Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2012604

Amazon Review Analysis Based on Comprehensive Product Score

With the development of the internet, online shopping has become a popular trend all over the world. Sunshine company hopes to make an appropriate marketing strategy by analyzing the marketplace. In this paper, we are called to analyze the marketplace and inform the marketing director their online the sales strategy and specific products with potentially important design features.

In task 1, we preprocess the original data set firstly including detecting and deleting abnormal data, and supplementing missing data. Then we further describe the characteristics so as to preliminarily carry out **Visual Analysis** on the data set.

In task 2.a, to identify data measures based on ratings and reviews the reviews, the key is how to quantify reviews. To get the best results, we selected three models for comparison including **FastText Model**, **ID-IDF+ Random Forest Model** and **Unsupervised Sentiment Analysis Model**. The result shows that **the FastText Model** is the most reliable one based on the evaluation index. Then, we explore ways to combine reviews ratings with star ratings and find the appropriate metric named **CPS(Comprehensive Product Score)** for interpreting the data.

In task 2.b, we use the CPS and time of each brand to construct the time series, and then we use ADF to test the stationarity of time series. **ARIMA Model** is what we built to analysis the time series. Result shows that the reputation of hair dryers, microwave ovens and pacifiers are all in a growth trend and pacifiers have the best reputation.

In task 2.c, our goal is to find the combinations of text-based measures and ratings-based measures. We calculate the CPS of each product and then use **NLTK** to make Sentiment Analysis. Based on this, we select two major categories of the products: potential success (CPS>50 and sentiment value>0)and potential failure(CPS<50 and sentiment value<0).And then, we analyze the key words, number of reviews, and star ratings for potential successes and failure. Results show that 4-or 5-star rating with the words in the reviews, such as “great”, “love”, “blow”, “good”, “like”, “best”, “powerful” and more than 46 reviews are the symbol of potential success when referring to a hair dryer.

In task 2.d, we use **bubble color mapped plots**, we select the hair dryer data set and study the variation trend for the number of reviews corresponding to different ratings between 2002 and 2015.We find that customers are more likely to write reviews when they see the 1- or 5- star ratings.

In task 2.e, we build model based on sentiment descriptors and function descriptors respectively. On the one hand, based on **Hierarchical Clustering**, we found that the tendency of sentiment descriptors is positively correlated with the sentiment score. On the other hand, by identifying the most successful potential goods and filtering the description function descriptors, we know that the corresponding product performance of function descriptors is positively correlated with the potential success of the product. Therefore, we demonstrate that specific quality descriptors of text-based reviews strongly associated with rating levels from two aspects.

Catalogue

1 Introduction 2
1.1 Problem Statement 2
1.2 Our Goals 2
2 Assumptions and Notations 2
2.1 Assumptions 2
2.2 Notations 2
3 Data Processing 3
3.1 Data Cleaning 3
3.2 Data Description 3
3.3 Data Visualization 4
4 Task2.a 6
4.1 Processing of reviews data 6
4.2 Model Ⅰ: FastText Model. 6
4.2.1 Select the training data set 7
4.2.2 Quantification of review ratings 7
4.2.3 Experimental process of Scoring 8
4.3 Method 2: TF-IDF + Random Forest 8
4.4 Method 3: Unsupervised Emotion Analysis Model 8
4.5 Comparison of experimental results 9
4.4 CPS (Comprehensive Product Score) 10
The CPS Calculation 10
5 Task2.b 11
5.1 Data Processing 11
5.2 Exploratory Analysis 12
5.3 The ARIMA Model Building 12
6 Task2.c 13
6.1 Calculate the CPI of each product 13
6.2 Sentiment Analysis 14
6.3 Combination Analysis 14
7 Task2.d 15
8 Task 2.e 16
7.1 Model 1: Sentiment Descriptors-based Model 16
7.2 Method 2: Function Descriptors-based Model 19
9 Strengths and Weaknesses 19
9.1 Strengths 19
9.2 Weaknesses 19
10 Summary 20
LETTER 21
Reference 23

1 Introduction

1.1 Problem Statement

In the online marketplace, the market analysis of products, which includes users' star rating of products and reviews on text information, is a very important work. On the one hand, based on user experience, it provides the possibility of market research for other users, and it is beneficial for users to make decisions to some extent. On the other hand, through the analysis of user experience, Marketing Director can timely grasp the focus of such products and then analyze the market situation, so as to make the products have greater economic benefits. Therefore, during the operation of amazon, Marketing Director hope to further understand the market situation by analyzing the market data.

On the platform of sunshine company, there are three online products - a microwave oven, a baby pacifier, and a hair dryer – that hope to form a marketplace analysis letter as well. Asked by sunshine company, we determine a set of goal for analyzing marketplace.

1.2 Our Goals

Based on our understanding of the problem, we set the following goals:

- Task 1: Use the given data to find the distribution of features for different types of goods -- star rating, number of reviews, length of reviews, word distribution of reviews
- Task 2.a: Bulid reviews analysis model, deal with reviews content quantificationally, and get reviews score. Explore ways to combine review ratings with star ratings, and determine the data metrics that best reflect market information.
- Task 2.b: 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.
- Task 2.c: Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product.
- Task 2.d: Visualize the relationship between star ratings and reviews over time. Give a clear conclusion about whether specific star ratings incite more reviews
- Task 2.e: Select specific quality descriptors of text- based reviews and define rating levels. Verify the correlation between specific quality descriptors and rating levels

2 Assumptions and Notations

2.1 Assumptions

- Assumed that the records in the dataset are complete and the data is not missing.
- Assume that there is a correspondence between ratings and reviews

2.2 Notations

Symbol	Definition
CPS	Comprehensive Product Score
S	star rating

Symbol	Definition
R	reviews
CI	Comprehensive Index
C	Clustering cluster
d	Cluster distance

3 Data Processing

In data analysis, there are usually missing values or outliers in the original data set, so we need to preprocess the original data set firstly. In addition, we performed a visual analysis of the pre-processed data to identify and describe the data supplied that will help Sunshine Company succeed in their three new online marketplace product offerings.

3.1 Data Cleaning

In this section, we focus on two types of data: Default Value and Outliers. In one sale record, if there is a large amount of data missing, we will delete this record, because it does nothing to our model and will affect the accuracy of the model; For individual missing data, we use interpolation to complete it. Outliers refer to individual values in the sample whose data deviates significantly from the rest of the observed values. We use box graph to detect outliers and use regression interpolation to deal with them.

