Problem Chosen
C

2020
MCM/ICM
Summary Sheet

Team Control Number
2010225

A Wealth of Data

Summary

With the advent of the era of big data, data analysis has become a powerful means of market competition. In this paper, we aim to analyze the sales data of three types of products in the market and user reviews to provide recommendations for the sale of products.

First of all, we analyze the difference in sales between products and the trend of change, which provides us with a direct view of the market situation.

Second, we perform word frequency statistics on user evaluations and use the results after removing stopwords as the basis for subsequent analysis.

Third, in word frequency statistics, we noticed that some words are related to the nature of the product. Statistics for these words are critical to improving a product or launching a new product.

Fourth, many words express user emotions in user reviews. These words have a lot to do with their star ratings.

Fifth, we analyzed the relationship between the star ratings of previous users and the reviews over some time.

Sixth, through the processing of natural language, we rate each comment according to its sentiment. We also compare scores with the user's star ratings to verify their accuracy.

Finally, we combined data from product sales records and user reviews to identify a way to measure products.

Keywords: data analysis;natural language processing;statistics

1 Introduction

1.1 Background

1.2 Program Restatement

1.3 Solution Introduction

2 Data Processing and Conclusion Analysing

2.1 Sales-Based Analysis

We screened sales data for three types of products, and selected the three products with the highest sales volume to draw sales-time line charts. The results are as follows.

