Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2308823

Comprehensive Exploration of Amazon Reviews and Ratings

Summary

This paper conducts data analysis and information mining from the following four aspects: review, the relationship between star rating and usefulness rating, product brand rating, product prediction reputation and the impact of star rating on the review, so as to put forward reliable sales strategies and product improvement suggestions.

During data preprocessing, we removed the fields irrelevant to the data (column processing) and deleted the evaluation irrelevant to the commodity (row processing) in the data set. We also extract keywords from the comment title and comment content, and make some word clouds. We can find that the most concerned aspects of users are price, quality, volume, safety, brand, appearance, etc., so that we can put forward some suggestions to businesses.

After preprocessing, we use Python's NLTK tool to analyze the text data and quantify it into emotional scores. We have adopted the methods of data visualization, descriptive statistics and correlation analysis to establish a reasonable mathematical model, which can comprehensively evaluate the stars, the number of likes and the content of comments from multiple perspectives.

Next, in order to continue to refine the scoring model, we studied the impact of the number of comments on the number of likes. The results show that most comments with more words will receive a lot of likes, which shows that comments with more words are more reliable. We will also consider whether to include Amazon reviewers, and believe that the comments made by Amazon reviewers are more reliable. We finally integrated all the subjective parts of the comment into one indicator, which reflects the overall score of the comment for the product.

Further, we analyzed the word-of-mouth of each product, described the seasonal characteristics of the three products, and predicted the future word-of-mouth trend of the products, based on the review date, which can provide suggestions for the listing time of Sungrow. We also analyzed the change of stars over time, and found that high stars can bring higher stars, and users will indeed be affected by the change of ratings.

Finally, we analyzed the relationship between the emotional score of comments and the star rating, and clearly found that there is a strong correlation between specific keywords and ratings. Specifically, comments containing positive words will lead to higher stars, while comments containing negative words will lead to lower stars.

We also performed sensitivity analysis on the model to prove the rationality of the model design. According to our research on the data, we finally provided some suggestions on product production and sales strategy for Sunshine.

Keywords: NLTK; Emotional analysis; Sales Strategy Formation

Contents

1 Introduction

1.1 Problem Background

With the development of economy, the progress of computer and electronic technology and the widespread use of the Internet, the e-commerce industry has developed rapidly. People's consumption mode and shopping mode are changing, and people are more and more like shopping online. At present, China's online shopping market is growing rapidly. As one of the world's well-known e-commerce enterprises, Amazon has rich experience, strong logistics system and very professional and advanced technology in the field of e-commerce.

Amazon is committed to becoming the most "customer-centric" company in the world. It has become the online retailer with the largest variety of goods in the world. Amazon and other sellers provide millions of unique new, refurbished and second-hand goods, including books, movies, music and games, digital downloads, electronics and computers, household and gardening products, toys, baby products, groceries, clothing, footwear, jewelry, health and beauty products, sports, outdoor products, tools, and automotive and industrial products.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to develop a model to identify key patterns, relationships, measures, and parameters in past customer supplied ratings and reviews associated with other competing products:

- Problem 1： By analyzing the three product data sets provided in the title, and analyzing the parameter relationship between star rating, comment and help rating, to help Sunshine Company succeed in its three new online market products.
- Problem 2： Identify data measures based on ratings and reviews that are most informative for Sunshine Company to track, once their three products are placed on sale in the online marketplace.
- Problem 3：Identify and discuss time-based measures and patterns within each data set that might suggest that a product’s reputation is increasing or decreasing in the online marketplace.
- Problem 4： Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product.
- Problem 5： Analyze the relationship between specific star ratings and customer comments. After seeing a series of low star ratings, are customers more likely to write some type of comments?
- Problem 6： Analyze the relationship between specific quality descriptors of text-based comments, such as "enthusiasm", "disappointment", and rating level.

1.3 Our Work

Our work mainly includes the following parts:

Firstly through the processing of product data sets, including removal of useless data, removal of invalid evaluation, data set transformation, comment merging, etc. This paper analyzes the relationship between star rating and evaluation of different products, and uses

python's NLTK package to segment, remove punctuation, remove stop words, extract stem, restore part of speech, and finally conduct emotional analysis.

Next we evaluate each comment with 1-5 points, 1 point indicates that the comment is completely negative, and 5 points indicates that the comment is completely positive. Analyzed the proportion of comments of each star of each category of products

Finally The comprehensive score of the product is obtained by modeling the star rating and comment emotion score, the number of evaluations obtained by products every day is counted and visualized, and the evaluation of different stars every day is also refined and visualized.