3.2 Data Description

To better understand the basic characteristics of the data, we calculated the average, standard deviation, maximum, minimum, median and upper and lower quartile of the three data sets provided, as shown in Tab.1. From Tab.1, we can find that Pacifiers were rated highest and microwaves were lowest. For review ratings, microwave ovens had the most reviews and pacifiers the least.

Tab. 1 Characteristics of the data

Type	Microwave			Hair dryer			Pacifier		
Variables	Star rating	Helpful votes	Total votes	Star rating	Helpful votes	Total votes	Star rating	Helpful votes	Total votes
Count	1615	1615	1615	11470	11470	11470	18939	18393	18393
Mean	3.444	5.6217	6.6694	4.116	2.1791	2.5632	4.305	0.8272	1.131
Std	1.645	27.771	29.262	1.301	14.241	15.382	1.190	5.7356	6.557
Min	1	0	0	1	0	0	1	0	0
25%	2	0	0	4	0	0	4	0	0
50%	4	1	2	5	0	0	5	0	0
75%	5	3	5	5	1	1	5	0	1

Max	5	814	848	5	499	575	5	283	306

3.3 Data Visualization

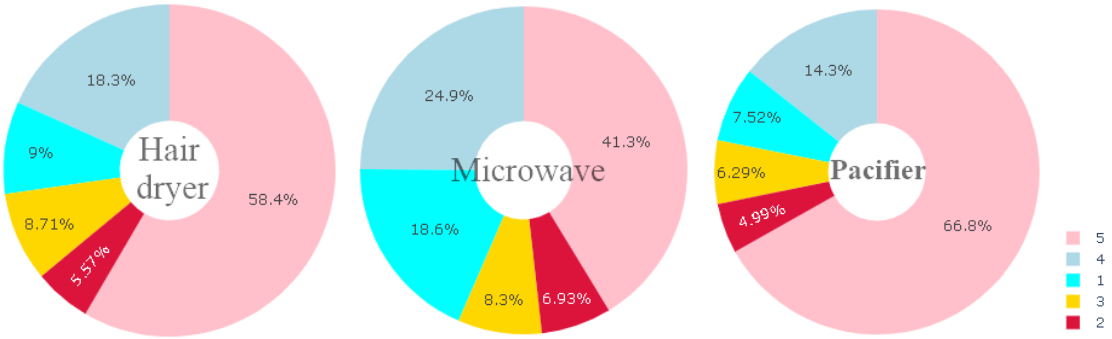

Fig.1 Distribution of star ratings (Hair dryer, Microwave oven, pacifier)

By looking at the above pie charts, we can infer that most of the Ratings are good for the pacifier. Around 66.8% people have given Alexa 5 Star rating, which means the product has given them what they want. The 4-star rating is second only to 5-star rating. The number of 2-star rating is the lowest. Of the three products, the microwave oven has the most negative ratings. About 33.83% people did not like the microwave oven and chose to give only 3-, 2- or 1-star ratings to the microwave oven.

Then, we will present our visualized results with the example of the microwave, and we will find the top 20 reviewed products and then try to analysis reviews of them, including the body length and rating distribution.

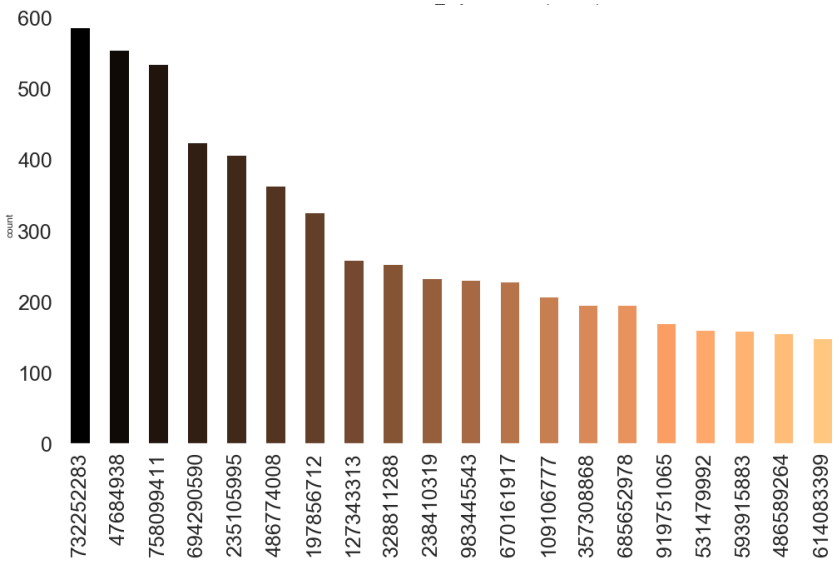

Fig.2 Distribution of review number in kinds(1-20st)

In the Fig.2 above, we list the top 20 reviewed products. These products all received more than 100 reviews. When customers buy products online, they can provide them with more reference information. It is quite clear that product No.732252283 is the most popular variation of hair dryer with near 600 units.

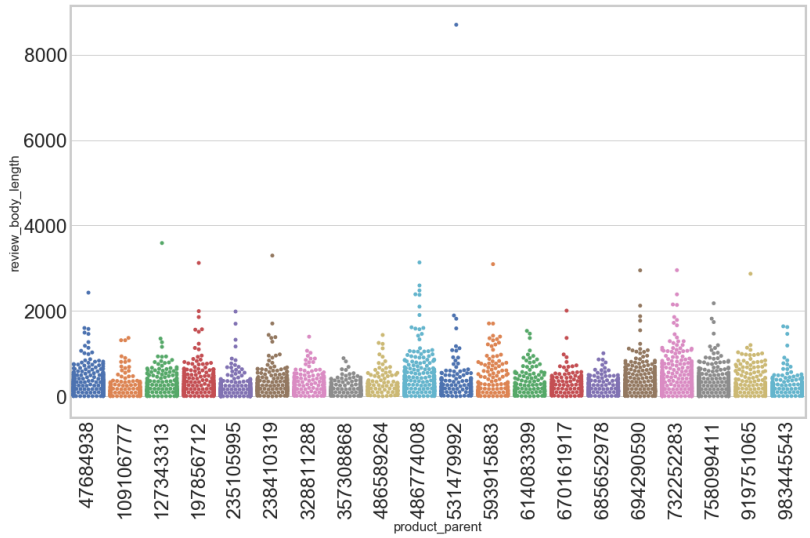

Fig.3 Distribution of Review_body_length in kinds (1-20st)

The above Bivariate plot shows warm plot between kind in hair dryer and length of the reviews. We would like to see for which of the model variations of the hair dryer people have written the longest reviews. By looking at the graph, we can easily spot that the longest review was written for No.486774008. At the same time, the numbers of reviews on Product No.47684938, No.694290590 and No.732252283 are also large. However, there are some review with no reference because they are outliers in this graph, such as No.531479992.