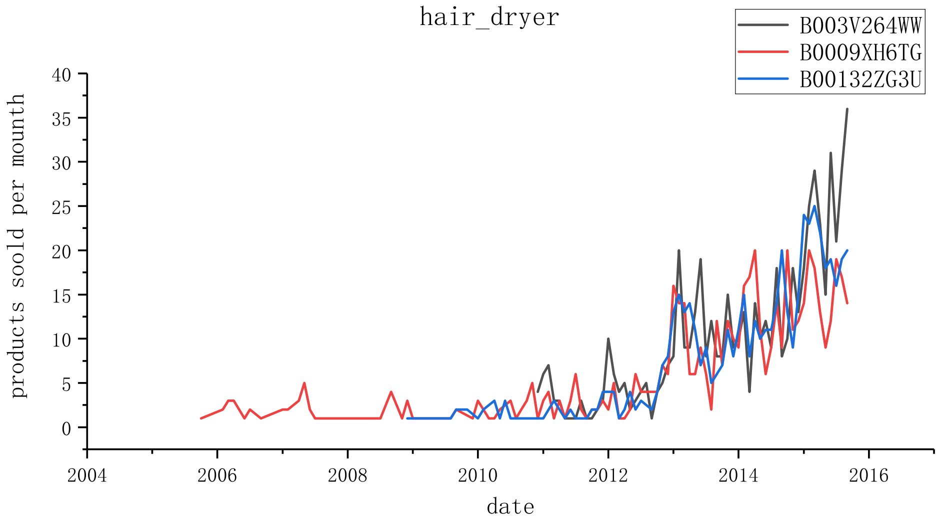

Figure 1: hair dryer sales

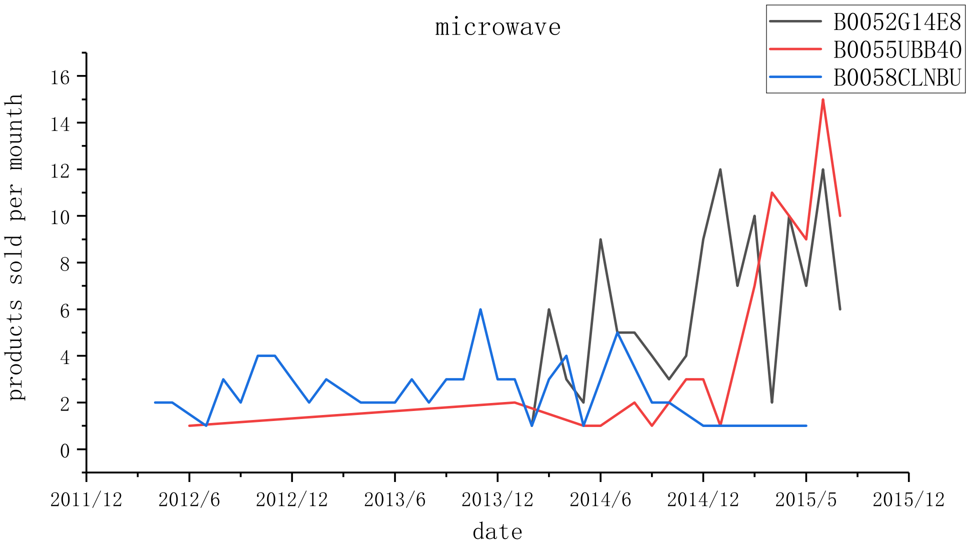

Figure 2: microwave sales

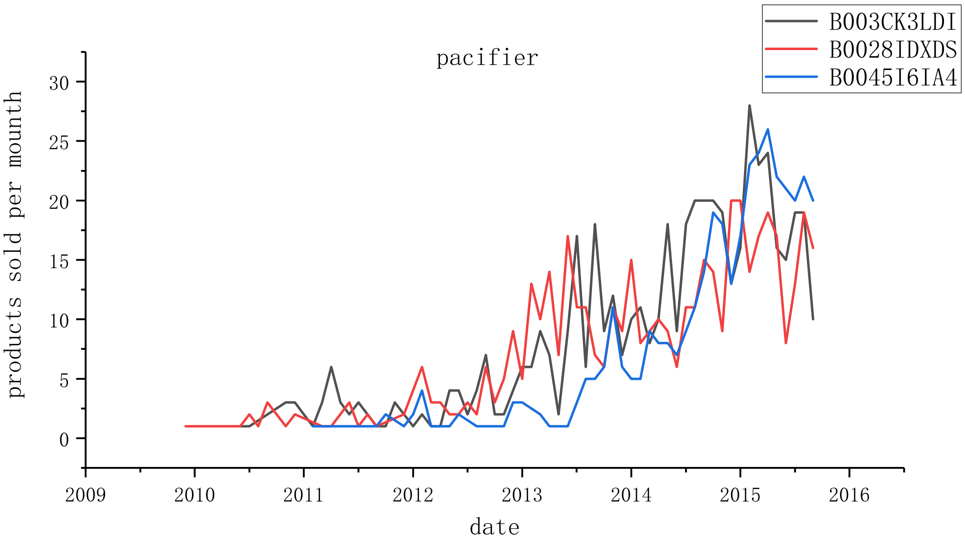

Figure 3: pacifier sales

By simply analyzing these images, we have reached some conclusions:

- The total sales of the three products are significantly different. The sales of hair dryers and pacifiers are much larger than microwave ovens, and the market competition is more intense.

- For hair dryers, some products have been sold early, but their sales have grown slowly. Other products have only begun to appear in recent years, but sales have grown rapidly, quickly surpassing older products. The total sales volume of hair dryers has increased rapidly in recent years and has great market potential.
- The overall sales of the microwave oven market are low with relatively large fluctuations. Some products maintain low monthly sales and still have a downward trend. Some new products have good sales as soon as they hit the market. This may be because buyers of microwave ovens value new features of the products more.
- The sales of pacifiers are also generally increasing, but most of the best-selling products have a long sales history. This may be because for products such as pacifiers, people prefer products that are safe and reliable, and are more willing to choose a brand that sells well and is well received.

However, sales-based analysis is not enough. Below we will analyze the data from some other angles.

2.2 Count the Word Frequency Statistics

To analyze the review of consumers, we divided the review into high score (3-4 points) and a low score (1-2 points). First of all, we synthesize the reviews of the three products and find out the word frequency. Then we deal with the high-grade reviews and low-grade reviews of the three products respectively. The python code is attached. For the python code, see the appendix.

2.2.1 Remove Stopwords

Because some commonly used words are used quite frequently, such as a the, he, etc. in English, almost all comments contain these words. If they are used as keywords, almost all comments will be indexed, and there is no differentiation, so these words are generally removed directly, not as keywords.

2.3 Advantages and Disadvantages of Three Products

After the word frequency is counted and the stop words are removed, the reason why people like/dislike the product is obtained by analyzing the words with high word frequency. By counting the number of times that people mentioned the reason, we summed up 5 factors of success or failure of these products.