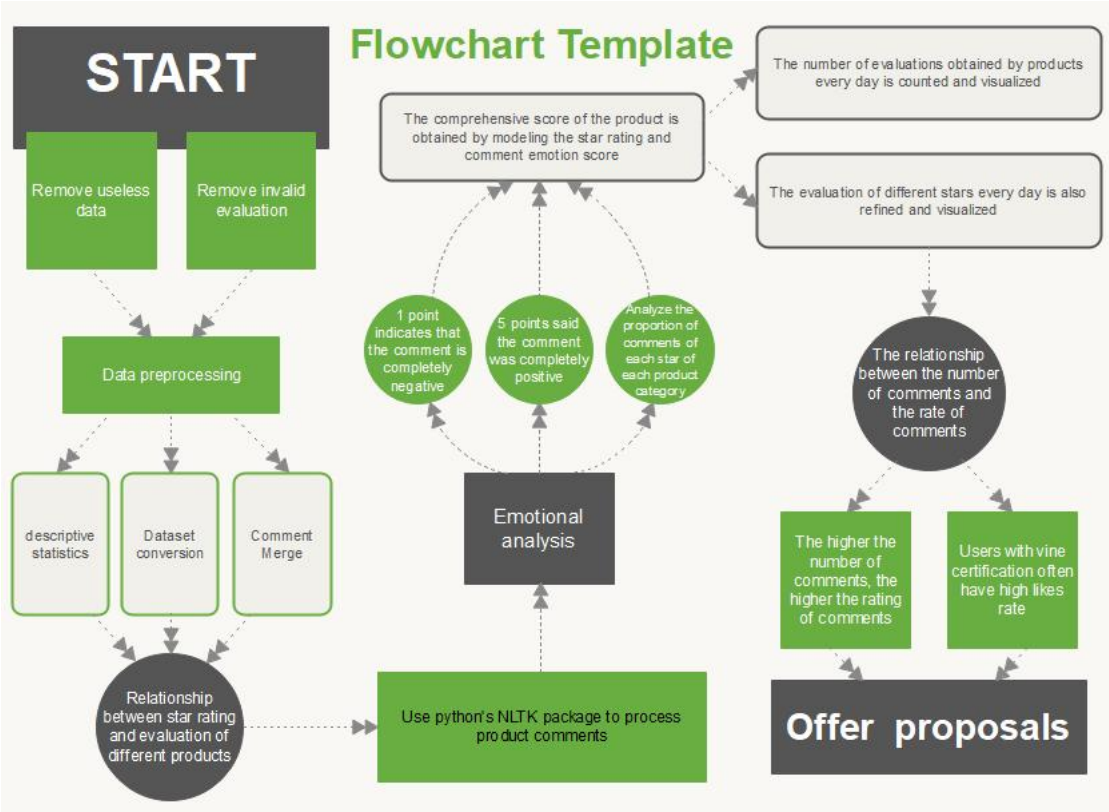

Figure 1: Flow chart

2 Assumptions and Justifications

Considering that practical problems always contain many complex factors, first of all, we

need to make reasonable assumptions to simplify the model, and each hypothesis is closely followed by its corresponding explanation:

- **Assumption 1: All data are assumed to be true.**

We assume that all data are real data from Amazon website, including product information, evaluation content, review date, etc.

- **Assumption 2: We assume that the data is continuous in time.**

We believe that all data have continuity in time, that is, there will be no lack of evaluation for a period of time, which is very important for our future model.

- **Assumption 3: We assume that users pay more attention to some evaluations.**

We assume that users pay more attention to the evaluation, positive or negative evaluation with high or low stars, ignore the neutral evaluation, and mainly focus on the evaluation with many likes.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description	Unit
r	Comprehensive evaluation, that is, the integration of various evaluation data	score
s	A star rating	star rating
d	Emotional score of a text comment	score
c	Credibility of a comment	-
l	Rate of comments	-
n	Number of likes to comment	-
N	Sum of likes of all comments	-
m	Number of words in a comment	-

4 Data Processing and Analysis

4.1 Data cleaning

4.1.1 Remove useless data

Before we build the model, we first remove useless data, the main data including the follow two parts:

- Commodity independent data

After preprocessing the data, we found that there are no keywords such as microwave in some product titles, and we think these products are not in the scope of discussion; Therefore, we process the data as follows: first, remove the fields irrelevant to the data analysis, such as commodity category, commodity category, etc; Secondly, the evaluation irrelevant to the product in the data set and the removal criteria are whether the product name we analyzed (such as microwave) is included in the product name; At the same time, remove the comments of people who do not have the vine label and who do not buy; There is blank data in the information provided to us by the database, and we will also delete this part of the information.

product_title
arksen portable mini washing machine 8 - 9lbs dorm camping rv compact laundry washer
ge mwf smartwater compatible water filter cartridge - refrigerator
samsung basket adjuster - dd97-00119b
koolatron coca-cola indoor/outdoor party fridge
dishwaser stainless steel film 26 x 3'
dr2009wglp 20" freestanding gas range with 2.62 cu. ft. manual clean oven 4 sealed burners electronic ignition and broiler door in
ge jgb870sefss 30 stainless steel gas sealed burner double oven range - convection
Danby DWC283BLS 3.5-Cu.Ft. 30-Bottle Free-Standing Wine Cooler, Black/Stainless
Broan 639 Wall Cap for 3-1/4 x 10 Duct for Range Hoods and Bath Ventilation Fans
IMAGE声 Mini USB-Powered Fridge Cooler for Beverage Drink Cans in Cubicle and Home office (Black)
Panda Small Mini Counter Top Portable Compact Washer Washing Machine 5.5-10lbs
Broan 412402 ADA Capable Non-Ducted Under-Cabinet Range Hood

Figure 2: Display of some useless data

- User-independent data

A large part of the data are not comments of Amazon reviewers and users who have not purchased goods. We name them as data irrelevant to users. We assume that the confidence level of these users' comments is 0, and convert 'N', 'Y' and other characters in the data set into logical numbers of '0', '1', and combine the title and body of the comment into a comment。

4.1.2 Descriptive statistics

Since the review texts are rich and complex, it is of great importance to dig out the topic words so as to clearly grasp what consumers really care about. And here we accomplish this task by applying NLP method for topic word extraction.

We first made descriptive statistics on the overall evaluation data of hair dryers, and found that helpful_ Votes, that is, the number of helpful votes is more, the average number of stars commented is 4.191, and the median is 5.