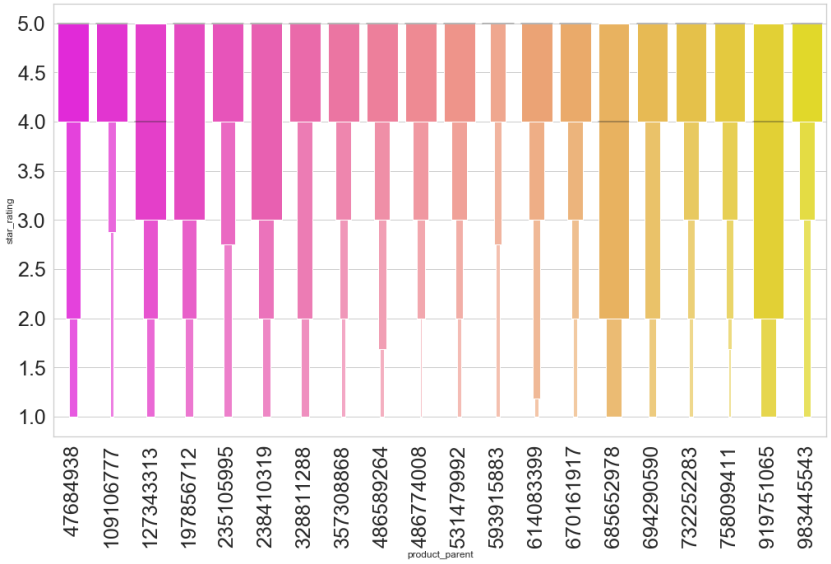

Fig.4 Distribution of star-ratings in kinds(1-20st)

From the graph above, we found that each product contained star rating from 1-star to 5-star, but overall, the number of 5-star ratings was significantly bigger than the number of 1-star ratings. In other words, from positive to negative reviews, the number showed a decreasing trend.

Last but not least, we analysis the word frequency of the reviews. The result (Fig.5) shows that besides the name of the hair dryer, the word “like”,”love”,”good”,”great”,”use”,”cord”,”heat”are

the most frequent words in the reviews so that we can get a rough idea about the reviews and what people think of the product.

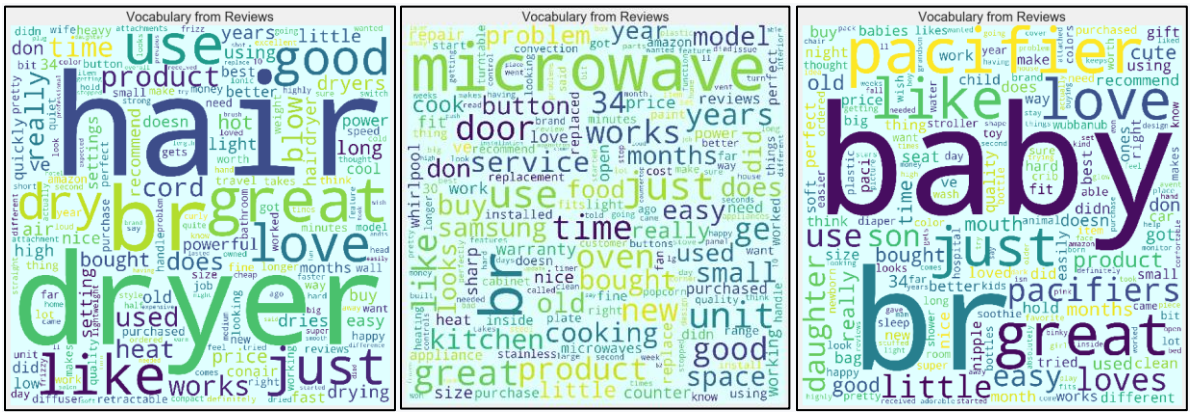

Fig. 5 Distribution of word from reviews(Hair dryer, Microwave oven, pacifier)

4 Task2.a

4.1 Processing of reviews data

1) Identify bulk ratings**;** Observe whether there is a phenomenon of users giving bulk ratings (more than 10), so that speculate whether ratings are genuine.

2) Remove punctuation , auxiliary verbs, blank space and symbols like smiley face; Delete the useless expression of review information, which is the key to reduce the amount of computation, and improve the computational accuracy of the model.

3) Convert all uppercase letters to lowercase letters; The characters are unified to facilitate the subsequent calculation.

4.2 Model I: FastText Model.

First we need to segment the comment data and remove the auxiliary verbs to get a cleaner review data. Then use a simple rule-based sentiment analysis model to define the comments for sentiment analysis. Finally, use the FastText model as a classifier for supervision to enable quantification Comment data.

Step1: Use regular matching expressions, remove all characters except letters, and use stemming or regular word reduction to eliminate or reduce the influence of words due to tense, singular and plural.

Step2: Use the existing open source English auxiliary verb collection of FaceBook to remove auxiliary verbs from the comment data in the previous step.

Step3: Use a simple rule-based sentiment analysis model to combine vocabulary features in reviews to identify sentiment intensity, so as to get the sentiment tendency of the review data

Step4: After processing the above steps, use the FastText model as a classifier to fit the relationship between review and star rating in the data.

Through the Amazon product reviews, classify text information, which can be converted to review score. We research different ways of text categorization, such as Word2vec model and FastText model. The difference between the two models is that Word2vec model enters a group of

words and acquire probabilistic description of text information, according to the classification of the predefined criteria. FastText model introduces n-gram characters, and further develops Hierarchical SoftMax which is used for probabilistic analysis. Compared with Word2vec model [1] the FastText model has higher training speed, occupies less memory space and is more convenient for parameter debugging.

Therefore, we adopted the FastText model to quantitatively score the review information of amazon products. The specific implementation process is shown in the following Fig.6.

