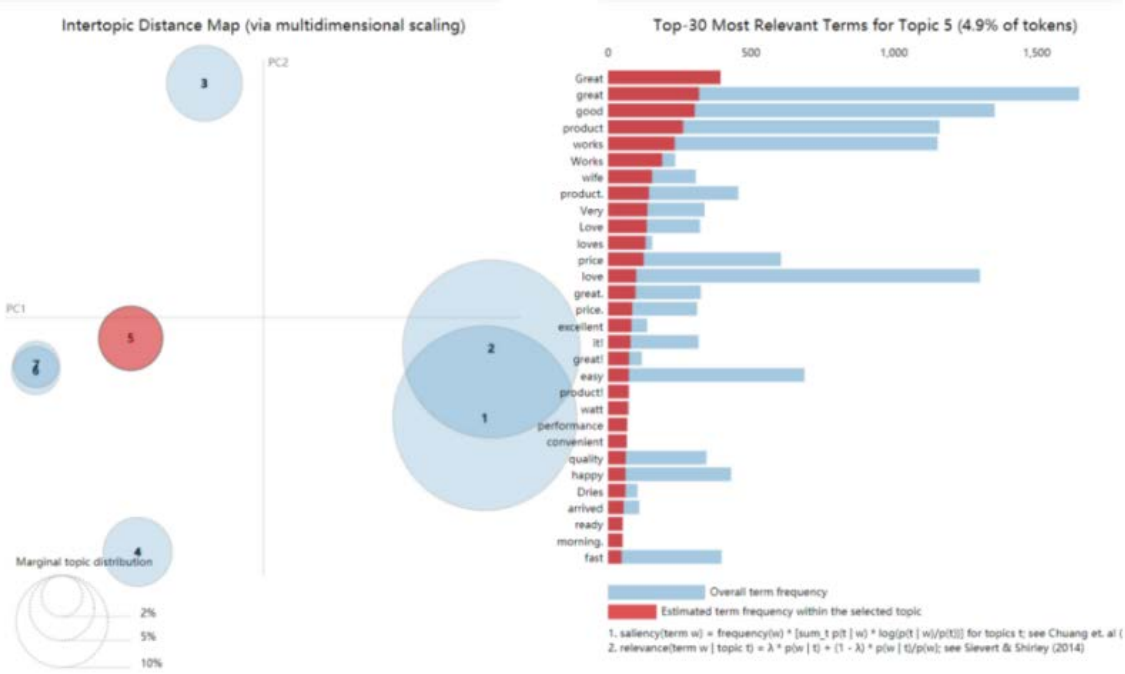

Figure 3. Seven themes were identified and their examples

By analyzing the related words of this topic, we can see that there are many words that express appreciation like "Great" and "excellent". Therefore, the theme clearly indicates that customers have highly praised the products.

2.2 Natural Language Processing

The three datasets provide three kinds of data, product_title, review_headline and review_body respectively, representing the title, comment title and comment content of the product respectively. For the comment title and comment content, the content of the free format text data 1 is rather mixed. In order to extract their effective characteristics, analyze the potential relationship between the emotion and the different data types reflected by them. We need Natural Language Processing. We have referred to the work of Filipe R. Lucinia and Leandro M. Tonetto, and have adopted the following preprocessing work for text data.

2.3 Spell correction for words in the text

W_f Used for spelling mistakes. W_t According to the Bayes formula, we can get the correct form corresponding to the wrong word.

$$P(W_t \mid W_f) = \frac{P(W_f \mid W_t) \times P(W_t)}{P(W_f)}$$

Here $P(W_t \mid W_f)$ expressing wrong words, W_f correct words are W_t probability. $P(W_f \mid W_t)$ Express will W_t write in error, W_f probability is expressed as the probability that the correct word is transformed from 1 to 2 letters 2, assuming that the probability of each form is equal. $P(W_t)$ Right word is W_t for probability, we replace it with the frequency of use (the frequency of words appearing in the corpus), and for every wrong word, W_f is a constant. Therefore, $P(W_f \mid W_t) \times P(W_t)$ biggest word is probably the right word.