Table 2: Hair dryer data description statistics

	star_rating	helpful_votes	total_votes	vine	verified_purchase
count	9740	9740	9740	9740	9740
mean	4. 191	1. 591	1. 900	0. 018	0. 982
std	1. 233	12. 454	13. 501	0. 134	0. 134
min	1	0	0	0	0
25%	4	0	0	0	1
50%	5	0	0	0	1
75%	5	1	1	0	1
max	5	499	575	1	1

Using the NLTK package of python, we have segmented the comments of the goods, removed punctuation, removed stop words, extracted stems, restored parts of speech, and finally carried out emotional analysis. We have given each comment a score of 1-5 points, with 1 point indicating that the comment is completely negative, and 5 points indicating that the comment is completely positive. The proportion of comments of each star of each category of products is analyzed. From Figure 2, we can see that the proportion of comments with 4 points is relatively large, indicating that customers give more positive comments

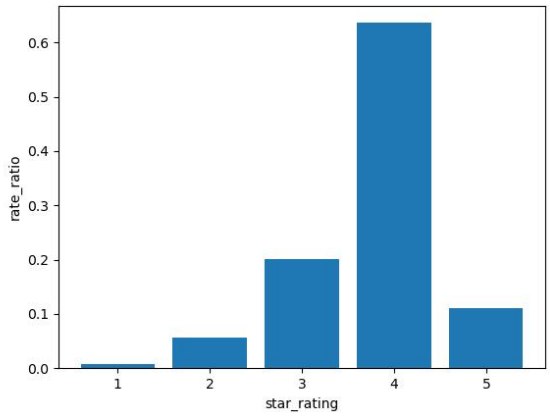

Figure 3: Overall evaluation proportion of commodities

5 Commodity evaluation model

By analyzing the three product data sets provided in the title, and analyzing the parameter relationship between star rating, comment and help rating, to help Sunshine Company succeed in its three new online market products. We try to build a model that can input a group of evaluation data and output the comprehensive evaluation, confidence and commodity keywords of the evaluation.

5.1 Establishment of model

① Comprehensive evaluation：

The comprehensive evaluation r is composed of the star rating s of the comment and the emotional score d of the comment content:

$$r = 0.3s + 0.7d \tag{1}$$

② Comment Confidence:

The comment confidence c is composed of the comment's likes rate l and the number of likes n. When the number of likes exceeds 10 and the rate of likes is less than 70%, we think the confidence level of this comment is 0. The specific functions are:

$$c = \begin{cases} 0, \ l < 0.7 \ and \ n \geq 10, \\ l \times (n+1), otherconclution \end{cases} \tag{2}$$

③ Product keywords:

The effective data is analyzed by word frequency, and 10 positive and negative keywords of each product are given, that is, the 10 product attributes that users most care about. Each comment is then keyword extracted, and the 10 keywords are matched to give the comment keywords.

5.2 Result analysis

First , after processing the data, because the comments also have an impact on the evaluation of the goods, we conduct emotional analysis on the comments of the goods and get positive, negative and neutral results. Here we use the NLTK toolkit in python for natural language processing, quantify the emotional degree, give the emotional score, and then

calculate the comprehensive evaluation and round it.

Below we show the results of some data processing. These are the proportion data of one row of each commodity.

Table 3: One-star ratio of each product

	product_parent	total_rates_count	star_rating	stars_rate_count	rate_ratio
52	127343313	236	1	1	0.004237288
413	868768702	114	1	1	0.00877193
359	748065701	92	1	1	0.010869565
338	694290590	360	1	1	0.002777778
333	685652978	159	1	1	0.006289308

Then, we analyze the changes of some main features over time: We visualize the review amount of three products each month and observe that the reviews amount has increased exponentially over time. Besides that, the averaged rating score per month is further calculated and displayed. The following figure shows the daily evaluation quantity of three products.

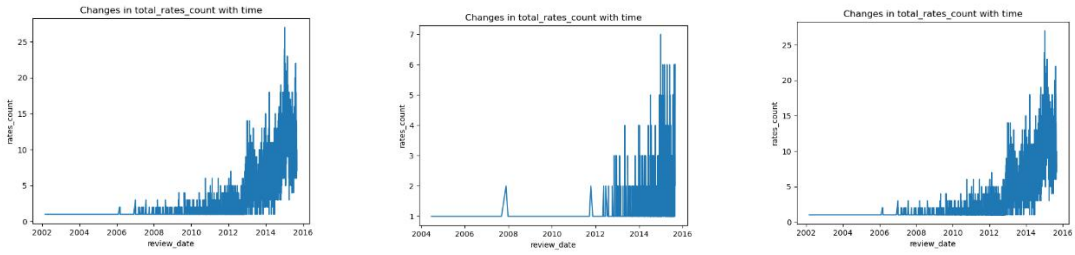

Figure 4: Daily evaluation quantity of three products

According to the statistics of the number of evaluations of each star over time, it can be seen that with the change of time, the overall trend of comments is on the rise, with more comments scoring 4 and 5 points.