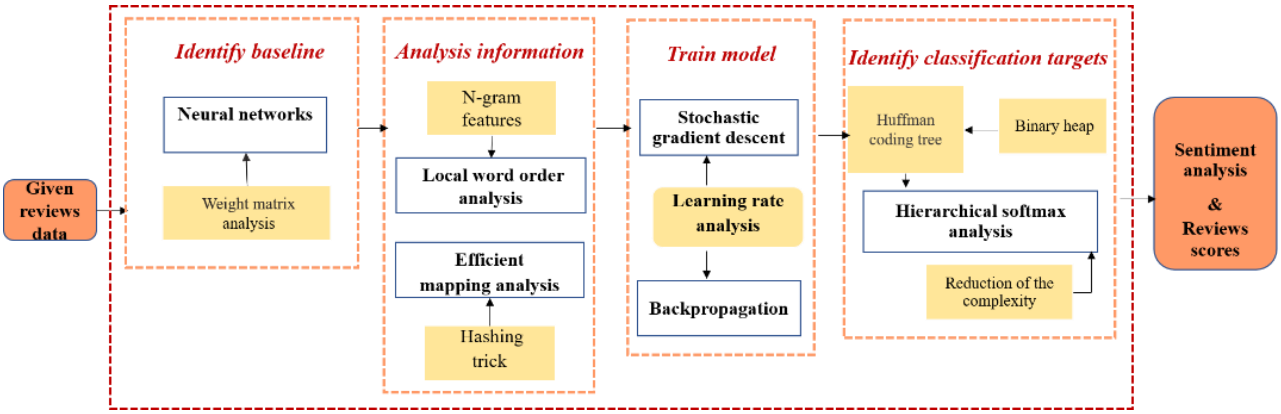

Fig.6 Process of FastText model

4.2.1 Select the training data set

Firstly, we divided the reviews into three categories: reliable reviews, general reviews and unreliable reviews. The specific partition rules are shown in the following Tab.2.

Tab.2 Classification for the reliable type of data

Type	Verified_purchase	The rate of helpfulness rating (>50%)
Reliable	Y	Y
General	Y	N
	N	Y
Unreliable	N	N

Then we use the reliable reviews as the training data set(90% for training and 10% for test) to built the model and use it to the general reviews to get its quantized value.

4.2.2 Quantification of review ratings

In the above process, the step Hierarchical SoftMax process uses the combination of Huffman coding tree and binary heap algorithm to vectorize different review information. In this process, the vectors of different categories are respectively analyzed by probability distribution to determine the classification target. Specifically, in the training process, when different word vectors pass through non-leaf nodes, the probability of left fork inclination or right fork inclination should be judged respectively to determine the probability distribution of the target word vector. Successfully, the multi-classification problem is converted into a binary classification problem, and the complexity of the simplified operation was reduced from O(Kd) to O($dlog_2(K)$), where

K is targets Numbers and d is hidden layer dimensions of neural networks[2].

In the process of classification, we will use the retention method to divide the word vector of amazon product review into training set and test set. By modeling the training set and predicting the test set, the feasibility of the model can be measured to some extent. The indexes of Testing Accuracy, precision, recall and f1-score were used to quantitatively evaluate the model. Based on the four indexes, the parameter values involved in the algorithm are set, and the model is tested and trained to demonstrate the rationality of the algorithm for the product review data set.

Finally, the probability distribution of the word vector data set is normalized, and the delimited interval is discretized, that is, different scores are set for product reviews. According to the weight analysis of probability distribution, the specific score value is obtained.

4.2.3 Experimental process of Scoring

We calibrate some parameters in the algorithm, and the specific parameter values are shown in the following Tab.3.

Tab.3 Specific parameter values

parameter	learning rate	epochs	negative simples	n-grams	context window
value	0.05	5	5	1	5

Among these parameters, we need to emphasize the necessity of negative sampling. Every time, one training sample is allowed to update a part of the weight to reduce the calculation amount in the process of gradient descent. Through experiments, the test time and algorithm performance are compared. Generally, the number of negative samples is set to 5.

4.3 Method 2: TF-IDF + Random Forest

The following Fig.7 is a demonstration of the specific solution method of the bag of words model.

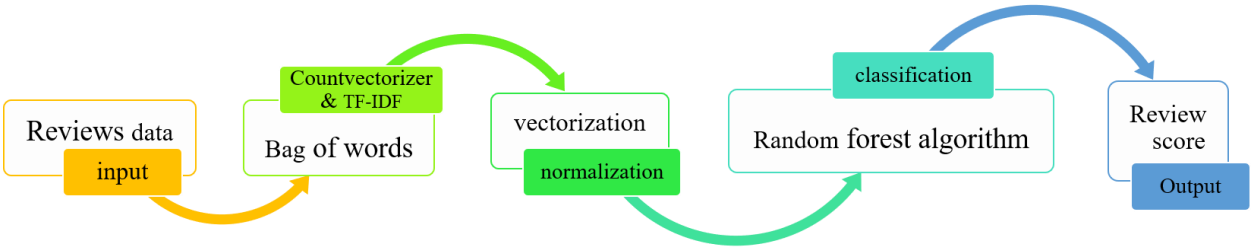

Fig.7 Process of bag of words model

The basic idea of bag of words model is to extract keywords from the review text of amazon products and count the word frequency. Furthermore, the word frequency is converted to the form of vector, and the vector is normalized and expressed as a set of vectors based on reviews.

4.4 Method 3: Unsupervised Emotion Analysis Model

In the process of language information dealing for amazon product review, both stemming both and lemmatization[3] are usually adopted. In terms of complexity, lemmatization method is relatively complex compared with the stemming process. However, the results obtained after the

lemmatization process are of certain significance, which can be converted to complete words. Therefore, we process the review information based on the lemmatization.

The following is a demonstration of the specific solving method of the lemmatization emotion analysis model.

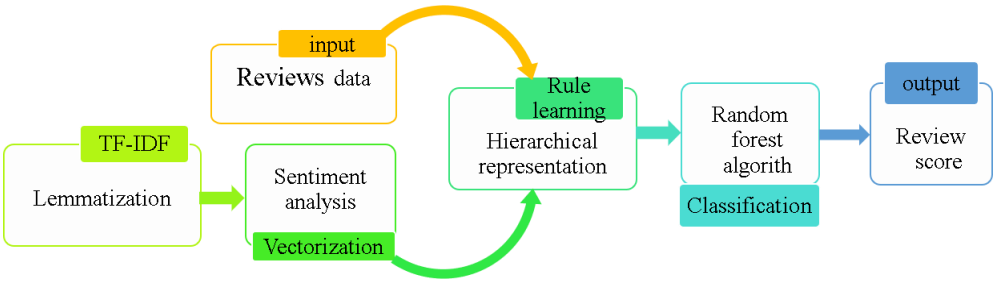

Fig.8 Process of the emotion analysis model based on lemmatization.

Step 1, conduct character processing on the review content, and then encode the extracted words based on the lemmatization model. The purpose of coding is to extract keywords (unsupervised learning).

Step 2, put the reviews into an emotion analyzer and score them according to the emotion level of the keywords. After the score, the emotional tendency of a word or sentence can be obtained.

