食品评论数据舆情分析

随着人们生活水平的提高和健康意识的增强,食品安全和质量问题越来越受到广大消费者的重视。互联网的普及和社交媒体的发展使得人们可以更加方便地表达对于食品的看法和感受,这也为食品评论数据舆情分析提供了丰富的数据来源。食品作为人们日常生活中的必需品,其销售和评价情况对于电商平台和食品制造商来说具有极高的参考价值。可以帮助企业和平台更好地了解市场动态和消费者需求,进而制定更精准的市场策略和产品改进方案。

本项目提供亚马逊食品评论数据利用数据可视化和情感分析进行分析,并提出相应的改进措施和建议。

学习目标

（1） 通过对食品评论数据进行分析,挖掘用户对食品的看法以及食品潜在的问题。

（2） 针对潜在的问题提出相应的措施和建议。

（3） 通过对评论数据进行情感分析,辅助确定评论的正负面情感。

1 了解食品评论数据现状和舆情分析

任务描述

近年来，食品安全问题频频发生，如添加剂超标、农药残留、过期食品等，这些问题严重影响了消费者的健康和权益。因此，对食品评论数据进行舆情分析，可以及时发现和解决食品安全问题，保障消费者的合法权益。

任务分析

（1） 了解食品评论现状。

（2） 了解食品评论数据基本情况。

（3） 了解舆情分析的方法和目的。

（4） 实现食品评论数据舆情分析的步骤和流程。

1.1食品评论现状

食品评论现状可以从多个维度进行分析，包括食品行业的发展趋势、食品安全

现状以及消费者对食品的关注和需求。随着消费者对健康的关注度提高，纯天然、低糖、低卡、低脂、低钠等健康食品受到追捧。品牌企业也积极回应这一需求，推出各种健康食品，并在产品包装和营销上强调健康属性。尽管我国的食品安全现状在近年来有所改善，但仍然存在一些安全问题，如食品原材料的农药残留、兽药使用，食品添加剂的滥用，以及假冒伪劣或变质的原料等。这些问题都影响着食品的安全和质量。随着网络的发展，越来越多的消费者会在网络上查看和发表食品评论。这些评论不仅影响着其他消费者的购买决策，也成为品牌企业改进产品和服务的重要参考。

1.2食品评论数据基本情况

本项目提供了亚马逊食品评论数据。该数据集中包含了约 568454 条用户评论，这些评论为分析提供了大量的样本。数据覆盖了从 1999 年 10 月至 2012 年 10 月的长达 13 年的时间，反映了这一时间段内消费者对食品的评价和喜好变化。每条评论包含了多个字段，用于记录不同类型的信息。该评论数据集是一个涵盖了较长时间跨度和大量用户评论的数据集，为食品行业的研究和市场分析提供了宝贵的资源。

1.3舆情分析的方法和目的

舆情分析的方法多种多样，通常包括以下几种：

（1） 数据收集。

（2） 数据预处理。

（3） 关键词分析。

（4） 情感分析。

舆情分析的目的包括以下几种：

（1） 发现潜在风险。

（2） 优化产品策略。

（3） 提升监管水平。

（4） 为政府提供数据支持。

2 预处理食品评论数据

任务描述

经观察，食品评论数据存在缺失值、异常值和重复值，且特征过多，需要对食品评论数据进行预处理。

任务分析

（1） 处理缺失值、重复值和异常值。

（2） 处理时间类型数据和文本数据。

（3） 特征构造。

2.1处理缺失值、重复值和异常值

数据读取成功后，利用 reviews.head()命令展示前几行数据。

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

图 2-1 部分原始数据展示

由观察可知，原始食品评论数据中存在缺失值、重复值和异常值。对食品评论数据进行缺失值处理，重复值处理，删去重复的文本，只保留第一个，异常值处理并查看剩余的异常值。

代码 2-1 缺失值、重复值、异常值的处理

```
#缺失值处理
reviews.isnull().sum()
reviews.dropna(inplace=True)
print(reviews.shape)
#重复值处理,删去重复的文本，只保留第一个
print(reviews.duplicated(subset=['Text']).sum())
reviews.drop_duplicates(subset=['Text'],keep='first',inplace=True)
print(reviews.shape)
# 异常值处理
reviews[reviews["HelpfulnessNumerator"]>reviews["HelpfulnessDenominator"]]
reviews.drop([44736,64421],inplace=True)
```

```
# 查看剩余的异常值
reviews[reviews["HelpfulnessNumerator"]>reviews["HelpfulnessDenominator"]]
reviews.reset_index(drop=True,inplace=True)
```

运行代码 2-1 得到的结果如下。

```
(568427, 9)
174851
(393576, 9)
```

经过分析，缺失值处理后形状展示为（568427, 9），重复值有 174851 个，经处理后形状展示为（393576, 9），异常值评论有两条，处理后无异常值。

2.2 处理时间类型数据和文本数据

首先对时间类型数据进行处理，再对文本数据进行处理。文本数据处理包括将所有字母转化为小写字母，删除非英文字符，去停用词等操作。

代码 2-2 时间类型数据和文本数据的处理

```
#时间类型数据进行处理
reviews['Time'] = pd.to_datetime(reviews['Time'],unit='s')
reviews['Time']
reviews['year'] = reviews['Time'].dt.year
reviews.head()
#文本数据处理
reviews['Text']
#将所有字母转换为小写字母
reviews["Text"]=reviews["Text"].str.lower()
reviews["Summary"] = reviews["Summary"].str.lower()

#删除非英文字符
import re
reviews['Text_cl'] = reviews['Text'].apply(lambda x: re.sub(r'[^a-z]+', ' ', x))
reviews['Summary'] = reviews['Summary'].apply(lambda x: re.sub(r'[^a-z]+', ' ', x))
#去停用词
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
sw = stopwords.words('english')
sw = sw+['br']
sw
```

```
reviews["Text_cl"] =reviews["Text_cl"].apply(lambda x:" ".join(x for x in str(x).split() if x not in sw or x == "not"))
reviews["Summary"] =reviews["Summary"].apply(lambda x:" ".join(x for x in str(x).split() if x not in sw or x == "not"))
reviews['Text_cl']
```

