**Predicting the marketing strategy of e-commerce products**

**based on text big data**

By Ding Lichao and Hu Zijiao

2021/10/30

**1 Introduction**

Nowadays with more and more merchants joining the e-commerce camp, competition in Internet marketing has become increasingly intense. Meanwhile, as more and more customers participate in online reviews and interactions, comprehensive analysis of reviews and ratings plays an increasingly important role in understanding customers’ pain points and specifying future strategies. Specifically speaking, in Amazon’s design, customers are allowed to choose a number from 1 to 5 to express their satisfaction level with the product, write any text-based messages as reviews and freely vote for other reviews they consider helpful. These are the main data sources for related companies to gain insights into the markets in which they participate, the timing of that participation, and the potential success of product design feature choices.

**2 Assumption**

*●****Assumption 1.*** *The online marketplace operates stably. And there were no situations such as an outbreak of an epidemic which would seriously affect the production chain of online shopping.*

●***Assumption 2.*** *The ratings and reviews depict customers’ real experience and feeling about their purchased products. The sentiment in the review text reflects one’s feelings on the products.*

●***Assumption 3.*** *The vast majority of individual differences of customers e.g., economic status and educational level, are ignored.*

**3 Data acquisition**

We download three products’ data sets which are hair dryer,microwave and pacifier.

Each data set contains the following:

●marketplace（string）

●customer\_id（string）

●review\_id（string）

●product\_id（string）

●product\_parent（string）

●product\_title（string）

●product\_category（string）

●star\_rating（int）

●helpful\_votes（int）

●total\_votes（int）

●vine（string）

●verified\_purchase（string）

●review\_headline（string）

●review\_body（string）

●review\_date（bigint）

**4 Data Clean**

We need to:

●Delete reviewers didn’t buy the products actually.

●Delete the column of marketplace and product\_category.

●Rearrange index.

Then we will get three new data set after cleaning.

Essential Code：



### 4 Training Model to Predict Star-rating and Vader Sentiment Analysis

In this part, we analysized the relationship between reviews’ texts and star-ratings. We train a model (from xgboost import XGBClassifier) and it can predict the star-ratings the users will give based on the reviews.

Input: reviews

Process:

Rebalance data

star\_count = [pdf[pdf.star\_rating == num].star\_rating.count() for num in range(1, 6)]

max\_star\_count = max(star\_count)

pdf = pd.concat(

    [pdf[pdf.star\_rating == index + 1].sample(frac=max\_star\_count // star\_count[index], replace=True) for index in

     range(5)], ignore\_index=True)

Split data

cv = CountVectorizer(

    max\_features=5000,

    encoding="utf-8",

    ngram\_range=(1, 3),

    stop\_words={"english"})

X = cv.fit\_transform(pdf.review\_body).toarray()

y = pdf['star\_rating']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

    X, y, test\_size=0.25)

Fit data in a matrix and train model

eval\_set = [(X\_train, y\_train), (X\_test, y\_test)]

eval\_metric = ["auc", "merror"]

model = XGBClassifier()

model.fit(X\_train, y\_train, eval\_set=eval\_set, eval\_metric=eval\_metric, verbose=True)

Make predictions for test data

y\_predictions = model.predict(X\_test)

y\_predictions = [round(value) for value in y\_predictions]

Evaluate predictions

accuracy = accuracy\_score(y\_test, y\_predictions)

print("Accuracy: %.2f%%" % (accuracy \* 100.0))

Outpt:

Star-rating (Users may give based on this review)

print(model.predict(cv.transform([

    "This is the worst product. I will never buy this again!",

    "I love it so much.",

    "Looks fine to me, but too expensive."

]).toarray()))

These three reviews will get scores:1,5,3

Accuracy: 88.41%

Vader Sentiment Analysis:

We call the Library:

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

import csv

input\_file = open('review\_body\_pacifier.txt',encoding='utf-8')

output\_file = open("review\_sentiment.txt", mode="w",encoding='utf-8')

csv\_writer = csv.writer(output\_file)

sentences = input\_file.readlines()

analyzer = SentimentIntensityAnalyzer()

for sentence in sentences:

    vs = analyzer.polarity\_scores(sentence)

    csv\_writer.writerow([sentence,vs['compound']])

    output\_file.write(str(sentence)+str(vs['compound'])+'n')

input\_file.close()

output\_file.close()

It can analysize each reviews’ sentiment score, A positive number indicates positive emotions and a negative number indicates a negative emotions.