2.4 Identify the special meaning phrases

Identify the special meaning phrases such as names, place names, and organization names, and replace them with words such as "person", "location" and "organization". We use Stanford NLP toolkit 3 to identify the text content.

2.5 Unify word form and filter words through parts of speech

In order to prevent the same word from being unable to be classified as a case of difference in size, we capitalize all capital letters in the text with lowercase letters. In addition, we also use the Python NLTK toolkit to label the parts of speech. We only retain the adjectives, adverbs and nouns in the text. Delete other words that are less effective in reflecting textual features. Finally, we restore the retained vocabulary to its basic form (for example, restore the complex number to singular).

2.6 Remove and delete specific vocabularies

We remove repetitive comment headings and comment content, and delete the stop words in the text. In addition, we discard the words with less than 2% frequency in the text, and extract 150 keywords from each dataset.

3. The establishment and solution of the model.

3.1 Comment model

3.1.1 Reviews lexical relations network

We extracted the key words of the comment text. However, after reviewing the text, we still have a large vocabulary. After that, we cannot directly analyze the correlation and distribution of the key

words. In order to solve this problem, we use the commentary keyword to construct the co-occurrence matrix, and analyze the lexical network relationship through Ucinet, a social network analysis software.

In the study of the visualization of co-existing network, Zhou Shuanlong pointed out that the network structure formed by the association of words according to certain rules presents complex network characteristics. By constructing the collinear network, we can analyze lexical clustering, lexical relations and lexical centrality.

The co-occurrence matrix means that by counting the number of word co-occurrence in a pre-specified size window, the number of co-occurrence words around the word is used as the vector. Of the current word. We can represent the co-occurrence matrix as follows: $CO = \{(w_i, w_j) \mid w_i, w_j \in W_s\}$

Among them, $CO = \{w_1, w_2, ..., w_n\}$ the keyword list is expressed. w_i Key words. n Indicate the number of keywords. (w_i, w_j) Express w_i and w_j the number of collinear edges, that is, the number of two keywords appearing together in the comment.

After obtaining the collinear matrix, we use Ucinet's built-in Net draw drawing software to generate commentary keyword network diagram to visualize the relationship between keywords and the network distribution of keywords.

Finally, we calculate the centrality of each keyword through Ucinet. C_r, proximity centrality C_c, ability centrality C_a, degree centrality C_d. We will focus on the four centrality of the keywords and the frequency of words appearing in the commentary. f_w As an index, the weight of the five indicators is calculated by entropy method. The formula is as follows:

$$\partial_j = \frac{d_j}{\sum_{j=1}^{m} d_j}$$

Among them, d_j it indicates the degree of redundancy of information entropy. m Indicates the number of indicators. Here, $m = 5$.

Then, based on the weight of the index obtained in the formula, we define the influence of keywords as follows:

$$W_e = \partial_1 \bullet C_r + \partial_2 \bullet C_c + \partial_3 \bullet C_a + \partial_4 \bullet C_d + \partial_5 \bullet f_w$$

In the follow-up to the paper, we will use the influence of keywords to divide the topic of the comment vocabulary.

3.1.2 Word2Vector word vector model based on comment text

In quantitative analysis, we are faced with the problem of transforming text format data into data that can be used for numerical analysis. Therefore, we use the Word2Vector model to transform comment text into word vector.

Word2vec includes two language models, CBOW and skip-gram. It adopts two training methods: level softmax and negative sampling. It is a two-layer neural network. Word2vec maps each word to the word vector in low dimensional space from the hidden layer. Besides, it introduces the context of words in the training process of word vectors, and it can express the distance between words (the similarity between words).

When analyzing the text data, we hope to dig out the correlation between the commentaries, so that we can cluster the comment words. If we translate words into points in the space, we can imagine the distance between the two points.

At the same time, each comment contains multiple words. Each word has different meanings and emotions. How to judge the overall mood tendency of each comment is a very important problem.