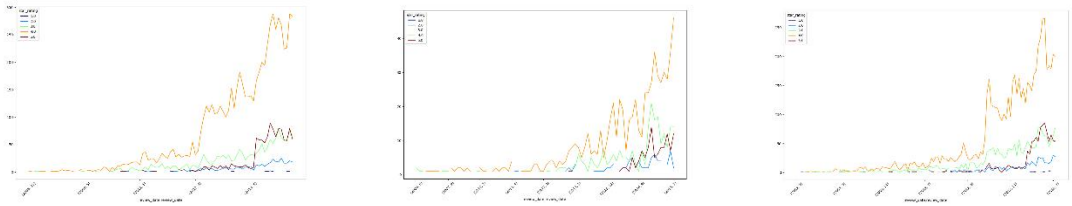

Figure 5: The number of evaluations of each star changes over time

We use python's NLTK package to segment words, remove punctuation, remove stop words, extract stems, restore parts of speech, and finally conduct emotional analysis and get the following word cloud.

Figure 6: Positive and negative word clouds of hair dryer

6 Relationship between ratings and comments

6.1 Ratings and Reviews Based Data Measures

① Research on the number of comments

By analyzing the relationship between the number of comments and the number of likes, we can see that the number of words is more than the number of likes. We take hair dryer as an example.

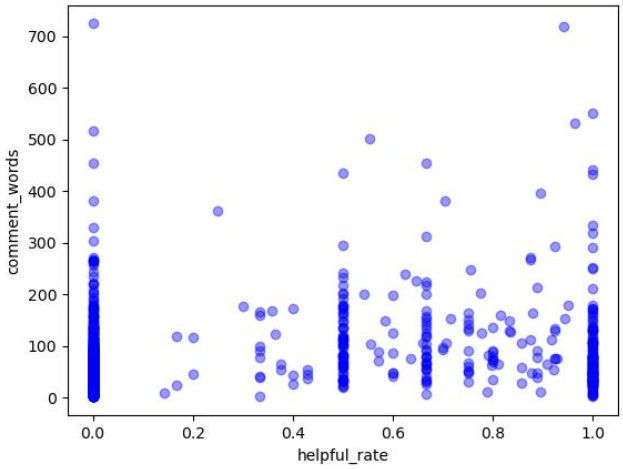

Figure 7:Relationship between comments and likes

② Vine's comments are a little more persuasive

Draw a scatter chart of the number of likes obtained by vine. The number of likes is the abscissa and the number of people is the ordinate. vine is one color and non-vine is another color. Observe the distribution. It can be seen that Vine's comments are a little more restrictive

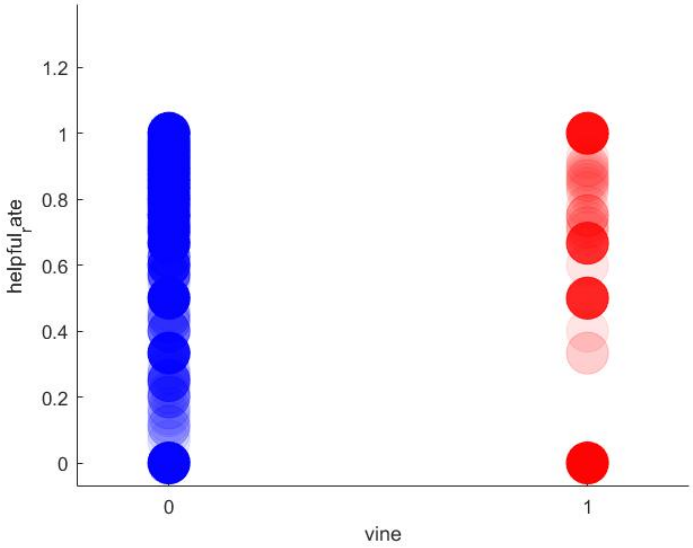

Figure 8:Relationship between vine and helpful_rate

Normalize the confidence, N is the sum of all likes, and v is the 0-1 variable of whether the user is vine:

$$v=\begin{cases}1, \text{ The user is vine,}\\ 0, \text{ The user is not vine}\end{cases} \tag{1}$$

$$c=\begin{cases}0,\ l<0.7\ and\ n\geq 10, \\ (l+0.1v)\dfrac{n}{N},\ \text{other conclution}\end{cases} \tag{2}$$

The average number of comments is $m_{avg}=52.046$, the maximum number of comments is $m_{\max}=1579$, and the number of comments m and emotional score d constitute 70% of the comprehensive evaluation:

$$r=0.3s+0.7(d+\frac{m-m_{avg}}{m_{\max}}) \tag{3}$$

Confidence is actually a coefficient of comprehensive evaluation, giving the function of comprehensive evaluation:

$$r=\left[0.3s+0.7(d+\frac{m-m_{avg}}{m_{\max}})\right]c \tag{4}$$

6.2 Reputation Metric

By analyzing vine and helpful_ We can see that the reputation of products in the online market fluctuates with time. The reputation of baby pacifiers gradually gets better with time, while the reputation of microwave ovens tends to get worse with time. The reputation of hair dryers changes from good to bad with time.