Roughly speaking, this process is based on the different feelings of words or sentences, which can then lead to different ratings or inclinations of reviews. The keyword extraction involved is subjective to some extent, and the selected keywords can be either with high frequency of occurrence, or with high score through an evaluation method.

In the above white-box testing process, the method adopted for the classification of the two modes is Random forest algorithm[4]. The specific parameters we set here include Max_depth (the maximum depth of the decision tree), Min_samples_split (the value of the minimum number of leaves), and N_estimators (the number of decision trees), as shown in the following Tab.4.

Tab.4 Specific parameter values

parameter	Max_depth	Min_samples_split	N_estimators
Value of method one	100	8	100
Value of method two	90	2	50

4.5 Comparison of experimental results

Here, we use all the data dry data for experiment. In the process of explaining the rationality of the algorithm, we rely on the three models discussed above. Taking the hair dryer as an example, the data set is divided into training set and test set, and the ratio of the training set is 9:1. For multi-classification, there have many methods to calculate its metrics. We select one-vs-reset kind to realize the multi-classification using traditional classification. So we use mean to calculate the four metrics. The results obtained by measuring the four indexes are as followed.

Tab. 5 Comparison of experimental results for different models

The Evaluation Index / Model	Testing Accuracy	Precision	Recall	f1-score

FastText	0.7054	0.7054	0.7054	0.7054
ID-IDF+Random Forest	0.5652	0.5652	0.5652	0.5652
Unsupervised Emotion Analysis Model	0.6298	0.7298	0.6298	0.6298

After a comprehensive comparison of the above three models, we chose the FastText model as the basis for determining the evaluation score due to the four evaluation index.

The reviews were rated on a data set of hair dryers, pacifiers and microwave ovens. In three data sets. We randomly selected two reviews for rating, and the rating results are shown below.

Tab.6 The rating results associating with review

	review_id	review_body	review_rating
Hair dryer	R24QP0W6EEOHUS	Seems to be working well, was easy to mount on the wall.	3
	RVQC05SNQCV7U	It broke the third time I used it piece of crap	1
Pacifier	R2M4WTN0Z3EJZ5	went well with the baby one-piece tuxedo. excellent	5
	R1JMEC4M3M2GJE	I wish I could've picked the design.	4
Microwave	R2X9MN0B0MVUZT	Works Really Well	5
	R3UMLQR0WFBJ11	Doesn't work on a pure sine wave inverter.	2

4.4 CPS (Comprehensive Product Score)

S(star rating), R(reviews) , C(Comprehensive) and CI(Comprehensive Index) are all column vector, which is an n × 1 matrix.

$$C_n = \alpha \times S_n + \beta \times R_n \quad (1)$$

Then we find α as 0.3 and β as 0.7 based on Factor Analysis. When we analyzed the distribution of comment values of some brands, we found some anomalies.We judge that although the calculation results of factor analysis show that comments account for a large weight in the dimensionality reduction process, there may be a phenomenon of malicious review in the actual situation. Therefore, we manually modified the results of factor analysis:

$$CI_n = f(C_n) \quad (2)$$

f is the operator that we defined; it is used to rounding to discretize the data into integers.

We define a variety named CPS referring to NPS, which is an indicator built by Amazon. NPS measures customer experience and predicts business growth. This proven metric transformed the business world and now provides the core measurement for customer experience management programs the world round[5].

The CPS Calculation

Calculate the CPS using the index CI that we built above. It has combined the star rating, reviews and helpfulness rating. Its biggest advantage is that it improves the problem of only considering users' ratings and ignoring the information contained in users' reviews. We define three types based on CI. They are Promoters, Passives and Detractors.

The users are grouped as follows based on CI.

Promoters (score 5): They are fanatical loyalists who will continue to buy and introduce the product to others.

Passives(score 4): They are used to dealing with your company and are satisfied, but have no enthusiastic recommendation. They are easily attracted by other competitors.

Detractors(score 1-3): They are not satisfactory or have no loyalty to your company. They have a mediocre or even bad relationship with your company, and most of the bad reviews comes from them.

$$\mathrm{CPS} = \frac{(Promoters - Detractors)}{Total\ users} \times 100 \tag{3}$$

Then we find α as 0.3 and β as 0.7 based on Factor Analysis. When we analyzed the distribution of comment values of some brands, we found some outliers (Fig.9). The graph on the left shows the R of users who did not buy, while the right one shows R of users who did buy. There is a significant difference in the distribution of the two graphs. By comparing the overall trend of the data set analyzed in the first part, we judge that although the calculation results of factor analysis show that comments account for a large weight in the dimensionality reduction process, there may be a phenomenon of malicious review in the actual situation. Therefore, we manually modified the results of factor analysis: α=0.4，β=0.6.

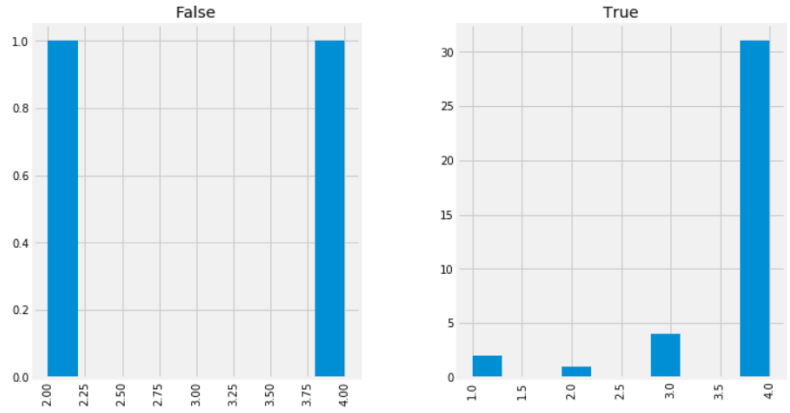

Fig.9 Distribution of star rating by non-purchaser and vice-purchaser

Taking hair dryers as an example, we calculated the CPS of vice-purchaser and non-purchaser. The results showed that the CPS of vice-purchaser (64.55) was greater than that of non-purchaser(37.00), indicating that the higher the CPS is, the higher the sales volume will be.

5 Task2.b

In this part, we try to identify and discuss time-based measures and patterns within each data set. The model we built is Time Series Analysis.

5.1 Data Processing

Since we need to observe the trend of the whole data set and each brand, while products with few purchase records do not have time series, we need to make descriptive statistics on the distribution of purchase records of brands and choose the data with time series. By looking at the distribution of review for each product, we can see that there is a turning point at about 10. Therefore, we removed products with less than 10 reviews, meaning that they didn't participate in the time series analysis.