The word vector generated by word2vec can be directly added or subtracted. A very simple example is in two-dimensional space.

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$$

Therefore, by training the word2vec model, we can judge the text sentiment of this comment by adding the words of a comment.

3.1.3 Comment on topic classification

As discussed in the last section, we will comment on the conversion of text into vocabulary n. Then, we clustering the comment vectors. The time complexity of the K-means algorithm is O (n). The algorithm is very efficient for dealing with a large number of high-dimensional data. Therefore, we choose the K-means algorithm. However, the algorithm is sensitive to noise and outliers, and the clustering effect depends on the initialized clustering centers. First of all, we further filter the low-frequency words in the text. Secondly, we choose the influence of keywords *we* tallest k vocabulary (higher centrality and word frequency) is used as the initial clustering center.

We will $\{x_1, x_2, ..., x_m\}$ defined as a label free training set, each of which is x_i they all represent one. n Vector of dimension words, i. e. $x_i \in R^n$. M representing the number of words $\{\mu_1, \mu_2, ..., \mu_k\}$ express k a keyword with high influence k cluster centers of different categories. $\{c_1, c_2, ..., c_m\}$ To store and rank i, the index of the clustering center nearest to the data. The objective function of the algorithm is:

$$J(c_1, c_2, ..., c_m, \mu_1, \mu_2, ..., \mu_k) = \frac{1}{m}\sum_{i=1}^{m} \| x_i - \mu_{c_i} \|^2$$

Among μ_{c_i} representative and x_i nearest clustering center.

We assume $\{\ell_1, \ell_2, ..., \ell_k\}$ for the final clustering center after cycle, the cosine similarity of the cluster center and each commentary word is calculated.

$$s_i = \frac{x_i \cdot \ell_i}{\| x_i \| \| \ell_i \|}$$

We can get cosine similarity with each cluster center. s_i The number of words with the highest number. K by analyzing the clustering results, we can get the best comments on topic number and subject category.

3.1.4 Perceiving comment emotion

It is worth noting that the data set of the problem does not give the corresponding sentiment labels. Therefore, it is impossible to classify emotions by traditional methods such as naive Bias. Therefore, we propose an unsupervised sentiment classification algorithm based on word2vec.

First, in order to quantify sentiment, we refer to M. Lutfullaeva, M. Medvedeva and E. Komotskiy's theses to create a commentary tone dictionary and choose words that express positive or negative attitudes (emotional states). They were grouped according to the nature and degree of the emotion. Each group of tone words was assigned a parameter to convey emotion (-2 to +2). The negative number of emotional parameters expressed the negative attitude of the user to the product, and the positive number of emotional parameters expressed the positive attitude of the user to the product.

Secondly, for every comment $\mathrm{Re} = (\omega_1, \omega_2, ..., \omega_p)$ among them ω_i represents a single word. p Representing the total number of words in a comment. We call the word2vec model that has been

trained to calculate the vocabulary in each emotion level. ω_i The similarity of a single word can be expressed as:

$$S_{\omega_i} = (\sum_{j=1}^{4} s_j^1, \sum_{j=1}^{4} s_j^2, \sum_{j=1}^{4} s_j^3, \sum_{j=1}^{4} s_j^4)$$

Among them, s_j^k represents the first in a dictionary. K first in the emotional level. J vocabulary and ω_i. Then, we add and process the word similarity in a sentence. Finally, we get the similarity between a sentence and four emotion levels, and select the most similar emotion category as the category of the sentence.

$$S_{sentence} = \max(\sum_{i=1}^{p} S_{\omega_i})$$

4. Comment Model testing and problem analysis

4.1 Reviews lexical relations network

For the three data sets of hair dryer, microwave oven and nipple, we selected 100 commentary keywords respectively, and generated the comment vocabulary network diagram through Ucinet software.