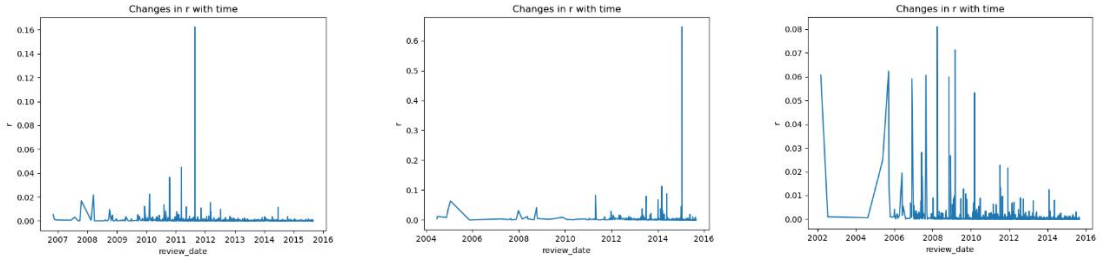

Figure 9:Changes in r with time

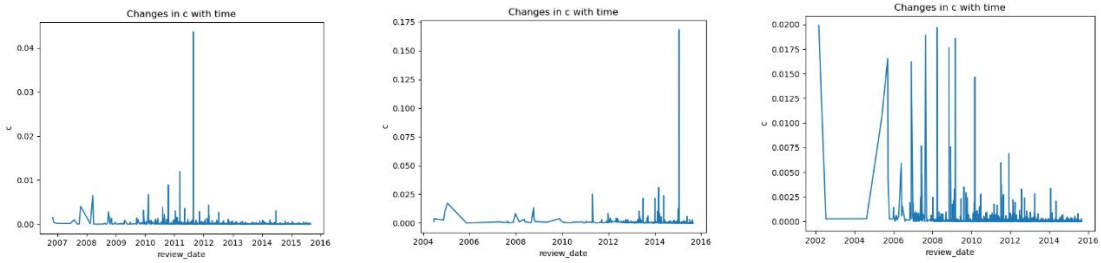

Figure 10: Changes in c with time

6.3 Correlation between Affective Words and Star Ratings

Since different people have different rating standards, their mappings from reviews to ratings are various. Therefore, chances are that: for some people, specific quality descriptors of their text-based reviews such as enthusiastic and disappointed are strongly associated with rating levels, for others the relations are relatively looser to different extents. As a result, analyzing the alignment of rates and review texts is of great importance.

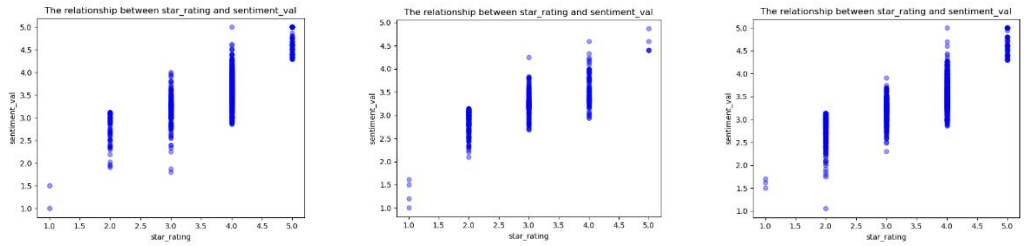

Figure 11:Relationship between star_rating and sentiment_val

From the figures above, it is easily seen that there is a great correlation between star_rating and sentiment_val. Extreme reviews with more neutral words tend to consist of more property words and have longer lengths. With further observation of detailed examples, we find that this is largely because some reviewers are objective enough to give factual descriptions about products instead of piling up their strongly affective attitude towards them. For these customers, the extents of their affective words in reviews are not closely related to ratings.

7 Sensitivity Analysis

1. Comments from users who are neither Amazon commentator nor purchase goods：

By comparing with normal comments, we find that comments are very different and words with high frequency are strange, which can prove that even if a small number of users leave comments on Amazon after buying on other platforms, most of these users' comments have low credibility and can be removed from the data.

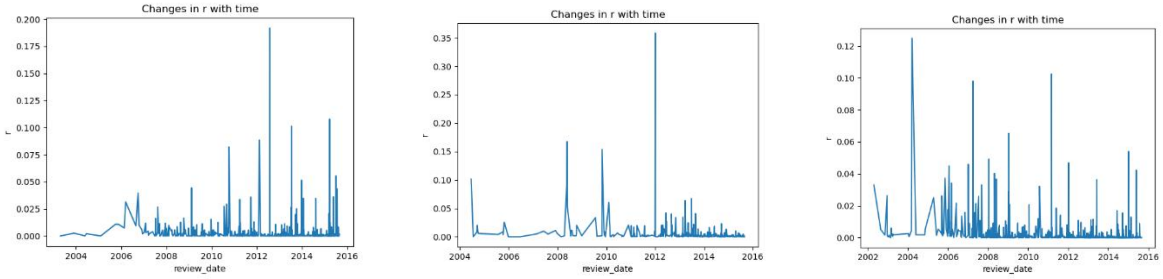

Figure 12: Changes in r with time

By analyzing the change chart of confidence over time, we can see that the reputation of the product in the online market shows a fluctuating state with the increase of time. The reputation of the baby pacifier gradually gets better with the increase of time, while the reputation of the microwave oven has a trend of getting worse with the increase of time. The reputation of the hair dryer changes from good to bad with the change of time.

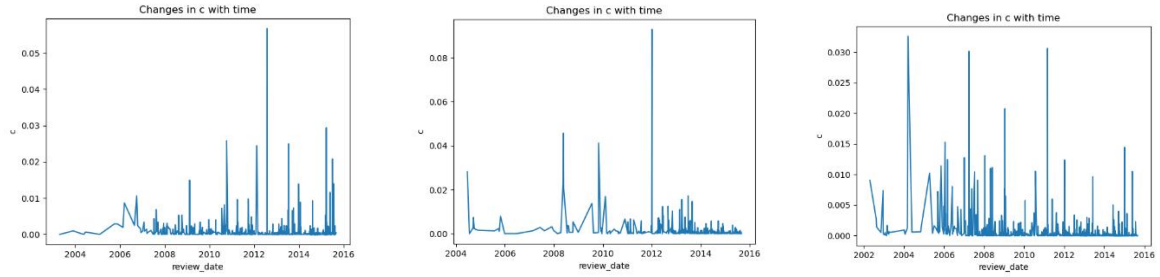

Figure 13: Changes in c with time

2. Verify the correctness of the model：

Manually added 1000 real data obtained from Amazon, and predicted the user's score by using the model by entering the comment content and the number of likes. It can been seen that the accuracy can be up to 50%