经处理后，数据发生了变化，图 2- 2 展示了 2.2 小节数据处理后的评论形式。

```
0         bought several vitality canned dog food produc...
1         product arrived labeled jumbo salted peanuts p...
2         confection around centuries light pillowy citr...
3         looking secret ingredient robitussin believe f...
4         great taffy great price wide assortment yummy ...
                              ...
393569    great sesame chicken good not better resturant...
393570    disappointed flavor chocolate notes especially...
393571    stars small give one training session tried tr...
393572    best treats training rewarding dog good groomi...
393573    satisfied product advertised use cereal raw vi...
Name: Text_cl, Length: 393574, dtype: object
```

图 2-2 经代码 2-2 处理后的评论形式

由图 2- 2 可以观察到，数据中的评论字母全部转换成了小写，不存在非英文字符，数据处理成功。

2.3特征构造

构造评论有用性特征。

代码 2-3 评论有用性特征构造

```
#构造评论有用性特征
import numpy as np
result=np.where(reviews['HelpfulnessDenominator']==0,np.nan,reviews['HelpfulnessNumerator']/reviews['HelpfulnessDenominator'])
reviews['Usefulness'] = np.where(result>0.5,'useful',np.where(np.isnan(result),'unknown','useless'))
reviews
```

评论有用性特征构造完毕后，图 2- 3 为部分数据展示情况。

	Id	ProductId	UserId	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text	year	Text_cl	Usefulness
0	1	B001E4KFG0	A3SGXH7AUHU8GW	1	1	5	2011-04-27	good quality dog food	i have bought several of the vitality canned d...	2011	bought several vitality canned dog food produc...	useful
1	2	B00813GRG4	A1D87F6ZCVE5NK	0	0	1	2012-09-07	not advertised	product arrived labeled as jumbo salted peanut...	2012	product arrived labeled jumbo salted peanuts p...	unknown
2	3	B000LQOCH0	ABXLMWJIXXAIN	1	1	4	2008-08-18	delight says	this is a confection that has been around a fe...	2008	confection around centuries light pillowy citr...	useful
3	4	B000UA0QIQ	A395BORC6FGVXV	3	3	2	2011-06-13	cough medicine	if you are looking for the secret ingredient i...	2011	looking secret ingredient robitussin believe f...	useful

图 2-3 经代码 2-3 处理后的数据

由图 2- 1 和图 2- 2 对比可以观察到，经过数据处理，成功在数据集中添加了 year，Text_d，Usefulness 三列数据，方便了后续可视化处理与情感分析。

3 数据可视化

任务描述

数据可视化能够将大量、复杂的食品评论数据转化为易于理解的图形和图像，使得数据呈现更加直观、生动。这有助于人们更快地理解和把握数据的内在规律和趋势。通过数据可视化，可以更容易地发现食品评论数据中隐藏的模式、趋势和关联关系。数据可视化可以揭示数据中可能存在的异常值、缺失值或错误数据，从而及时发现潜在问题。这些问题可能影响到食品评论数据的准确性和可靠性，进而影响到决策的正确性。

任务分析

（1） 评分占比分析。

（2） 评论有用性分析。

（3） 用户情况分析。

（4） 评论内容分析。

（5） 热评产品分析。

3.1评分占比分析

对食品评论数据进行评分占比分析，展示 reviews 数据框中 Score 列的不同值的分布情况，autopct 参数会在饼图中显示每个扇区的百分比。

代码 3-1 评分占比处理

```
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']=['SimHEI']

tmp1 = reviews['Score'].value_counts()
plt.pie(tmp1,labels=tmp1.index,autopct='%1.1f%%')
plt.title('评分占比')
plt.show()
```

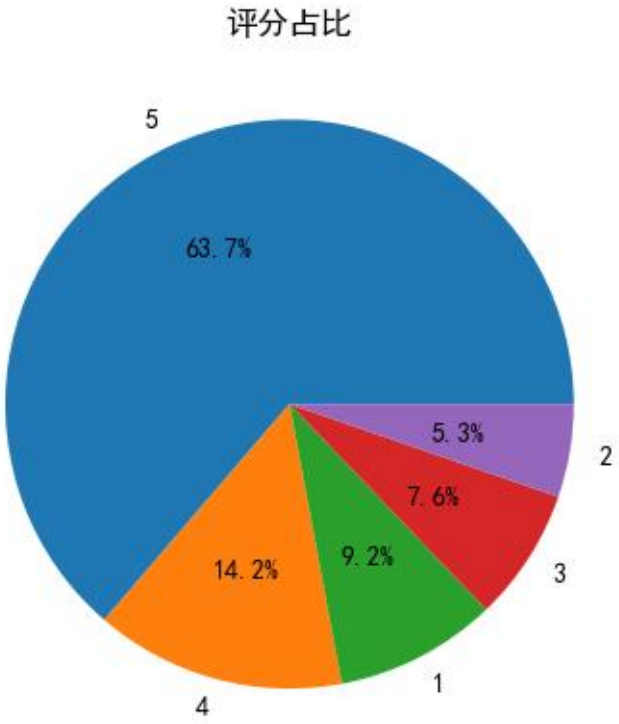

图 3-1 评分占比图

图 3- 1 即为代码 3- 1 的运行结果，每个扇区的大小代表了对应评分的数量可以看出一共有 5 种评分，1-5 分，在所有的评论中其中 5 评分占比最多达 63.7%，表明消费者对食品具有较高的满意度。相反，2 分评论占比最少仅占 5.3%，表明消费者对极不满意的产品相对较少。1、3 和 4 分分别占比 9.2%、7.6%和 14.2%。这种分布揭示了整体的评论趋势是呈正面的。

3.2评论有用性分析

使用 matplotlib 库来展示数据集中'Usefulness'列的不同值的分布情况，计算'Usefulness'列中每个值的出现次数，绘制饼图，显示每个类别的百分比。

代码 3-2 评论有用性分析处理

```
#评论有用性的占比环形图
tmp2 = reviews['Usefulness'].value_counts()
plt.pie(tmp2,labels=tmp2.index,autopct='%1.1f%%')
centre_circle = plt.Circle((0,0),0.5,fc='white')
fig = plt.gcf().gca().add_artist(centre_circle)
plt.title('评分有用性占比')
plt.show()
```