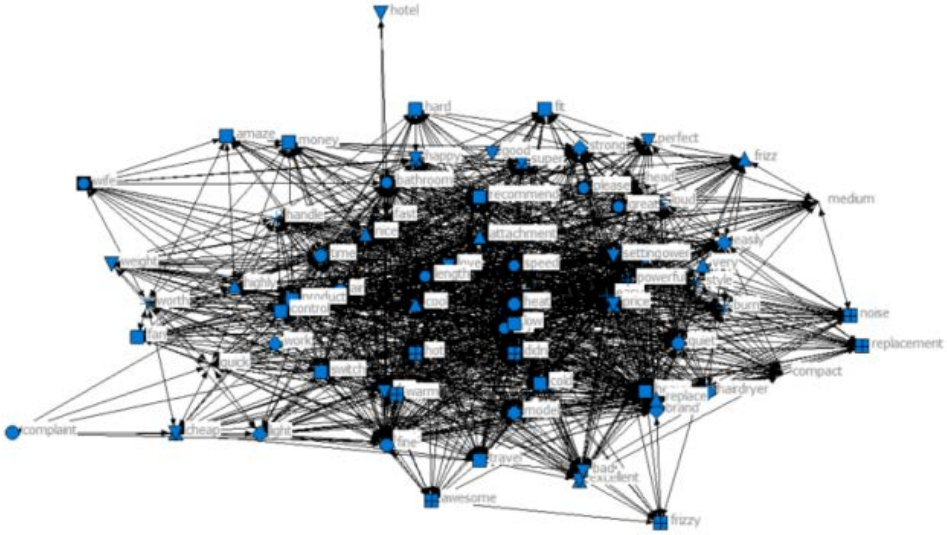

Figure 4. Network diagram of microwave oven Reviews

From this we can see that the description content of product characteristics, such as "heat", "speed", "length" and so on, are at the center of the comment network. At the same time, we can clearly see that the distance between product performance evaluation words such as "powerful", "quite" and "great" is relatively close.

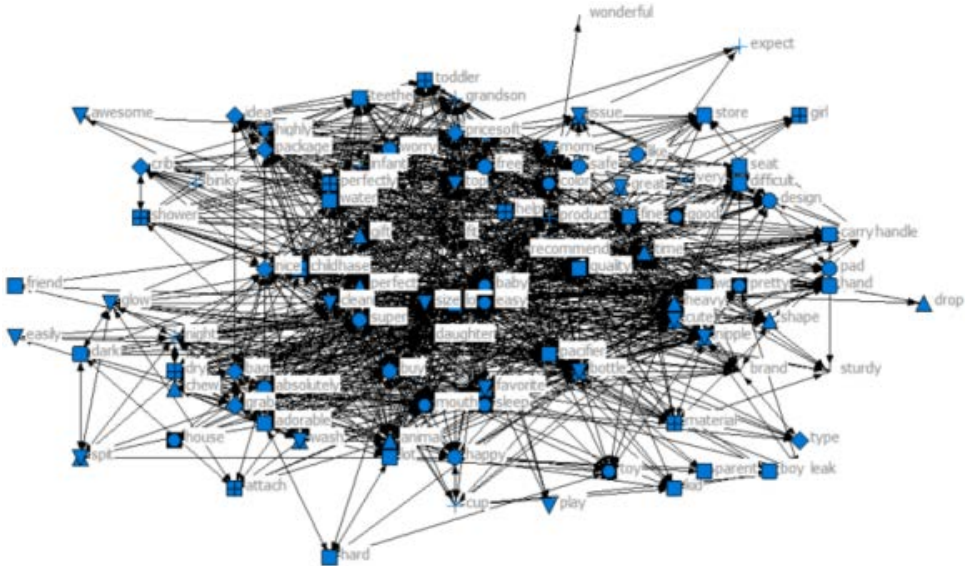

Figure 5. Pacifier network diagram

It can be seen that in the evaluation of the nipple product, the "fit" has the highest centrality, which reflects users' attention to the performance of the products when purchasing baby products. At the same time, we can clearly see that there are many positive words in the reviews of the nipple products, which are closer to the vocabulary describing the high performance products.