Data / Star_rating	5	4	3	2	1
forecast	132	621	103	47	35
Actual	120	669	146	46	19
Accuracy	90.9%	92.8%	70.5%	97.9%	54.3%

8 Model Evaluation and Further Discussion

8.1 Strengths

- Full utilization of information: When designing our measures, not only do we combine

star ratings and review contents, but also we consider other useful attributes in the provided data;

- Great model performance: The high correlation between trends of our quantified reputation and actual sales suggests our models have great performance and that our measures are accurate;
- Good choice of methodology: We reveal the strong connections between positive words and high ratings, negative and low ratings, thanks to our choice of methodology;

8.2 Weaknesses

- We simply assume the sales volumn to be proportional to number of reviews, which may be too simplistic to be realistic
- We do not take marketing strategies of Amazon like sales promotion into consideration when analyzing specific ratings and descriptors.

9 Conclusion

Through the processing of product data sets, including removal of useless data, removal of invalid evaluation, data set transformation, comment merging, etc. This paper analyzes the relationship between star rating and evaluation of different products, and uses python's NLTK package to segment, remove punctuation, remove stop words, extract stem, restore part of speech, and finally conduct emotional analysis. Each comment is evaluated with 1-5 points, 1 point indicates that the comment is completely negative, and 5 points indicates that the comment is completely positive. The proportion of comments of each star of each category of products is analyzed, and the comprehensive score of the product is obtained by modeling the star rating and comment emotion score. The number of comments received by the product every day is calculated and visualized. At the same time, the evaluation of different stars every day is also refined and visualized. We also analyzed the relationship between the number of comments and the rate of liked comments, and found that the more comments, the higher the rate of liked comments. At the same time, we found that users with vine authentication often have higher rate of liked comments.

We extracted keywords from positive comments and negative comments respectively, got their own keywords, and made them into word clouds. Through the analysis of keywords, we found that products with favorable comments often have advantages in terms of price and quality, while products with poor comments often have disadvantages in terms of appearance, use, etc. Manufacturers can also make relevant improvements according to the word clouds of their different products.

10 The Letter to the Marketing Director of Sunshine Company

Dear director,

Considering today's increasingly fierce competition in e-commerce, accurately grasping customer needs and specifying appropriate marketing strategies are of vital importance to improve corporate profits and product visibility. As response to your company's requirement, we are here pretty glad to have the opportunity to introduce our research and suggestions to you, with the hope that it may give you some insights of the future strategies.

First of all, we analyzed the relationship between the star rating and the evaluation of different products, and used python's NLTK package to segment, remove punctuation, remove stop words, extract stem, restore part of speech, and finally carried out an emotional analysis, giving a score of 1-5 points for each comment, A score of 1 indicates that the comment is completely negative, and a score of 5 indicates that the comment is completely positive. The proportion of comments of each star in each category of products was analyzed, The comprehensive score of the product is obtained by modeling the star rating and comment emotion score, The number of evaluations obtained by products every day is counted and visualized, and the evaluation of different stars every day is also refined and visualized.

We also analyzed the relationship between the number of comments and the rate of liked comments, and found that the more comments, the higher the rate of liked comments. At the same time, we found that users with vine authentication often have higher rate of liked comments.

According to our analysis results, we formulate reasonable sales strategies for your company: ①We recommend that you put microwaves and hair dryers into the market when the reputation rises. While the reputation of baby pacifiers is on the rise, and it's best to put them on the market now. ② The more complete the product information is, the less loss will be caused by the unequal information between buyers and sellers. ③ We recommend that you increase your promotional efforts when there are more five-star ratings of your products to form positive feedback. ④ When your product's reputation declines, focus on one-star ratings and reviews.

Thanks for taking the time out of your busy schedule to read my letter. Hope our advice can help.

MCM Team # 2308823

References

[1] Amazon Highlights Brand Protection Report[J]. Manufacturing Close - Up,2020.
[2] Amazon-related phishing sites approach 900 on Amazon Prime Day[J]. M2 Presswire,2020.
[3] Proteomics; New Proteomics Study Findings Reported from Shanghai Jiao-Tong University (Structure-based design and confirmation of peptide ligands for neuronal polo-like kinase to promote neuroregeneration)[J]. Chemicals & Chemistry,2016.
[4] He-Li Cao,Hao Chen,Yu-Hui Cui,Heng-Li Tian,Jiong Chen. Structure-based design and confirmation of peptide ligands for neuronal polo-like kinase to promote neuroregeneration[J]. Computational Biology and Chemistry,2016,61.
[5] YADAVILLI VRPS SASTRY,SESHADRI KARTHICK. Explainable sentiment analysis for product reviews using causal graph embeddings[J]. Sādhanā,2022,47(4).
[6] Almuayqil Saleh Naif,Humayun Mamoona,Jhanjhi N. Z.,Almufareh Maram Fahaad,Khan Navid Ali. Enhancing Sentiment Analysis via Random Majority Under-Sampling with Reduced Time Complexity for Classifying Tweet Reviews[J]. Electronics,2022,11(21).
[7] Almalis Ioannis,Kouloumpris Eleftherios,Vlahavas Ioannis. Sector-level sentiment analysis with deep learning[J]. Knowledge-Based Systems,2022,258.