5.2 Exploratory Analysis

To make time series analysis, we need to test the stationarity of time series first. So, we choose ADF test to explore it. The ADF test expands the Dickey-Fuller test equation to include high order regressive process in the model[6].The table followed shows the original and final value of ADF before and after the difference.

Tab.7 The value of ADF in the three products

ADF	Hair dryer	Pacifier	Microwave oven
Original	1.03(-2.64)	-0.72(-1.53)	-1.12(-3.48)
Final	-2.01(-1.97)	-1.64(-0.72)	-11.6(-3.48)

*The Numbers in bracket represent the threshold under the 5% confidence interval

From the table, we can find that though the time series are not stationary originally (>5%), it changes into a stationary time series after difference (<5%). Based on the exploratory analysis, we can build the ARIMA Model to further investigate pattern of the time series.

5.3 The ARIMA Model Building

Since our data does not have a stable time series, we choose the ARIMA model to carry out time series analysis. It is a model which transforms the non-stationary time series into stationary time series and then regresses the dependent variable to its lag value and the present value and lag value of the random identifier.

Given a time series data X_t where *t* is an integer index and the X_t are real numbers(CPS), an ARMA(p',q) model is given by

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + \delta + \mu_t + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \cdots + \theta_q \mu_{t-q} \quad (4)$$

where the $\emptyset_i$ are the parameters of the autoregressive part of the model, the $\theta_i$are the parameters of the moving average part and the μ_t are error terms. The error terms are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean[7].

$$\Delta X_t = X_t - X_{t-1} = X_t - LX_t = (1-L)X_t \quad (5)$$

$$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1} = (1-L)X_t - (1-L)X_t = (1-L)^2 X_t \quad (6)$$

$$\Delta^d X_t = (1-L)^d X_t \quad (7)$$

where d is the degree of differencing and L is the lag operator.

$$w_t = \Delta^d X_t = (1-L)^d X_t \quad (8)$$

where w_t is a stationary series.

$$w_t = \emptyset_1 w_{t-1} + \emptyset_2 w_{t-2} + \cdots + \emptyset_p w_{t-p} + \delta + \mu_t + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \cdots + \theta_q \mu_{t-q} \quad (9)$$

The result of the model is as followed(Fig.9).We can easily find that the reputation of the pacifier is relatively stable and has been maintained at a high level, while the reputation of the microwave oven fluctuates greatly and remains at the lowest level. From the time series analysis, it can be found that the reputation of the three products in the future shows a trend of gradual growth and the growth rate of hair dryer is the fastest.

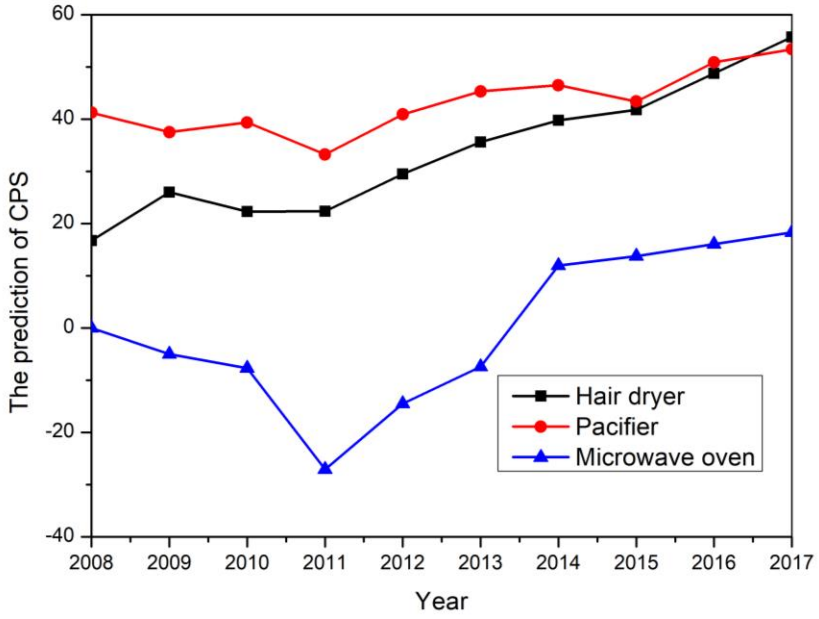

Fig.9 The prediction of CPS from 2008 to 2017

6 Task2.c

In this part, we try to determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product. Our main idea is shown as the followed Fig.10.

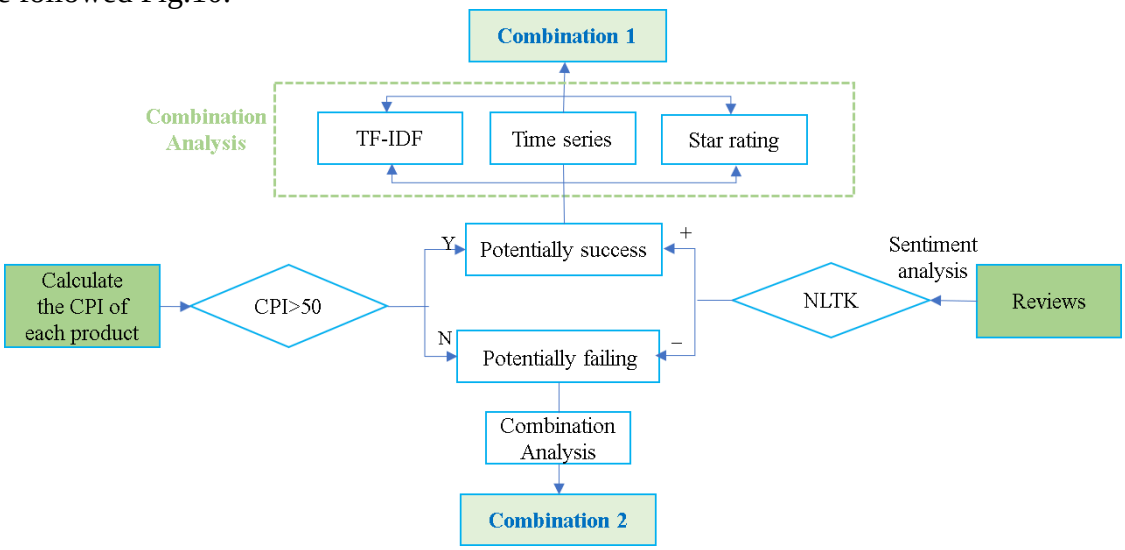

Fig.10 Flow chart of Task2.c

6.1 Calculate the CPI of each product

To analysis the factors of potential success or failure, we should identify the successful products and failing products from others. CPI is an index that we defined in 4.3, which is a

symbol based on ratings and reviews that are most informative for a company to track. So, we calculate the CPI of each product and select the ones whose CPI>50 firstly.