Then we calculated the top 10 most influential keywords for each type of product.

Table 1. The three most influential keywords in the three datasets

Hair dryer		Microwave oven		Pacifier	
Key	Influence	Key	Influence	Key	Influence
dry	47.5498	cook	48.3422	love	50.3295
powerful	34.6424	size	40.3244	baby	49.4698
great	32.6648	buy	35.4934	easy	48.0325
easy	30.2168996	light	31.2973	quality	45.0255
love	29.9312	like	30.2349	size	43.7748
fast	28.3608	service	29.3248	perfect	40.3444
switch	27.7646	price	27.6182	pacifier	37.9053
model	27.5462996	love	26.7393	like	36.6482
power	26.9164	easy	25.0724	free	33.7853
compact	25.5056996	clean	24.7558	daughter	30.6294

From this we can see that buyers of hair dryer pay more attention to the drying ability and usage mode of hair driers. For microwave oven, consumers are more concerned about the size, weight, price and related services of the products. Besides, pacifier buyers are keen to discuss the feelings of infants and focus on the quality and size of nipples.

4.2 Word2Vector word vector model based on comment text

In view of the textual relevance of the comment headlines and reviews, we merged the commentary titles and commentaries of the three datasets respectively as input training word2vec models.

Table 2. Word2vec training parameters

sg	size	window	min_count	negative	sample	workers
1	5	5	5	3	0. 001	4

By calling word2vec, we can see the most relevant word with a word.

Table 3. Examples of correlation between words

Word	Most relevant words				
good	great	decent	terrific	ecellent	perfectly
hot	warm	blows	cold	blow	softly

4.3 Comment on topic classification

Through many tests, we use the K-means algorithm to cluster the reviews of the drier data sets and the microwave oven dataset respectively into 3 topics. For the Pacifier dataset, we cluster 4 themes.