绘制结果如下：

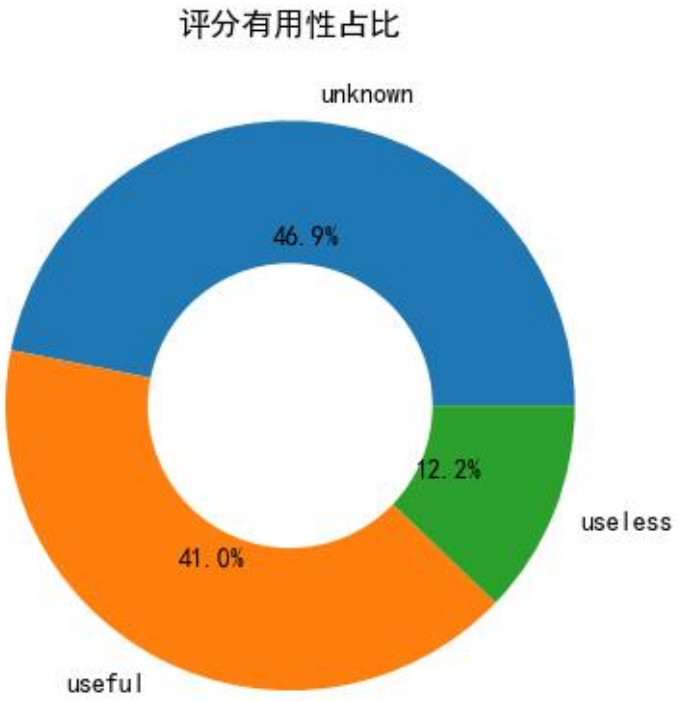

图 3-2 评分有用性占比图

图 3- 2 即为代码 3- 2、的运行结果，这个环形图呈现了评论是否有用的占比情况。从这个环形图可以看出约有 41.0%的评论被认为是有用评论,46.9%的评论没有用户点评是否有用,12.2%的评论被认为是无用评论。用户对于评论质量的评判存在显著差异，且有大部分的评论有效性有待进一步验证。

继续使用 matplotlib 库来展示不同评分下评论的有用性分布情况，首先按评分和有用性进行分组，并计算每组的评论数量，然后提取不同有用性评级的评论数量，创建图形。

代码 3-3 各评分评论有用性分析处理

```
#各评分中评论有用性分布条形图
tmp3 = reviews[['Score','Usefulness']].groupby(by=['Score','Usefulness'],as_index=False).agg(count=('Usefulness','count'))
x = tmp3['Score'].unique()
y_useful = tmp3[tmp3['Usefulness']=='useful']['count'].values
y_useless = tmp3[tmp3['Usefulness']=='useless']['count'].values
y_unknown = tmp3[tmp3['Usefulness']=='unknown']['count'].values

fig,ax = plt.subplots()
width = 0.25
```

```
rects1 = ax.bar(x-width,y_useful,width,label='useful')
rects2 = ax.bar(x,y_useless,width,label='useless')
rects3 = ax.bar(x+width,y_unknown,width,label='unknown')

ax.set_title('各评分评论有用性')
ax.legend()
plt.show()
```

绘制结果如下：

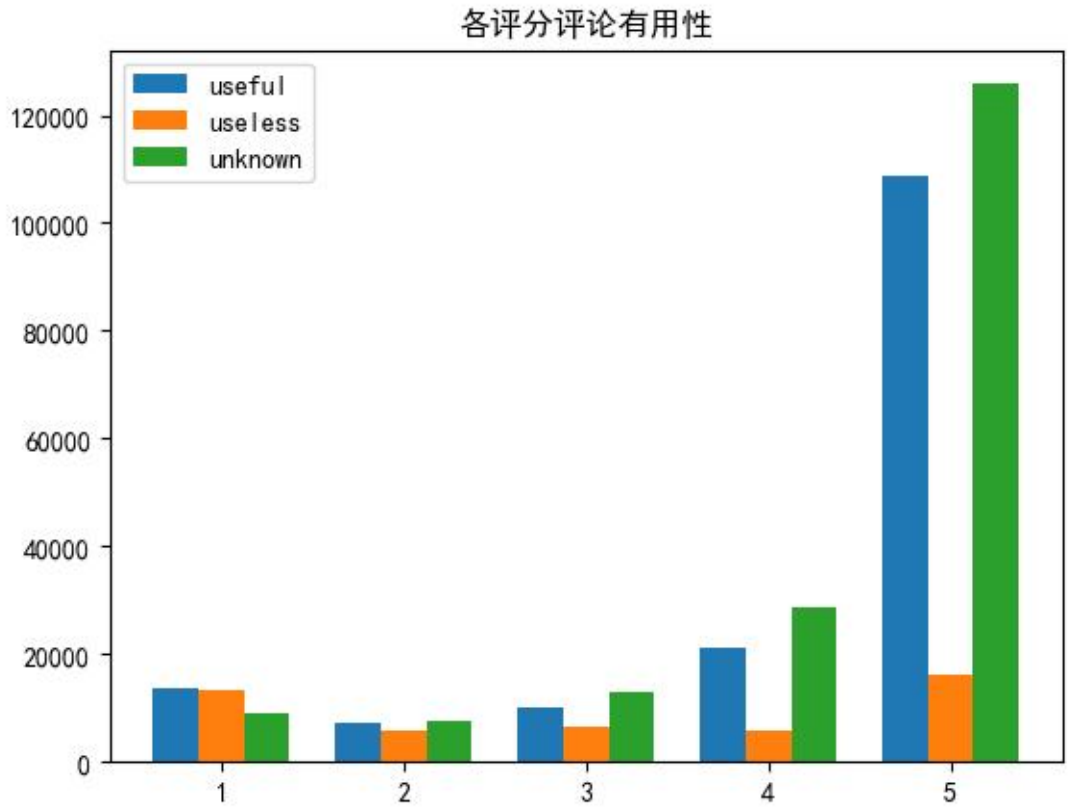

图 3-3 各评分评论有用性分析图

图 3- 3 即为代码 3- 3 的运行结果，条形图分为三组，分别代表有用、无用和未知的评论数量。每组条形图在 x 轴上的位置根据评分值确定，且相互之间有 0.25 宽度的间隔。这个条形图呈现了各评分评论是否有用的情况。从这个条形图可以看出 5 分的评论中被认为是有用的评论是最多的，同时，无用的评论与有用的评论数量相差较大。1-4 分评论中的有用评论数量远低于 5 分评论的有用性评论。