2.3.1 Pacifier

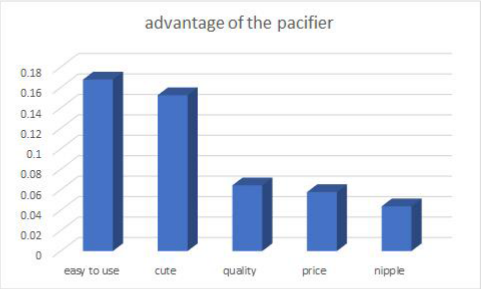

Figure 4: The proportion of reasons people like the pacifier

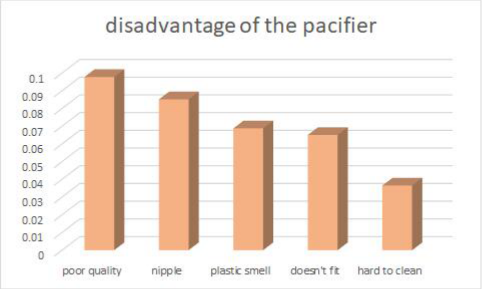

Figure 5: The proportion of reasons people dislike the pacifier

As shown in the chart, ease of use is the most outstanding advantage of the product. About 17% of high scoring customers mentioned this feature in their comments. After that is cute, quality, price and pacifier design.

On the other hand, about 9.5% of people who don't like the product complain about the poor quality. Although 4% of people who like the product think that the nipple fits well with there baby's mouth, 7% think that the nipple leaks, or is too short and hard. Besides, people can't bear the plastic smell, which is an important reason why they don't like it. Furthermore, some parents think that the pacifier doesn't fit well. At last, about 3% of people find it hard to clean the pacifier, which makes them give low ratings.

2.3.2 Hair Dryer

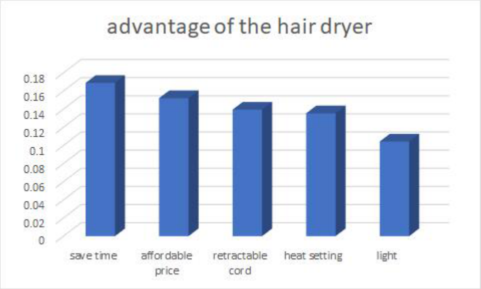

Figure 6: The proportion of reasons people like the hair dryer

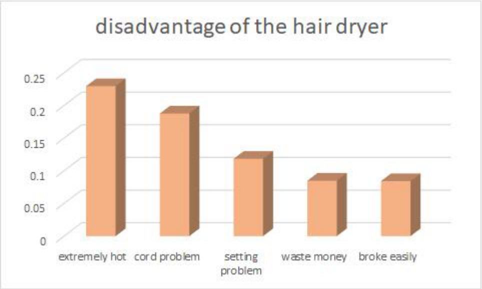

Figure 7: The proportion of reasons people dislike the hair dryer

According to the figure,16% of people think it saves their time, while 14% were attracted by the affordable price.12% think the retractable cord is a good design. Besides, the hairdryer is light to carry, so 10% of customers give a high rating. At last, the heating setting design is also attractive.

From the figure, we can know that about 21% of people can't stand the extreme hot of the hair dryer.17% complained that the cord broke quickly.10% find that the setting mode has some problem.7% think buying the hairdryer is a waste of money, and the same percentage of people wrote that the hairdryer broke quickly.

2.3.3 Microwave

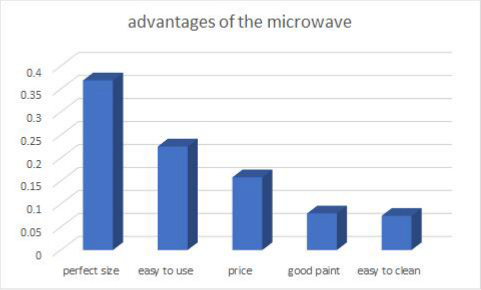

Figure 8: The proportion of reasons people like the microwave

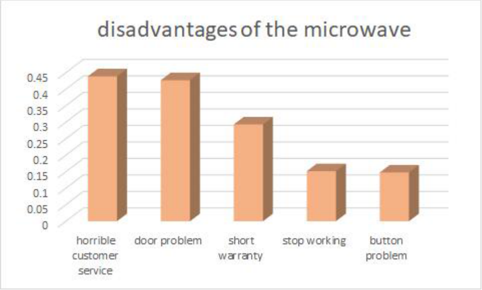

Figure 9: The proportion of reasons people dislike the microwave

About 35% of customers love its size. They said they can put it anywhere in the kitchen and it fits well.21% thinks it's easy to use.15% people think that the price is fair. 6% enjoy its paint. And some other people find the microwave easy to clean.