Appendices

Appendix 1

Introduce: Emotional analysis

```
import numpy as np
import pandas as pd
import nltk
import matplotlib.pyplot as plt
from textblob import TextBlob
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer


# 情感分类器
def classifier(x):
    testimonial = TextBlob(x)
    testimonial.sentiment
    a = testimonial.sentiment.polarity
    if a < -0.1:
        return 'neg'
    elif a > 0.5:
        return 'pos'
    else:
        return 'neu'


def classifier_val(x):
    testimonial = TextBlob(x)
    testimonial.sentiment
    a = testimonial.sentiment.polarity
    return a * 2 + 3


def calculate_word(text):
    return len(text.split())


# 读入数据
# TODO(ppy): 在下面更改产品
product = 'pacifier'
input_file = 'E:\\njupt\\Problem_C_Data\\' + product + '.csv'
```

```
print(input_file)
data = pd.read_csv(input_file, encoding='utf-8')
data.head()

# 将标题和正文连接
data['review_headline_body'] = data['review_headline'] + ' ' + data['review_body']

# 删除无关字段
data = data.drop(columns=['marketplace', 'customer_id', 'review_id', 'product_id', 'product_category'])
data['product_title'] = data['product_title'].str.lower()

# 将是否购买转化为 0, 1
data.replace('n', '0', inplace=True)
data.replace('N', '0', inplace=True)
data.replace('y', '1', inplace=True)
data.replace('Y', '1', inplace=True)

# 将字符串转化为 float
data[['star_rating', 'helpful_votes', 'total_votes', 'vine', 'verified_purchase']] \
    = data[['star_rating', 'helpful_votes', 'total_votes', 'vine', 'verified_purchase']].astype('float')
data['product_parent'] = data['product_parent'].astype('object')

# 删除无用评价
data = data[~((data['vine'] == 0) & (data['verified_purchase'] == 0))]
# TODO(ppy): 如果产品为吹风机的话使用下面的一行
data = data[data['product_title'].str.contains(product)]
# data = data[data['product_title'].str.contains('hair dryer')]

# 转化时间
data.loc[:, 'review_date'] = pd.to_datetime(data.loc[:, 'review_date'], format='%m/%d/%Y', errors='coerce')

# 删除缺省值
data = data.dropna(subset=data.columns, how='any')

# 统计评价的单词数量
data['words'] = data.apply(lambda x: calculate_word(x['review_headline_body']), axis=1)

# 分析情感
data['sentiment'] = data.apply(lambda x: classifier(x['review_headline_body']), axis=1)
```

```
data['sentiment_val'] = data.apply(lambda x: classifier_val(x['review_headline_body']), axis=1)

# 结合情感和评分
data.loc[:, 'star_rating'] = round(data.loc[:, 'star_rating'] * 0.3 + data.loc[:, 'sentiment_val'] * 0.7)
# data.loc[(data['star_rating'] == 3.0) & (data['sentiment'] == 'pos'), 'star_rating'] = 4.0


# 计算评价比例
df1 = data.groupby(['product_parent'], sort=True).size().reset_index(name='total_rates_count')
df2 = data.groupby(['product_parent', 'star_rating'], sort=True).size().reset_index(name='stars_rate_count')
merge12 = pd.merge(df1, df2, on='product_parent', how='outer')

# 计算每种商品每一级评价比例
merge12['rate_ratio'] = merge12['stars_rate_count'] / merge12['total_rates_count']
# 计算各星比例
most_rated_5 = merge12.loc[merge12.loc[:, 'star_rating'] == 5, :].sort_values(by='stars_rate_count')
most_rated_5.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\每种商品的五星比例' + product + '.csv')
most_rated_4 = merge12.loc[merge12.loc[:, 'star_rating'] == 4, :].sort_values(by='stars_rate_count')
most_rated_4.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\每种商品的四星比例' + product + '.csv')
most_rated_3 = merge12.loc[merge12.loc[:, 'star_rating'] == 3, :].sort_values(by='stars_rate_count')
most_rated_3.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\每种商品的三星比例' + product + '.csv')
most_rated_2 = merge12.loc[merge12.loc[:, 'star_rating'] == 2, :].sort_values(by='stars_rate_count')
most_rated_2.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\每种商品的二星比例' + product + '.csv')
most_rated_1 = merge12.loc[merge12.loc[:, 'star_rating'] == 1, :].sort_values(by='stars_rate_count')
most_rated_1.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\每种商品的一星比例' + product + '.csv')

temp_x = merge12.loc[merge12['product_parent'] == most_rated_5.iloc[-1]['product_parent']]['star_rating']
```

```
rate = [most_rated_1['stars_rate_count'].sum() / most_rated_1['total_rates_count'].sum(),
        most_rated_2['stars_rate_count'].sum() / most_rated_2['total_rates_count'].sum(),
        most_rated_3['stars_rate_count'].sum() / most_rated_3['total_rates_count'].sum(),
        most_rated_4['stars_rate_count'].sum() / most_rated_4['total_rates_count'].sum(),
        most_rated_5['stars_rate_count'].sum() / most_rated_5['total_rates_count'].sum()]
plt.bar(temp_x, rate)
plt.xlabel('star_rating')
plt.ylabel('rate_ratio')
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\查看商品整体的评价比例' + product + '.png')
plt.show()

# 统计每天的评价数量
df3 = data.groupby(['review_date'], sort=True)['star_rating'].size().reset_index(name='rates_count')