Output: Sentiment score of each reviews



**5 Analytic Hierarchy Process Based on Information Entropy**

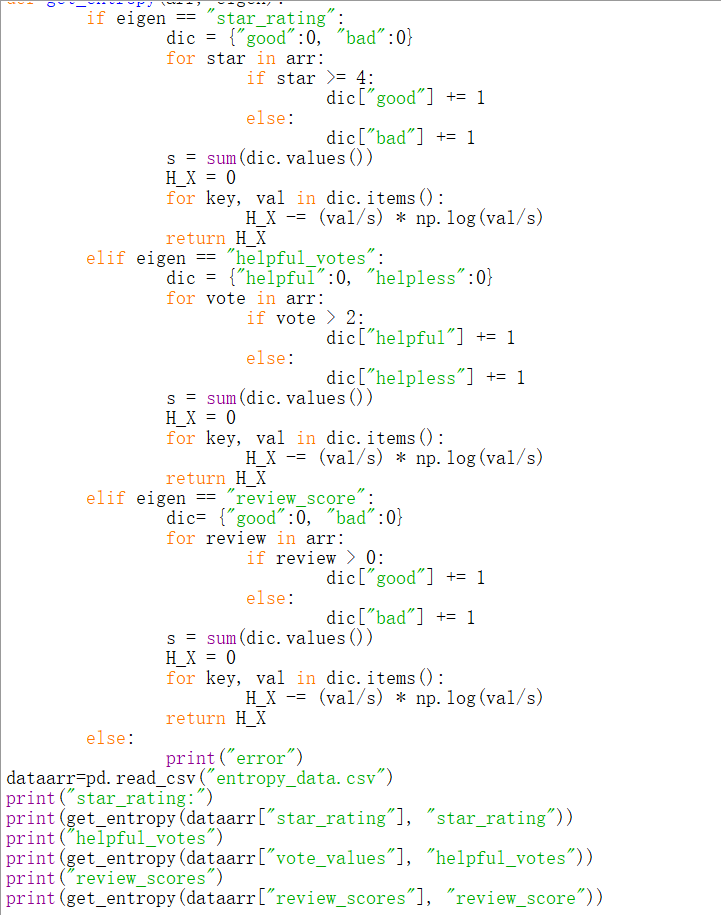
Information entropy is a value that measures the amount of information carried by a set of data. We use the information entropy of star\_rating, help\_factors and review\_score as a reference to evaluate the amount of information contained in each factor. In order to deﬁne the score that describes product evaluation, we establish an AHP model (The Analytic Process). By comparing the importance to the score among the three factors, we assign certain weight to each factor, so as to get the evaluation system.We established a AHP model, which focuses on the following three aspects:

1. Star\_rating: A means used by consumers to score products in an evaluation system. The website will show the average score on the product page, and each comment will show the score. When a consumer purchases a product, the score will affect the consumer’s purchase intention and thus affect the sales volume of the product, so it is taken into account.
2. Useful evaluation coefﬁcient H: In the evaluation system, the index that reﬂects the quality of the review. Consumers will browse product reviews when purchasing products, and vote on reviews when they think the reviews are useful. The more votes a comment has, the more reference it will have.
3. Review: In the evaluation system, the way consumers describe their experience of pur- chasing products. If consumers have a desire to buy this product, they will know more about the product information in detail, including comments. The quality of the comments will increase or decrease consumers’ desire to buy, so they will be taken into consideration. Here, review has been quantiﬁed as a review\_score.

In order to reduce the influence brought by the subjective effect of AHP model, we use infor mation entropy to help determine the weight of three factors. Information entropy is a measure that represents the uncertainty of random variables.