条形图绘制完毕，接下来绘制历年评分层级和趋势图，创建一个折线图，展示不同年份的评分分布和评论数量的变化趋势。

代码 3-4 历年评分星级和评论趋势分析处理

```
tmp4 = reviews[['year','Text']].groupby(by='year',as_index=False).count()
tmp4_1 = reviews[['year','Score']].groupby(by=['year','Score'],as_index=False).size()
```

```
pivot_data = tmp4_1.pivot(index='year',columns='Score',values='size').fillna(0)


plt.figure(figsize=(12,8))
for i in pivot_data.columns:
    plt.plot(pivot_data.index,pivot_data[i],label=f'{i}')
plt.plot(tmp4['year'],tmp4['Text'],label='评论趋势',marker='o')
plt.title('历年评分层级和趋势')
plt.legend()
plt.show()
```

绘制结果如下：

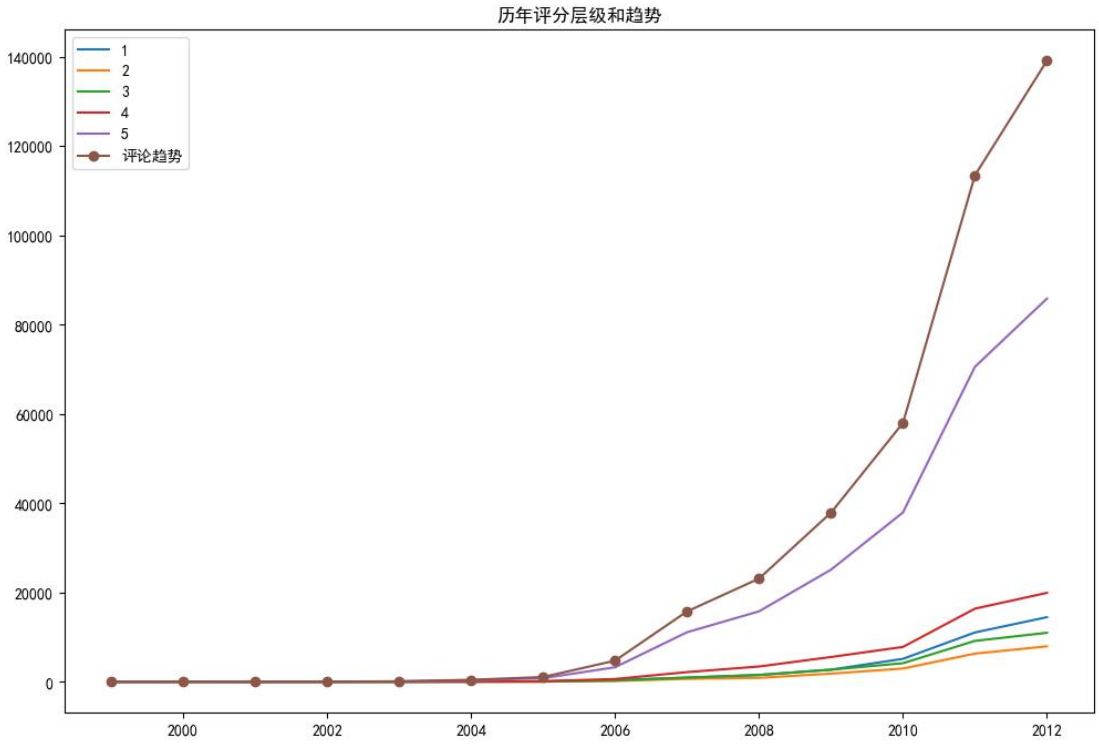

图 3-4 历年评分层级和趋势图

图 3- 4 即为代码 3- 4 的运行结果，对于每个评分等级，都会有一条折线显示其在不同年份的评论数量。另外，还有一条折线显示了每年的总评论数量。

这张折线图中，棕色的折线代表的是评论趋势，从这张图可以看到评论数量逐年增加，在 2010 年呈迅猛上升。反映了消费者在线评价行为的加速普及，以及用户的活跃度，这也与互联网的普及、网购等的普及有关。各评分的折线图总体来看都呈一个上升的趋势，其中 5 评分这条折线的上升趋势明显比其他评分的上升趋势更快。

3.3用户情况分析

对于评论次数大于 100 的用户进行观察，分析他们对产品的评价好坏，以便对产品进行改进。创建一个条形图，其中展示至少发表了 100 条评论的用户的评论数量。条形图的高度代表了每个用户发表的评论数量。

代码 3-5 用户情况分析处理

```
#用户情况分析
user_count = reviews['UserId'].value_counts()
selected_user = user_count[user_count>100]

plt.figure(figsize=(10,6))
selected_user.plot(kind='bar')
plt.title('常评论用户')
plt.show()
```