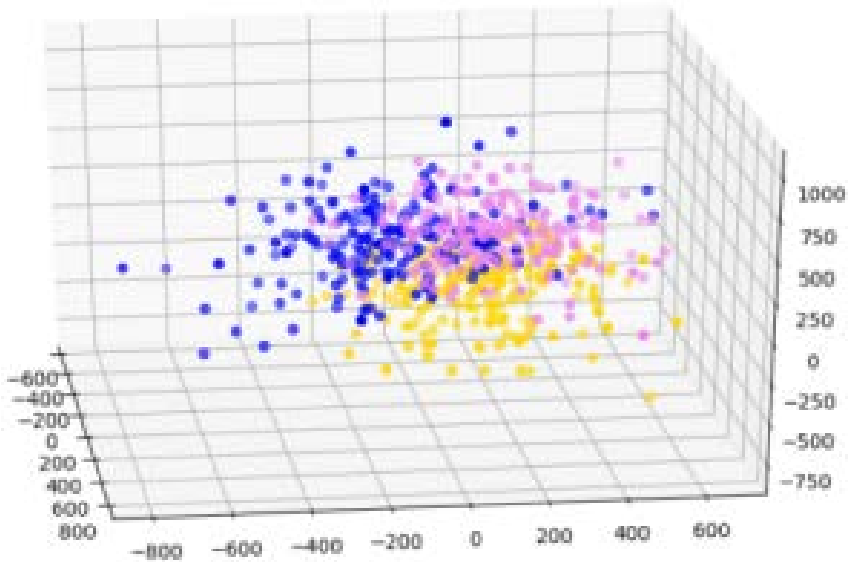

Figure 6. Cluster results of hairdryer Reviews

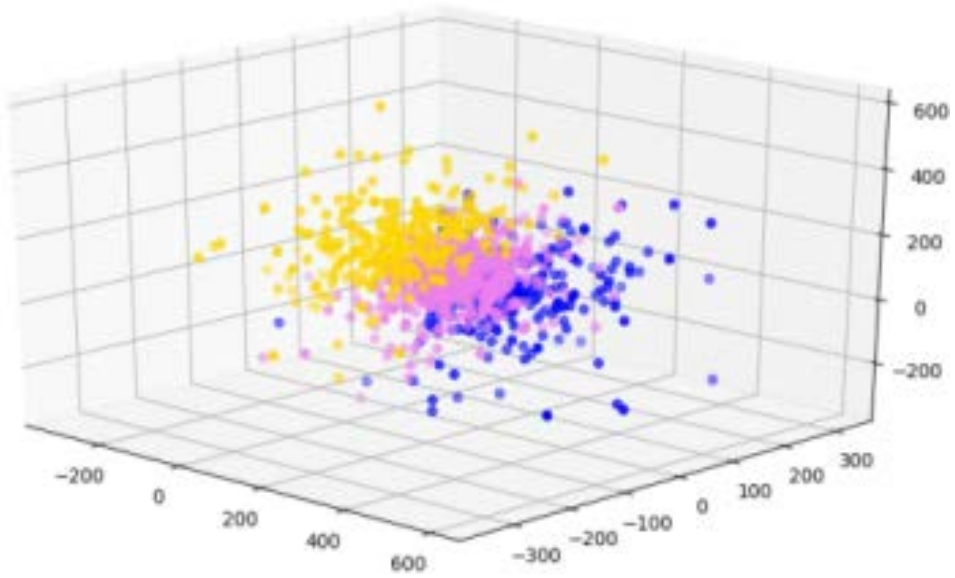

Figure 7. Clustering results of microwave oven Reviews

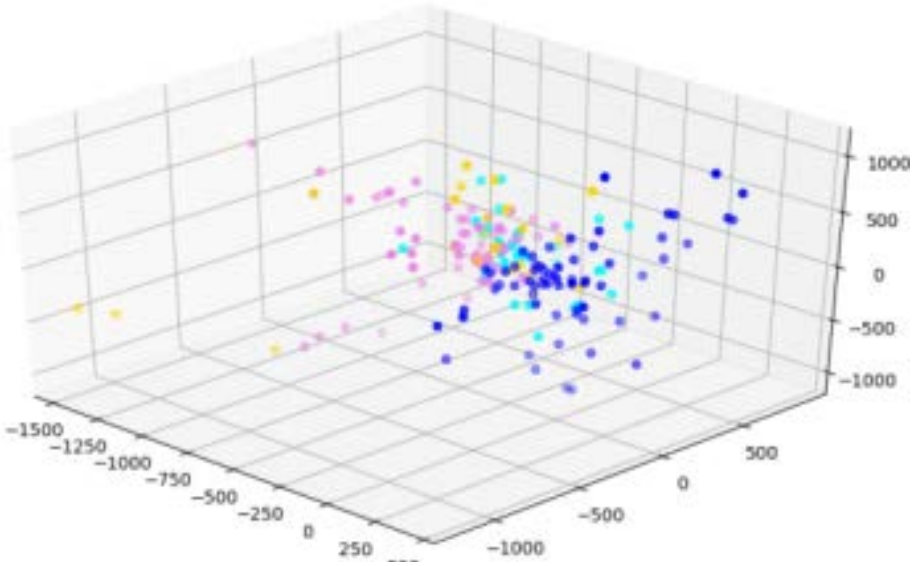

Figure 8. Clustering results of nipple Reviews

They are divided into three product types: the theme of the review, the key words and the coordinates of the center points of each topic.