However, problems still exist. Customers complain that the customer service is horrible. Besides, the door of the microwave has many problems. Furthermore,20% think warranty time is too short. About 9% of people said that their microwave stop working or the button broke;

2.4 Relations between Specific Quality Descriptors and Rating Levels

We choose 8 specific quality descriptors. And their frequency is shown in the table.

Table 1: The number of times the word was mentioned

	love	perfect	nice	cute	disappointed	waste	bad	junk
high rating	8193	2591	2483	2435	165	65	339	20
low rating	187	64	185	189	432	292	260	149

We calculated the probability of these words appearing in high rating reviews and low rating reviews, respectively.

$$p(word\_in\_high\_rating) = \frac{\frac{num\_in\_high}{high\_num}}{\frac{num\_in\_high}{high\_num} + \frac{num\_in\_low}{low\_num}} \tag{1}$$

- $word\_in\_high\_rating$:Probability of words appearing in high scoring comments
- num_in_high :The number of times words appear in high rated comments
- num_in_low :The number of times words appear in low rated comments
- $high\_num$:Number of high rated comments

- low_num :Number of low rated comments

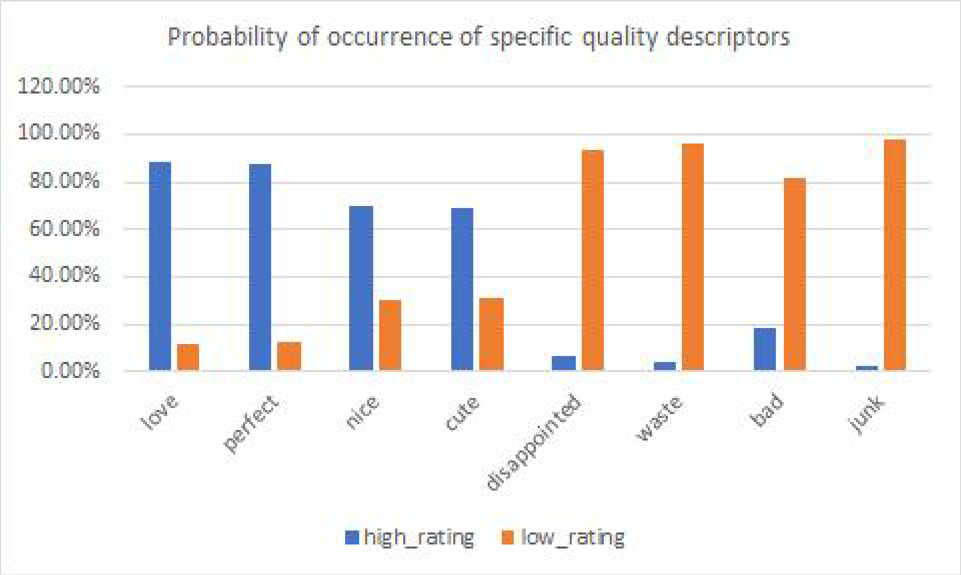

Figure 10: probability the word appear in high/low rating comments

As we can see, if the word 'love' or 'perfect' appears in the review, Users nearly have a 90% chance to give high scores. On the contrary, if the word 'disappointed' is in the review, We can almost conclude that the user will give a low rating.

So it is obvious that specific quality descriptors are strongly associated with rating levels.

2.5 Do specific star ratings incite more reviews?

2.5.1 Theory

To analyze whether some specific star ratings lead to more reviews, we have analyzed the highest selling products for hair dryers, microwave ovens, and pacifiers. First, we counted the number of reviews per month for a certain product and the number of reviews per month for five ratings of the product and then used the correlation coefficient formula to calculate the correlation coefficient between the total number of reviews and a certain number of star ratings. The correlation coefficients corresponding to the scores are compared. When a certain correlation coefficient is large, the more the certain star ratings in a certain period, the more the user's comments will be. That is to say, the number of certain star ratings and the total number of star ratings are in a linear relationship. That way, we have every reason to believe that this rating will incite more reviews.