绘制结果如下：

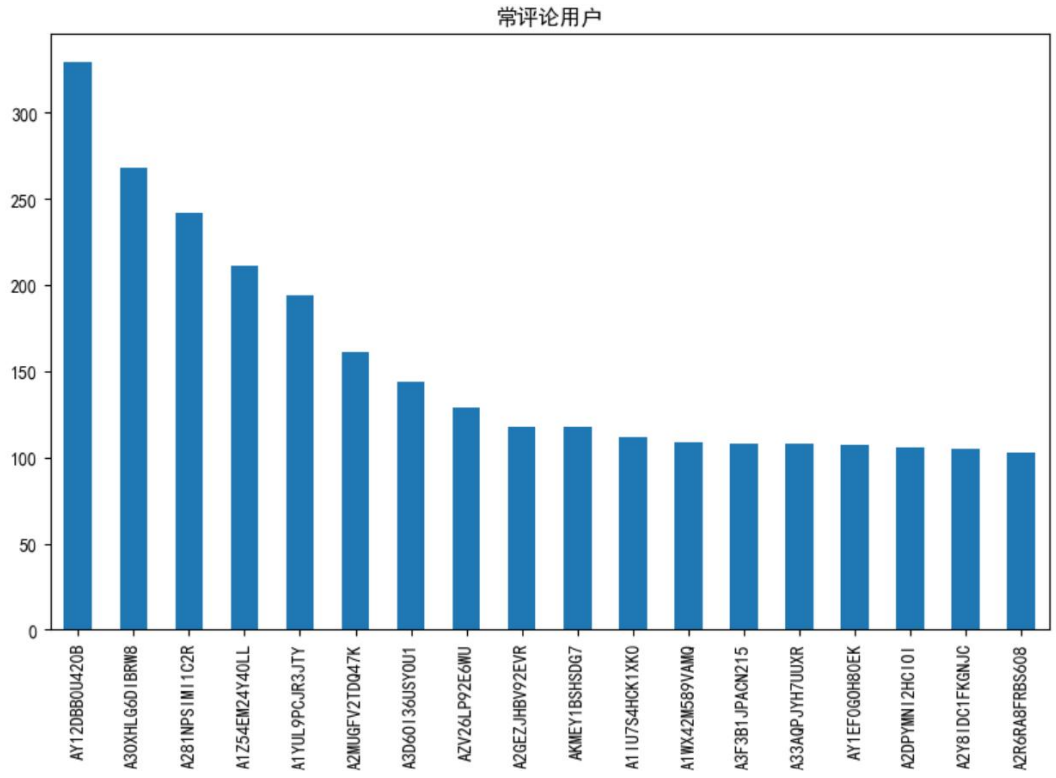

图 3-5 常评论用户展示图

图 3- 5 即为代码 3- 5 的运行结果，这张条形图横坐标轴代表用户名称，纵坐标轴代表用户评论数量，评论最多的用户评论数量达到了 300 多，远高于其他用户评论数量。常评论用户柱状图可以看出评论最多的用户是 AY12DBB0U420B,可能是重点客户。在评论数量较多的客户中，存在重点客户和有价值的客户，都是商家需要重点关注的用户。

3.4评论内容分析

在此绘制评论中高频词的词云图进行分析，仅选取评论的前 200 条。

代码 3-6 评论内容分析处理

```
#评论内容分析
from wordcloud import WordCloud
from collections import Counter
import matplotlib.pyplot as plt

def getWordCloud(text, title, i=None):
    common_words = Counter(" ".join(text).split()).most_common(i)
    word_freq = dict(common_words)
    wordcloud = WordCloud(width=450, height=250, background_color="white",
min_font_size=10).generate_from_frequencies(word_freq)
    plt.figure(figsize=(10, 6))
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis("off")
    plt.title(title, fontsize=15)
plt.show()
# 评论中前 200 的高频词云图
getWordCloud(reviews["Text_cl"],"评论高频词",200)
```

图 3-6 评论高频词图

图 3-6 展示了评论高频词，选择评论文本前 200 个高频词，反应客户对食品总体的看法，评分基本都是正面情绪的词，如“good”、“like”、“best”等词表示对产品的

满意，还有“coffee”、“tea”这些代表产品的词，另外还有“not”表示否定的负面情绪词。

3.5热评产品分析

对热评产品进行分析，可以知道消费者热衷于这些产品的原因，从而加以推广利用至其他产品。在此仅对前十个热评产品的评分星级分布进行分析。

代码 3-7 热评产品的评分星级

```
#前十个热评产品的评分星级分布
plt.rcParams['font.sans-serif']=['SimHEI']
ind = reviews.groupby('ProductId').size().sort_values(ascending=False)[:10].index
print(ind)

tmp = reviews[reviews['ProductId'].isin(ind)]
tmp5 = tmp[['ProductId','Score']].groupby(by=['ProductId','Score'],as_index=False).agg(count=('Score','count'))
import seaborn as sns
fig,ax = plt.subplots(figsize=(12,12))
sns.barplot(x='count',y='ProductId',hue='Score',data=tmp5)
plt.title('前十个热评产品的评分星级分布')
plt.show()
```