6.2 Sentiment Analysis

In this part, we use NLTK(Natural Language Toolkit), which has been called “an amazing library to play with natural language”, to make sentiment analysis. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum[8] .It not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

We used python to conduct sentiment analysis on the reviews of each product. Combined with the CPI calculated in the previous part, products with CPI>50 and positive sentiment analysis were classified as products with potential to success, while brands with CPI<50 and negative sentiment analysis were classified as brands with potential failure.

6.3 Combination Analysis

After classifying the commodities, we used TF-IDF to extract the keywords in the reviews of successful and failed products respectively, and the principle of which has been described in section 4.3.The result of hair dryers can been seen in Fig.11 and Fig.12.

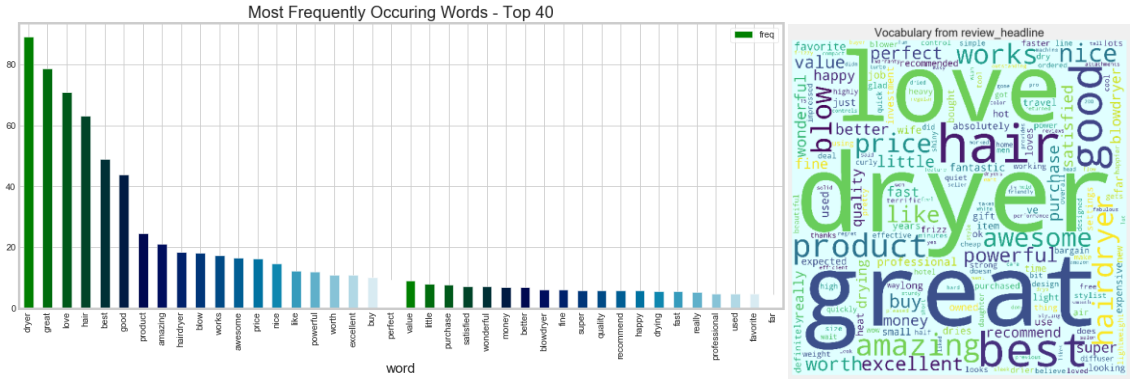

Fig.11 The most frequency words in the headline of the reviews

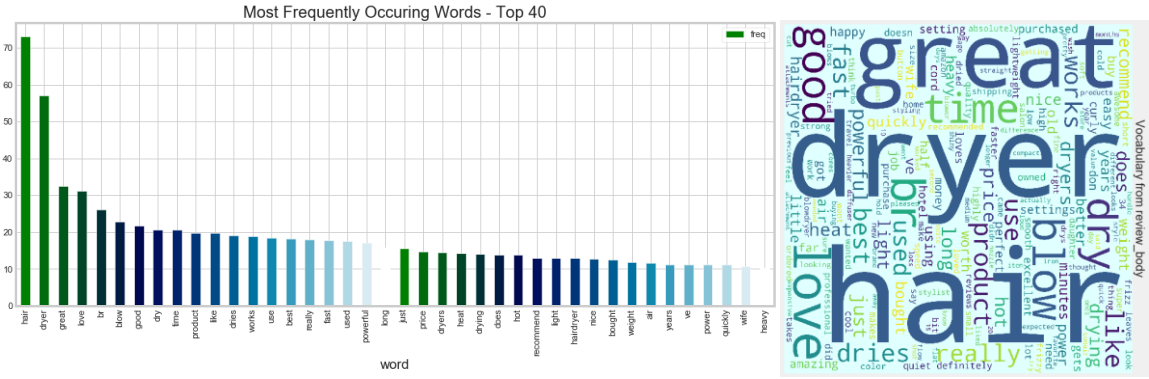

Fig.12 The most frequency words in the headline of the reviews

Besides, we calculated the averages of ratings and reviews for successful products, which were 4.5 and 46 respectively. Based on the investigation, we can find that 4-or 5-star rating with the words in the reviews, such as “great”, “love”, “blow”, “good”, “product”, “like”, “works”, “best”, “powerful” and more than 46 reviews are the symbol of potential success when referring to a hair dryer.

The symbol of potential failure can be calculate in the same way above. The result is that When the hair dryer has 3-,2- or 1-star rating with the words in the reviews, such as “cord”, “stopped”, “retractable”, “dry”, “dangerous”, “hazard”, “mistake”, “poor”, “defective” and less than 32reviews, it may lead to potential failure.

Finally, we make a summary of the result of the microwave oven and the pacifier in the Tab.8 followed.

Tab.8 The result of the Microwave oven and the Pacifier

Product	Potential Type	Keywords in Reviews	Star rating	Number of the review
Microwave Oven	Success	“paint”, “product”, “worked”, “great”, “like”	5	>73
	Failure	“just”, “repair”, “warranty”, “poor”, “bad”, “quality”, “broke”, “terrible”, “worst”, “died”, “damaged”	1,2	<32
Pacifier	Success	“baby”, “loves”, “love”, “cute”, “great”, “son”	5	>87
	Failure	“medicine”, “hard”, “son”, “just”, “mouth”, “nipple”, “bad”, “disappointing”, “wrong”, “poor”	1,2,3	<39

7 Task2.d

To visualize the relationship between star rating and the number of reviews over time. By using bubble color mapped plots, we select the hair dryer data set and study the variation trend for the number of reviews corresponding to different ratings between 2002 and 2015.

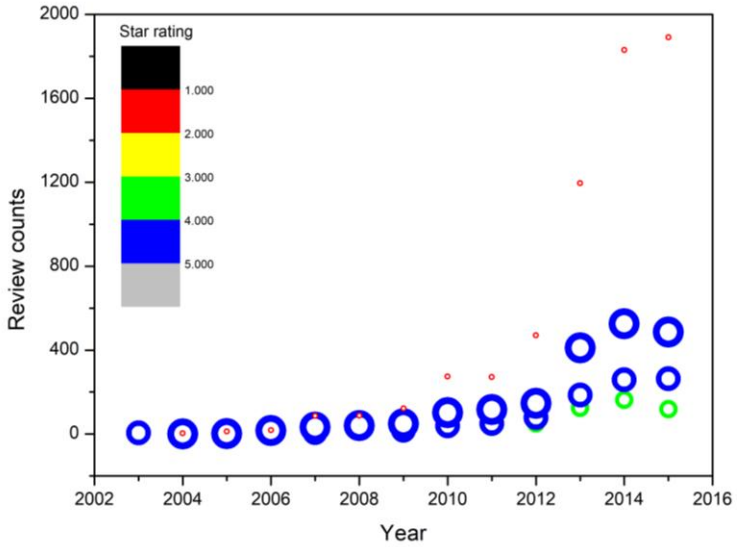

Fig.13 The relationship between review number and star ratings over time

In this Fig.13, different colors correspond to different ratings, and the size of bubbles represents the level of star rating. The trend of the bubbles over time indicates the variation trend of reviews number over time. The relationship between the size of the bubble and the trajectory of the bubble shows the relationship between the star rating and the number of reviews over time.