2.5.2 Data Processing

Here is the table of the correlation coefficient for every star rating:

Table 2: correlation coefficient for every star rating

product	star1	star2	star3	star4	star5
hair dryer	0.68	0.46	0.69	0.85	0.97
microwave	0.58	0.38	0.6	0.57	0.72
pacifier	0.35	0.36	0.5	0.61	0.98

Here is the graph based on the above table:

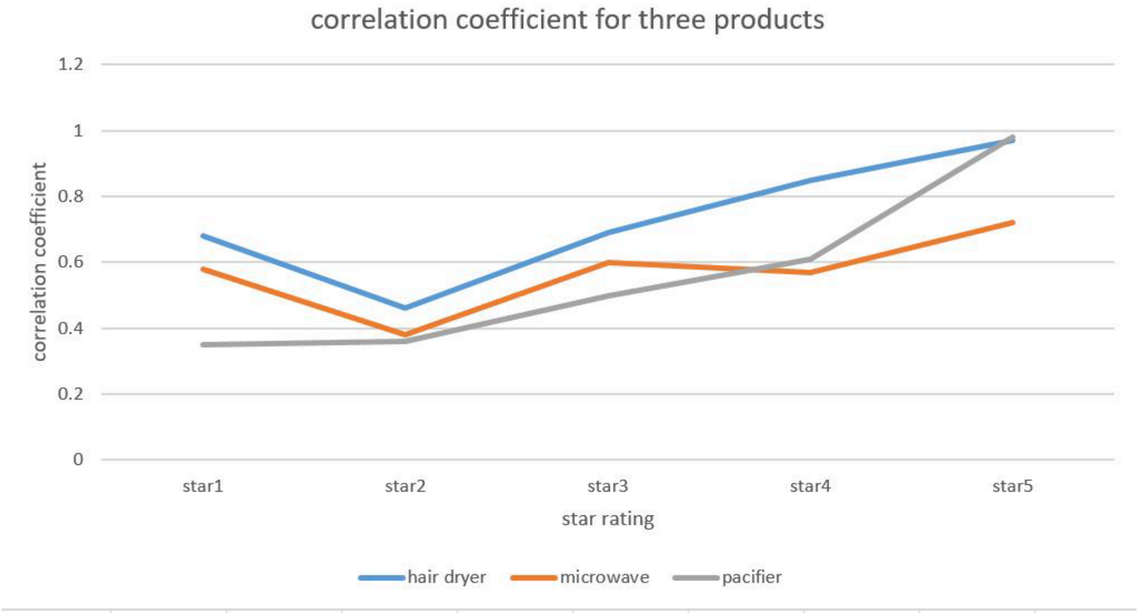

Figure 11: correlation coefficient incite

2.5.3 Conclusion

As can be seen from the above chart, the correlation coefficient with a star of 5 is the largest. This means that a 5-star rating will lead to more reviews.

2.6 Using DNN to Calculate the Value of the Reviews

The emotion analysis system is one of the most classical applications in natural language processing. We use the algorithm developed by Han Xiao and Yao Lu[?] to map the degree of whether a customer likes or dislikes the product to the range of [0,1], to quantitatively reflect the attitude expressed by the comment.

2.6.1 Preprocessing

The original text data often has many parts that affect the final classification effect. This part of data or text needs to be cleaned at the beginning of text classification, otherwise, it will easily lead to the so-called "trash in, trash out" problem[?]

In addition to the steps of missing value processing, de-duplication processing and noise processing included in the data cleaning of general classification problems, we also clean up the following data:

- Non-text data: HTML tags, URL addresses, and other non-text content are often attached to the text, so it is necessary to clear this part of the content that is not helpful for classification.
- Meaningless text: Besides, the rest of the text, such as advertising content, copyright information, and personal signature, should also be filtered out, which should not be learned as features.