绘制结果如下：

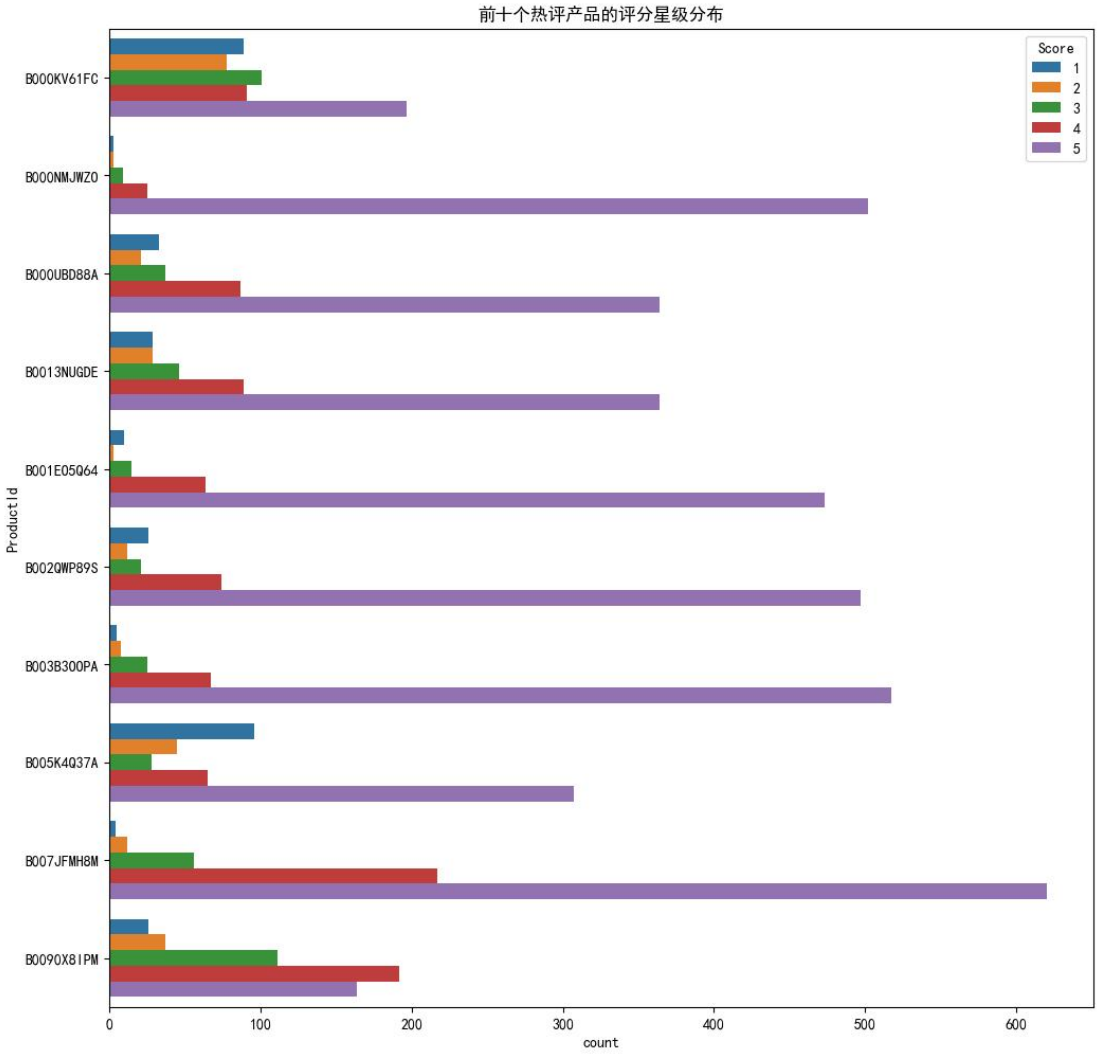

图 3-7 前十个热评产品的评分星级分布图

图 3- 7 即为代码 3- 7 的运行结果，展示了前十个热评产品的评分情况，不同颜色代表了不同的评分星级，纵坐标表示产品的名称，横坐标表示不同评分的产品数量。由图可以看出紫色占比最多，即 5 评分占比最大，也就是说热评产品中大部分都是 5 分。

接下来对热评第一的产品进行分析，绘制热评第一产品的关键词词云图。

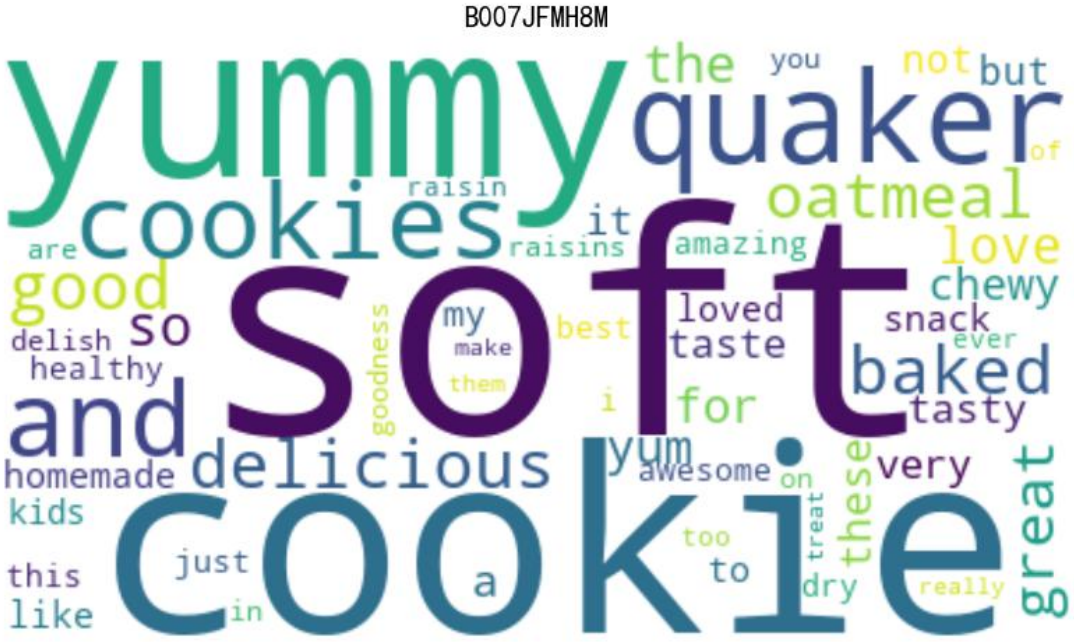

图 3-8 热评第一产品的关键词词云图

图 3- 8 展示了热评第一的产品关键词词云，该产品可以看出是一款饼干，其中大部分词汇是正面的，如“yummy”、“delicious”、“great”等，表示用户对于这种饼干有积极的评价。可以看出大部分用户对产品是满意的，也存在小部分不满意（“not”）。

接下来将 1-2 分归为低评分，3 分为中评分，4-5 分为高评分，绘制各评分评论内容的词云图。

代码 3-8 热评第一的产品低中高评分词云分析

```
# 1-2 低评分
# 3 中评分
# 4-5 高评分
getWordCloud(pro_top1.loc[pro_top1["Score"]<=2,"Text_cl"],"低评分")

getWordCloud(pro_top1.loc[pro_top1["Score"]==3,"Text_cl"],"中评分")
getWordCloud(pro_top1.loc[pro_top1["Score"]>3,"Text_cl"],"高评分")
```