Table 4. Comment topics

Hair dryer							
Review topic	Topic vocabulary					Cluster center	
1	speed	heat	wattage	time	portable	(81.24, -62.43, 190.17, 107.91, -208.19)	
2	affordabl	cost	investment	buying	life	(212.25, 54.82, 21.83, -163.99, -21.67)	
3	best	love	terrible	wonderful	nice	(138.76, 1.85, 69.29, -17.95, -162.84)	
Microwave oven							
Review topic	Topic vocabulary					Cluster center	
1	worry	fantastic	nice	thank	amaze	(41.85, 155.02, -157.27, -8.68, -67.73)	
2	fry	stove	electronic	boil	storage	(105.09, -91.09, 195.92, -164.31, -24.38)	
3	moderate	fancy	reliability	cupboard	locate	(-12.88, 183.77, 154.04, -11.13, 352.76)	
Pacifier							
Review topic	Topic vocabulary					Cluster center	
1	service	wonderful	perfect	favorite	handy	(-552.23, 5.65, 116.21, -142.54, -917.94)	
2	pattern	pacifier	milk	picture	pad	(-39.98, 116.45, -0.20, 650, -562.69)	
3	boy	daughter	son	buy	price	(-523.53, -23.09, 147.27, 469.46, -240.90)	
4	safety	baby	teeth	easy	bear	(-574.4394, 465.80, -243.76, 34.97, -461.48)	

We can conclude that the reviews of hairdryer products mainly focus on three aspects of performance, price and use experience of the hairdryer. The theme of the microwave oven reviews is the use experience, cooking function, and consumers focus on the reliability of the products. The theme of the nipple is divided into four aspects: the use experience, the product function and appearance, the children (using the baby) and the product performance. Users are very concerned about the safety of products.

4.4 Perceiving comment emotion

We turn the reviews into 4 types of emotions (from negative to positive). Through correlation analysis, we find that there is a certain correlation between the sentiment and ratings, but the trend is not the same.

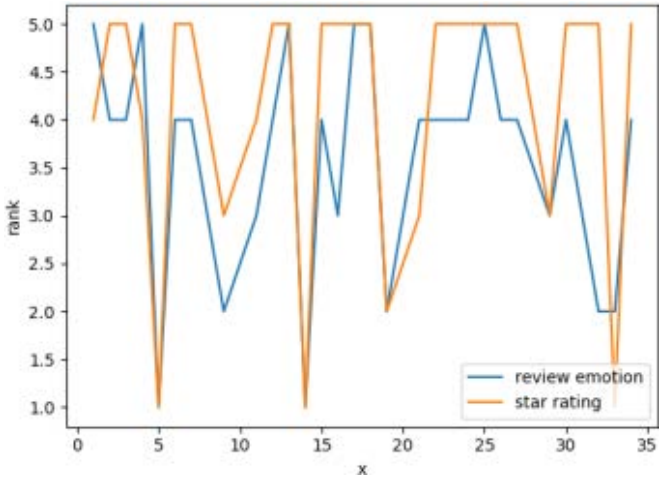

Figure 9. Comment and rating relationship map

References

[1] Liu Hua, Li jingqiang. The relationship between online reviews of experiential products and consumer wishes and sales volume-with price as the adjusting variable [J]. China's circulation economy, 2020, 34 (02): 83 - 91.

[2] Zhang Yanliang, Li Xiaozhe. Analysis of Tmall Store Online Comments Based on Case Reasoning [J]. Modern Electronic Technology, 2020, 43 (02): 57 - 59+63.

[3] Jingyong Xia, Gou Heping, Liu Qiang, Chen Lili. Research on Emotional Analysis of Online Curriculum Reviews Based on Topic Model [J]. Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), 2020, 34 (01): 54 - 56+61.

[4] Li Xiaojie, Jing Zhao, Wang Xiaoxin, Xing Chen. Analysis of Factors Affecting Online Shop Selection of Selected Products [J]. Jiangsu Shanglun, 2019(12): 42 - 44+47.

[5] Wei Sujuan. "The Impact of Online Shop Image of internet plus on College Students' Purchase Intention [J]. Journal of Wuyi University, 2019, 38 (12): 79 - 84.

[6] Guo Bin. Research on the Impact of Mobile Online Shopping Experience on Consumers' Intention of Repeated Purchase [D]. Nanjing University of Posts and Telecommunications, 2019.

[7] Dong Jingjing. Study on the Influence of Online Experiential Interaction between Consumers and Merchants on Their Purchase Intention [D]. Jilin University, 2019.