2.6.2 Algorithm

In the algorithm,Dynamic Convolution Neural Network is used as the sentiment prediction algorithm.The network layers are set as follows:

- Embedding layer: set as the lowest layer, to get the vector representation for individual words to form the matrix representation of the sentence
- Convolutional Layer: extracts local features by performing 2d convolution on the sentence matrix
- K-max pooling layer: extracts kth strongest signals on a per-feature basis
- Folding layer: adds interaction among features by "folding" the input matrix
- Logistic regression layer: the final layer that makes the prediction by assigning scores to each output label

2.7 Verifying the Accuracy of Review-Based Ratings

In the previous subsection, we analyzed reviews using natural language processing and gave a score for each record. This subsection we will verify the accuracy of the previously given score.

We believe that the correlation coefficient between review-based ratings and star ratings can well reflect the accuracy of the reviews analysis. In order to calculate the correlation coefficient, we need to first calculate the covariance of these two data.Covariance can be easily calculated by the following equation:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Cov (X, Y) represents the covariance of X and Y, and E (X) represents the expectation of X.

After getting the covariance, we used the following formula to calculate the correlation coefficient:

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$$

Here r (X, Y) represents the correlation coefficient between X and Y, and Var [X] represents the variance of X.

We calculated the total correlation coefficients of the three categories of products, and the correlation coefficients of high star rating (4-5) and low star rating (1-2). The results are as follows:

Table 3: correlation coefficient between review-based ratings and star ratings

		correlation coefficient
hair dryer	high star rating	0.4201
	low star rating	0.4934
	all	0.5964
microwave oven	high star rating	0.3978
	low star rating	0.5381
	all	0.6753
baby pacifier	high star rating	0.4372
	low star rating	0.4557
	all	0.5797

As can be seen from the table, review-based ratings are positively related to star rating satisfaction. In the case where the review-based score is relatively low, especially when the star rating is in the middle level, the correlation coefficient is relatively high. Overall our review-based ratings are acceptable.

2.8 User Evaluation

2.8.1 Theory

To get a user evaluation of the top three hair dryers, microwaves, and pacifiers, we came up with the following measurement method. To make full use of reviews and ratings, we combined text-based ratings and star-based ratings, each giving different weights, and finally got a most informative measure. The formula is as follows:

$$average\ rating = 0.8 \times star\_rating + text\_rating$$

2.8.2 Data Processing

First of all, let's see the chart of the top three hair dryers (product_id: B003V264WW, B0009XH6TG, B00132ZG3U):