图 3-9, 3-10, 3-11 是热评第一的产品各评分词云图。

低评分

图 3-9 热评第一产品的低评分词云图

中评分

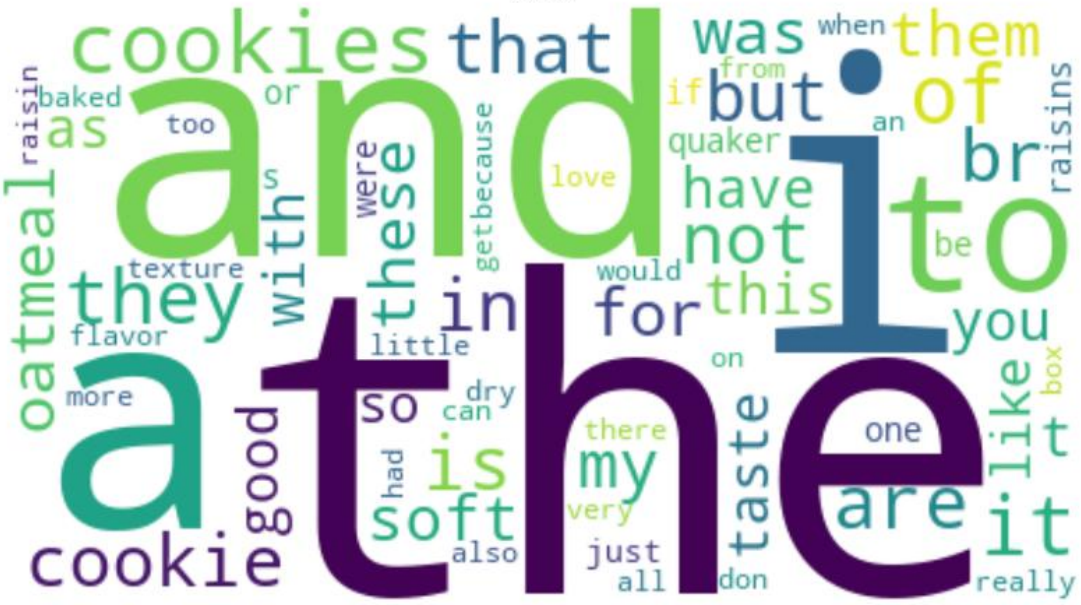

图 3-10 热评第一产品的中评分词云图

高评分

图 3-11 热评第一产品的高评分词云图

在这些词云图中可以看到“cookie”、“oatmeal”、“taste”、“soft”等等这些词语，从这些词语可以看出用户是从味道、成分和营养方面对产品进行评价。从低评分可以看到“sugar”、“texture”、“dry”等词，表示用户对这些方面不满意，商家可以着重看待这几方面的问题并加以改进。

4 情感分析

4.1情感打分

为了辅助去判断评论文本的正负面情感，进行情感打分处理。

代码 4-1 情感打分处理

```
#情感打分
import nltk
nltk.download('vader_lexicon') # 确保这行代码在尝试使用 SIA 之前执行

!pip install vaderSentiment

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

sia=SentimentIntensityAnalyzer()
res={}
```

```
for i,row in tqdm(reviews.iterrows()):
    text=row['Text_cl']
    my_id=row['Id']
    res[my_id]=sia.polarity_scores(text)
```

绘制代码结束后，观察打分的情况，在此截取 7 条数据进行观察。

```
{1: {'neg': 0.0, 'neu': 0.503, 'pos': 0.497, 'compound': 0.9413},
 2: {'neg': 0.092, 'neu': 0.801, 'pos': 0.106, 'compound': 0.0762},
 3: {'neg': 0.165, 'neu': 0.563, 'pos': 0.272, 'compound': 0.7926},
 4: {'neg': 0.0, 'neu': 0.854, 'pos': 0.146, 'compound': 0.4404},
 5: {'neg': 0.0, 'neu': 0.369, 'pos': 0.631, 'compound': 0.9468},
 6: {'neg': 0.088, 'neu': 0.681, 'pos': 0.231, 'compound': 0.811},
 7: {'neg': 0.0, 'neu': 0.559, 'pos': 0.441, 'compound': 0.9463},
```

图 4-1 部分评论的情感打分情况

由图 4-1 可以看到，每条评论共有四个打分情况，neg 代表消极的，neu 代表中性的，pos 代表积极的，compound 代表综合评分。这样可以清晰的观察到每条评论的四个打分情况。

与之前数据拼接在一起后，数据呈现如下状态，可以观察到，在 Id 这一列的后面增加了新的四列，分别是 neg，neu，pos，compound。

	Id	neg	neu	pos	compound	ProductId	UserId	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text	year	Text_cl	Usefulness
0	1	0.000	0.503	0.497	0.9413	B001E4KFG0	A3SGXH7AUHU8GW	1	1	5	2011-04-27	good quality dog food	i have bought several of the vitality canned d...	2011	bought several vitality canned dog food produc...	useful
1	2	0.092	0.801	0.106	0.0762	B00813GRG4	A1D87F6ZCVE5NK	0	0	1	2012-09-07	not advertised	product arrived labeled as jumbo salted peanut...	2012	product arrived labeled jumbo salted peanuts p...	unknown
2	3	0.165	0.563	0.272	0.7926	B000LQOCH0	ABXLMWJIXXAIN	1	1	4	2008-08-18	delight says	this is a confection that has been around a fe...	2008	confection around centuries light pillowy citr...	useful
3	4	0.000	0.854	0.146	0.4404	B000UA0QIQ	A395BORC6FGVXV	3	3	2	2011-06-13	cough medicine	if you are looking for the secret ingredient i...	2011	looking secret ingredient robitussin believe f...	useful
4	5	0.000	0.369	0.631	0.9468	B006K2ZZ7K	A1UQRSCLF8GW1T	0	0	5	2012-10-21	great taffy	great taffy at a great price. there was a wid...	2012	great taffy great price wide assortment yummy ...	unknown

图 4-2 处理后的部分数据展示

4.2探究评分星级与情感评分的关系

探究不同的评分与评分星级的关系，设置 x 轴为原来的评分星级，y 轴为情感评分。

代码 4-2 评分与评分星级关系探究处理

```
fig,axs = plt.subplots(2,2,figsize=(16,10))
sns.barplot(data=new_data,x="Score",y='pos',ax=axs[0,0])
sns.barplot(data=new_data,x="Score",y='neu',ax=axs[0,1])
sns.barplot(data=new_data,x="Score",y='neg',ax=axs[1,0])
sns.barplot(data=new_data,x="Score",y='compound',ax=axs[1,1])
```

```
axs[0,0].set_title('Positive')
axs[0,1].set_title('Neutral')
axs[1,0].set_title('Negative')
axs[1,1].set_title('Compound')
plt.tight_layout()
plt.show()
```