df3.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\统计每天的评价数量' + product + '.csv', header=None)
x = df3.loc[:, 'review_date']
y = df3.loc[:, 'rates_count']

plt.plot(x, y)
plt.xlabel('review_date')
plt.ylabel('rates_count')
plt.title('Changes in total_rates_count with time')
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\统计商品的总评价数随时间的变化' + product + '.png')
plt.show()

# 统计每个分数的评价数量随时间的变化
plt.figure()
reviews_grp = data.groupby([data['review_date'].dt.year, data['review_date'].dt.month, data['star_rating']])[
    'product_parent'].agg('count').unstack()

bar = reviews_grp.plot(figsize=(15, 9), rot=45, colormap='jet')
fig1 = bar.get_figure()
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\统计商品每个分数的评价数量
```

```
随时间的变化' + product + '.png')
plt.show()

# 统计点赞率
data['helpful_votes'].fillna(0)
data['total_votes'].fillna(0)
data['helpful_rate'] = data['helpful_votes'] / data['total_votes']
data = data.fillna(0)

# %%计算置信度
# data.loc[(data['star_rating'] == 3.0) & (data['sentiment'] == 'pos'), 'star_rating'] = 4.0
data['c'] = (data['helpful_rate'] + 0.1 * data['vine']) * data['helpful_votes'] / data['helpful_votes'].sum()
data.loc[(data['helpful_rate'] < 0.7) & (data['helpful_votes'] >= 10), 'c'] = 0
df4 = data.groupby(['review_date'], sort=True)['c'].agg('mean').reset_index(name='c_mean')
x = df4.loc[:, 'review_date']
y = df4.loc[:, 'c_mean']
plt.plot(x, y)
plt.xlabel('review_date')
plt.ylabel('c')
plt.title('Changes in c with time')
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\置信度随时间的变化' + product + '.png')
plt.show()
# %% 综合评价
word_avg = data['words'].mean()
word_max = data['words'].max()
data['r'] = (0.3 * data['star_rating'] + 0.7 * (data['sentiment_val'] + (data['words'] - word_avg) / word_max)) \
            * data['c']

df5 = data.groupby(['review_date'], sort=True)['r'].agg('mean').reset_index(name='r_mean')

x = df5.loc[:, 'review_date']
y = df5.loc[:, 'r_mean']

plt.plot(x, y)
plt.xlabel('review_date')
plt.ylabel('r')
plt.title('Changes in r with time')
```

```
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\综合评价随时间的变化' + product + '.png')
plt.show()
# %% 星级和感情得分的关系
x = data.loc[:, 'star_rating']
y = data.loc[:, 'sentiment_val']
plt.scatter(x, y, c='blue', alpha=.4)
plt.xlabel('star_rating')
plt.ylabel('sentiment_val')
plt.title('The relationship between star_rating and sentiment_val')
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\星级和感情度的关系' + product + '.png')
plt.show()
# %% vine 与点赞率的关系
x = data.loc[:, 'vine']
y = data.loc[:, 'helpful_rate']
x.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\vine_' + product + '.csv', index=None)
y.to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\helpful_rate_' + product + '.csv', index=None)

# %%评论字数和点赞数的关系图
x = data.loc[:, 'words']
y = data.loc[:, 'helpful_rate']
plt.scatter(y, x, c='blue', alpha=0.4)
plt.xlabel('helpful_rate')
plt.ylabel('comment_words')
plt.savefig('E:\\njupt\\Problem_C_Data\\' + product + '\\评论字数和点赞数的关系图' + product + '.png')
plt.show()

# %% 关键词提取


data_neg = data[['review_headline_body']][data['sentiment'] == 'neg']
data_pos = data[['review_headline_body']][data['sentiment'] == 'pos']
data_neg.index = range(len(data_neg))
data_pos.index = range(len(data_pos))
data_neg['review_headline_body'] = data_neg['review_headline_body'].str.lower()
data_pos['review_headline_body'] = data_pos['review_headline_body'].str.lower()

# 分词
```

```
words = []
# TODO(ppy): 在下面的循环选则积极或消极
for index, row in data_pos.iterrows():
    words.extend(nltk.word_tokenize(row['review_headline_body']))
# 去除标点

inter_punctuations = [',', '.', ':', ';', '?', '(', ')', '[', ']', '&', '!', '*', '@', '#', '$', '%', '<', '>', '/',
                      '...', 'br']
words_dePunctuations = [word for word in words if word not in inter_punctuations]
# print(words_dePunctuations)

# 去除停用词
stops = set(stopwords.words('english'))
word_deStops = [word for word in words_dePunctuations if word not in stops]
# print(word_deStops)

# 提取词干(Porter 提取算法)
word_stem = []
for word in word_deStops:
    word_stem.append(PorterStemmer().stem(word))
# print(word_stem)

# 词性还原
word_reduction = []
for word in word_stem:
    word_reduction.append(WordNetLemmatizer().lemmatize(word))
# print(word_reduction)
ex_del = []
word_end = [word for word in word_reduction if word not in ex_del]

freq_dist = nltk.FreqDist(word_end)
standard_freq_vector = freq_dist.most_common(50)
print(standard_freq_vector)
# %% 输出清洗后的数据
print(data.describe())
data.describe().to_csv('E:\\njupt\\Problem_C_Data\\' + product + '\\描述性统计' +
product + '.csv')
output_file = 'E:\\njupt\\Problem_C_Data\\' + product + '\\' + product + '_clean.csv'
data.to_csv(output_file)
```