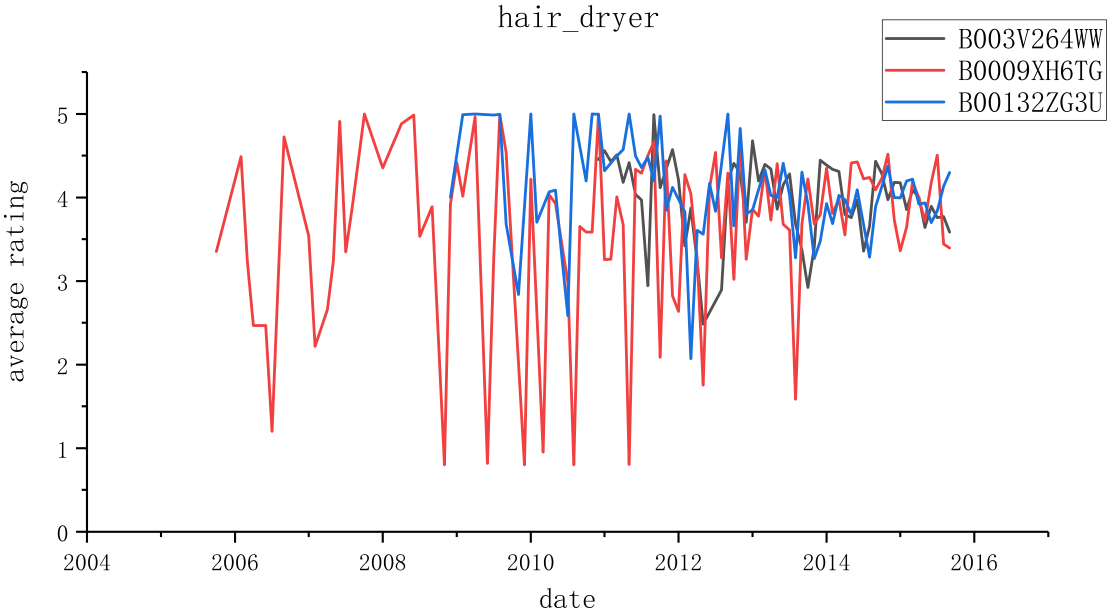

Figure 12: hair dryer average rating

Combining the above figure, we find that the average score curves of these three products are very similar. In the early stages, product ratings fluctuated greatly because there were fewer customers. For example, the product B0009XH6TG (red line) has fluctuated between 1 and 5 stars in the early stage, the product B003V264WW (blue line) has fluctuated between 3 stars and 5 stars in the early stage, and the product B00132ZG3U (green line) has also fluctuated between 3 stars and 5 stars. In the later period, the scores of the three products stabilized at about 4 points, and the fluctuations did not exceed 0.5 points. It can be seen that the ratings of these three products in the later period are similar.

Secondly, here is the chart of the top three microwave ovens (product_id:B0055UBB40, B0058CLNBU,B0052G14E8):

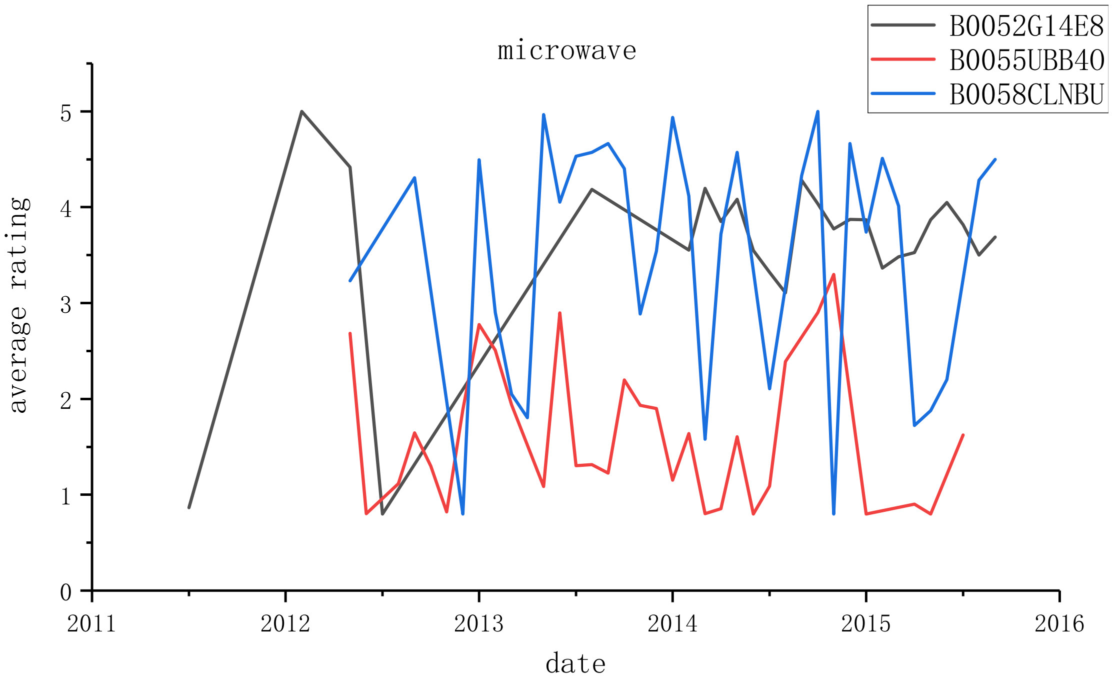

Figure 13: microwave average rating

This picture is not different from the above picture, because the microwave oven has fewer reviews, the average rating fluctuates greatly. First of all, the product B0055UBB40 (red line) has been fluctuating between one star and three stars. Secondly, the product B0058CLNBU (blue line) has been fluctuating between one star and five stars, and the blue line has been almost above the red line. The final product B0052G14E8 (green line) fluctuated between one star and five stars in the early stage, and stabilized at about four stars in the later stage. It can be seen that the market feedback of product B0058CLNBU (blue line) and product B0052G14E8 (green line) is better than product B0055UBB40 (red line).