The greater the value of information entropy, the greater the uncertainty of random variables. So we can use information entropy to measure the information contained in a random variable. According to the above principle, we calculated the information entropy of the three factors of the three products. The results are as follows:

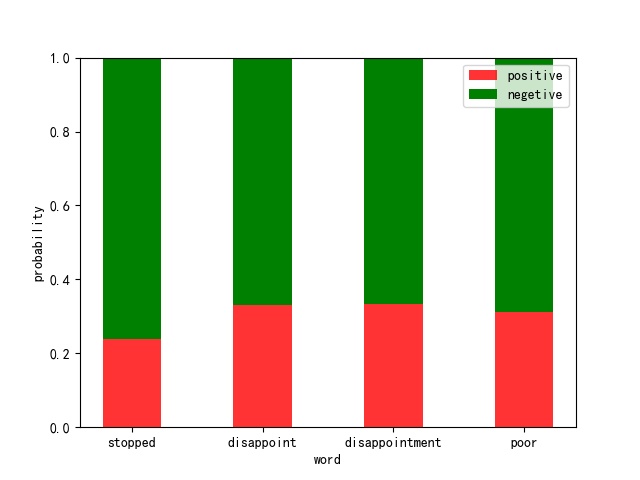
*star\_rating:  
0.4693001370880815  
helpful\_votes  
0.03781104985313568  
review\_scores  
0.49858726073827164*



Through the weight of the three factors, we can get the comprehensive satisfaction score of the products.

#### 6 Relationship between Specific Negative Description and Rating Leveling

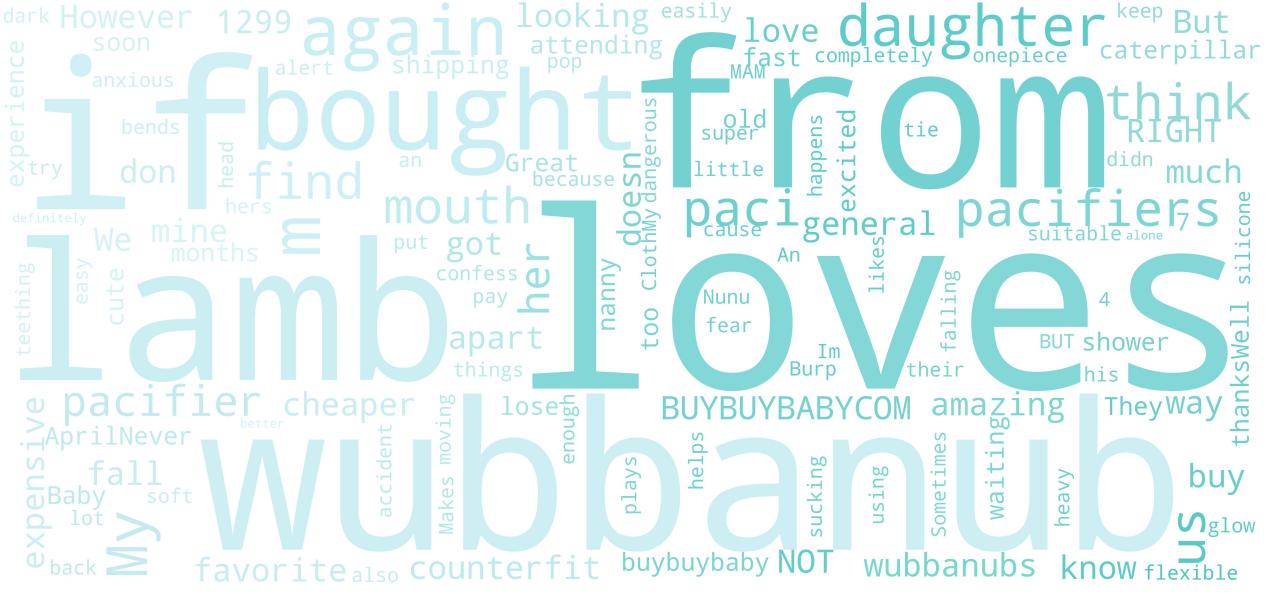
For several words with negative emotions, we counted the probability that these words appear in the reviews with low quality star rating. From the ﬁgure, we can see that these words have a relatively large degree of discrimination in scoring. If these words appear in comments, the corresponding scoring will often be linked to their negativity. However, not all words have similar discrimination, such as some unrelated words (price) or some neutral words (much). These words have no discrimination for scoring, so we think these words have no direct relationship with the rating level.



**7 Wordcloud of the Reviews**







**8 Conclusion**

Our conclusions are as follows:

Star rating and sentiment score of reviews play a decisive role in the overall evaluation of goods.Lower star rating usually relates with lower scores. Watch out for comments with certain negative words, which can lead to lower ratings.

Therefore, we highly suggest that company invite experiencers to experience and evaluate the products after its marketing, which will give consumers more reference information and promote their purchase.