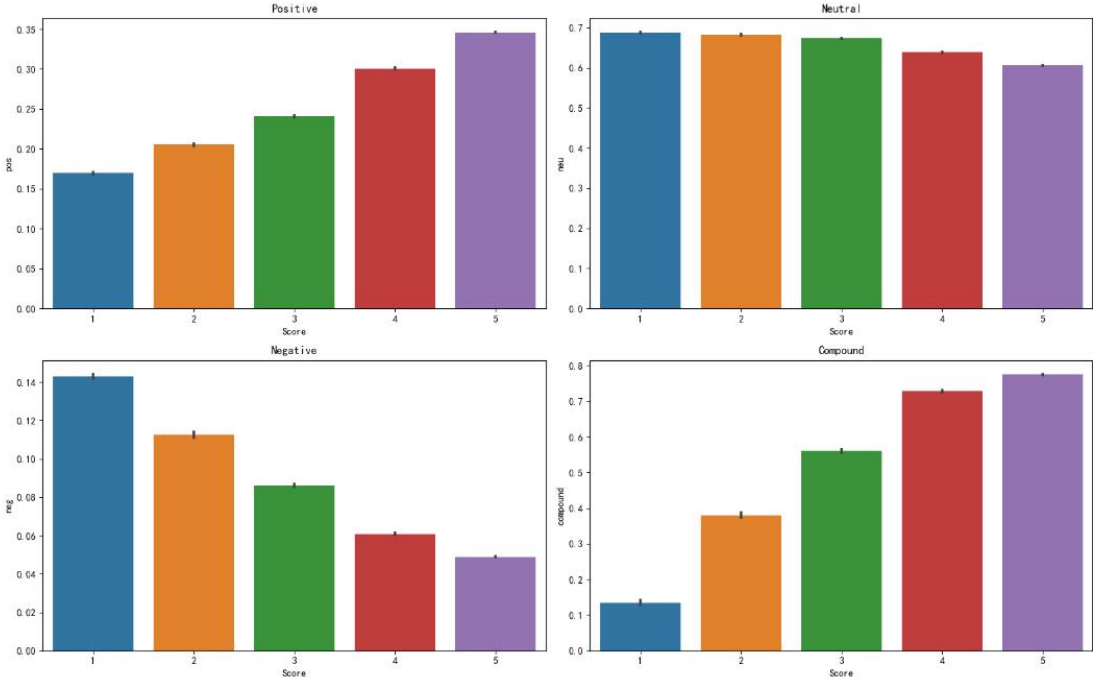

图 4-3 各评分星级与情感评分关系图

图 4- 3 即为代码 4- 2 的运行结果，可以观察到，积极的评分星级的分布是 5 分的是最多的，中评的评分星级 1-5 分分布均匀，消极的评分星级 1,2 分占大多数。这说明情感打分与先前评分星级是有关联的，可以总结出评分星级越高，情感越积极。

将评分星级进行等级划分，将小于等于 2 的评分划分为消极的，等于 3 则为中性的，否则划分为积极的，并将其赋值为新一列。同时对综合评分也进行划分，评分大于 0 划分为积极的，评分等于 0 则为中性的，否则为消极的，同样将其赋为新列。

通过对比新增的两列数据可以得出一个准确率，统计原来评分等级与情感分析中的评分等级是一致的数量除以所有评论数量总数就可以得到一致准确率。

代码 4-3 等级划分处理

```
#等级划分
new_data['Sentiment'] = new_data['Score'].apply(lambda score:'Negative' if score<=2 else 'Neutral'if score==3 else 'Positive')
new_data['nltk_Sentiment'] = ['Positive' if sentiment>0 else 'Neutral' if sentiment==0 else 'Negative' for
```

```
sentiment in new_data['compound']]

same_count = (new_data['Sentiment']==new_data['nltk_Sentiment']).sum()
round(same_count/len(new_data),2)
```

准确率运行结果如下：

```
0.79
```

由结果可以看出，一致准确率为 0.79。

打印出情感分析结果与原来结果不一样的数据。

	Text	Sentiment	nltk_Sentiment
0	product arrived labeled as jumbo salted peanut...	Negative	Positive
1	if you are looking for the secret ingredient i...	Negative	Positive
2	one of my boys needed to lose some weight and ...	Positive	Negative
3	my cats have been happily eating felidae plati...	Negative	Positive
4	i love eating them and they are good for watch...	Negative	Positive
...	...	...	...
80984	i thought this soup would be more like a chill...	Negative	Positive
80985	i just bought this soup today at my local groc...	Negative	Positive
80986	this soup is mostly broth. although it has a k...	Negative	Positive
80987	it is mostly broth, with the advertised 3/4 cu...	Negative	Positive
80988	i had ordered some of these a few months back ...	Negative	Neutral

80989 rows × 3 columns

图 4-4 Sentiment 与 nltk_Sentiment 结果不一致的数据

由图 4- 4 可知，第 0 条评论起初被评为消极的，但情感打分为积极的。

打印一条评论 tmp6['Text'][4]，观察其内容并进行分析。

```
'i love eating them and they are good for watching tv and looking at movies! it is not too sweet. i like to transfer them to a zip lock baggie so they stay fresh so i can take my time eating them.'
```

这条评论的意思是“我喜欢吃它们，尤其是看电影和看电视的时候，那很棒，它不是很甜。我喜欢把它们放在一个带拉链的袋子里，这样它们可以保持新鲜，我就可以慢慢吃了。”由此看来这条评论总体观察为积极的，而之前是消极的，可见经过改正，情感分析后的结果更加准确。

5 小结

本案例先对评论数据进行预处理,然后进行可视化分析,最后进行情感分析。通过情感分析，可以洞察消费者最真实的感受。在分析过程中，发现了许多可以改进的地方。例如可以密切关注负面评论,了解客户不满意的方面,并努力改进；根据正面评论中的亮点,继续强化和宣传产品或服务的优势。同时提升客户体验也非常重要，需要分析评论中提到的服务问题,如员工态度、回复的及时性等,并针对性地进行培训或流程优化。考虑引入新技术,如优化平台的功能等,以提高效率和客户满意度。在市场营销策略方面，利用社交媒体和在线平台积极推广正面评论和口碑,吸引更多潜在顾客。针对不同的客户群体制定个性化的营销策略,如针对年轻人推广新口味或潮流食品。