Finally, the chart of the top three pacifiers is as follow: (product_id:B003CK3LLDI, B0028IDXDS, B0045I6IA4):

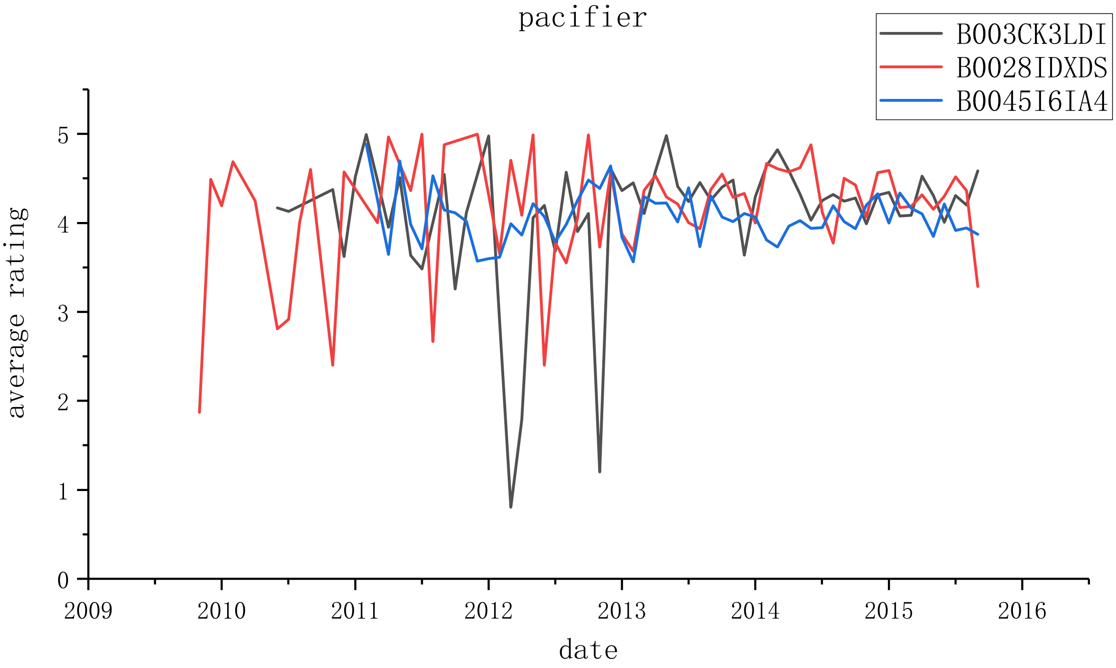

Figure 14: pacifier average rating

This picture is very similar to the picture about the hair dryer, both of which fluctuate in the early stage and stabilize in the later stage. In the later period, the three products were stable between four stars and 4.5 stars. In this regard, market feedback is better than hair dryers.

3 Conclusion

4 Letter

References

[1] http://xiaohan2012.github.io/twitter-sent-dnn/

[2] Hu M , Liu B . Mining and summarizing customer reviews[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004. ACM, 2004.

Appendices

Appendix A First appendix

some more text **Input python source:**

```
import collections
import re
def stopwordslist():
    #delete stop words
    stopwords = [line.strip() for line in open(’./stopwords.txt’).readlines()]
    return stopwords

def count_word(filename):
    """count filename word frequency
    :param filename:
    """
    # return value is dict,will count automatically
    word_counter=collections.Counter()

    with open(filename,’r’,encoding=’utf-8’) as f:
        for line in f:
            #Convert to lowercase
            line=line.lower()
            word_counter.update([word for word in re.split(’\s+’,line) if word !=’’])
    return dict(word_counter)

def get_top(filename,topk=10):
    """
    :param filename:the file you want to count word frequency
    :param topk:you want to return the first topk words with the highest frequency
    """
    word_dict=count_word(filename)
    topk_words=sorted(word_dict.items(),key=lambda x:x[1],reverse=True)
    return topk_words[:topk]


if ’__main__’==__name__:
    import sys
    stopwords = stopwordslist()
    fout=open(’output.txt’,’w’)
    if len(sys.argv)!=3:
        print(’Usage: {} filename topk’.format(sys.argv[0]),file=sys.stderr)
        sys.exit(0)
    top_words=get_top(sys.argv[1],int(sys.argv[2]))
```