As can be seen from the Fig., the number of reviews corresponding to different star scores fluctuates with the change of time. Overall, the number of reviews for different star ratings is increased over time, suggesting that online sales are gaining popularity. Therefore, it is particularly important to analyze the online marketplace.

We compare the growth trend of different star ratings, and find that the number of reviews increases fastest with the change of time when the star rating was one. For star rating of one, the growth rate is relatively gentle from 2004 to 2009, while the number of reviews increased sharply from 2010 to 2015. It indicates that customers are most likely to write some type of review after seeing a series of one-star ratings. For five-star rating, the growth of reviews number is slower than one-star rating, but still faster than that comparing to other star ratings. It also suggests that customers are more likely to write reviews when they see the highest ratings. That is to say, specific star ratings incite more reviews.

8 Task 2.e

In determining the CPS, we assume that there is a close relationship between the review text information and the rating level, but we did not provide the corresponding argument. Therefore, we further extract the text information, analyze the degree of correlation between the text information and the score level, and try to verify the rationality of the above hypothesis.

There are two ways to describe quality descriptors of text-based reviews, which use sentiment words and function words as quality descriptors.

- **Method 1:** Based on the cluster analysis[9] model of unsupervised learning and training, we extracted representative words from the clustering results. The emotion of the output words is analyzed and the emotional characteristics of the words are divided. Then, a simple rule learning model is used to calculate the emotional score of such words. When we consider the emotion score as rating levels, if the emotional words show a positive side, they also have a higher emotional score. We verify that there is a close correlation between the words describing the emotion color in the review text message and rating levels.
- **Method 2**: we aimed at the above bulided potentially successful or failing of the product as a result, select the results corresponding to the product category. Further, process text information of the product category reviews. If the content of the extract is a strong characteristic of function word, which is closely linked to potentially successful or failing product, it shows the functional word and rating levels are closely linked.

The following is a detailed discussion of the two methods

7.1 Model 1：Sentiment Descriptors-based Model

Firstly, we extract the word vector based on the FastText model and conduct clustering analysis on the word vector. The clustering method we use here is hierarchical clustering.

Taking the data set of hair dryer as an example, we classified 3 clusters by using hierarchical

clustering[10]. Note that we are clustering the entire dataset to make the results more visual and interpretable. Our results are shown in the following figure.

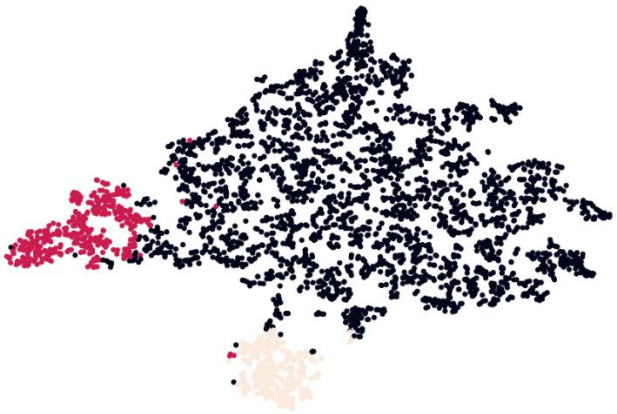

Fig.14 Result of hierarchical clustering

It can be seen from the figure, for hierarchical clustering, when the number of clustering clusters is 3, the clustering effect is very obvious, which indicates that the algorithm has a good effect on the data sets of products. In fact, if the whole data set is selected for clustering analysis to extract emotion words, the amount of data to be processed will be very large, so we only select the data set corresponding to the product with the most reviews in the hair dryer. Here, we select the hair dryer whose product_parent is 959834931 and conduct cluster analysis on it.

Hierarchical clustering regards each sample of the product data as a cluster. In each step, find the two clusters closest to each other, merge them and repeat until the preset number of cluster clusters is reached. For the given cluster C_i and C_j, the expression for calculating the distance[11] is:

$$d_{min}(C_i, C_j) = min\ \ dist(x, z), x \in C_i, z \in C_j \tag{10}$$

$$d_{max}(C_i, C_j) = \max\ dist(x, z),\ x \in C_i, z \in C_j \tag{11}$$

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|}\sum_{x \in C_i}\sum_{z \epsilon C_j} dist(x, z) \tag{12}$$

Here, we choose d_{max} as the cluster distance, which involves complete-linkage algorithm. The samples processed by FastText model are encoded to generate a tree graph as shown in the following Fig.15.

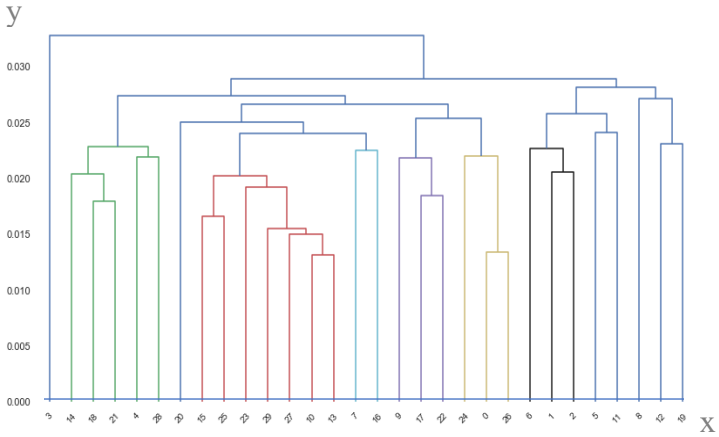

Fig.15 Distribution of Cluster distance (y) with sample number (x)

In the FastText model above, we extract the vocabulary of review information. For example, for a hair dryer with product_parent of 959834931, we first extract a series of terms for hierarchical clustering analysis. Then we used t-SNE to reduce the dimension of the high-dimensional data, so as to map the data to the floor plan.

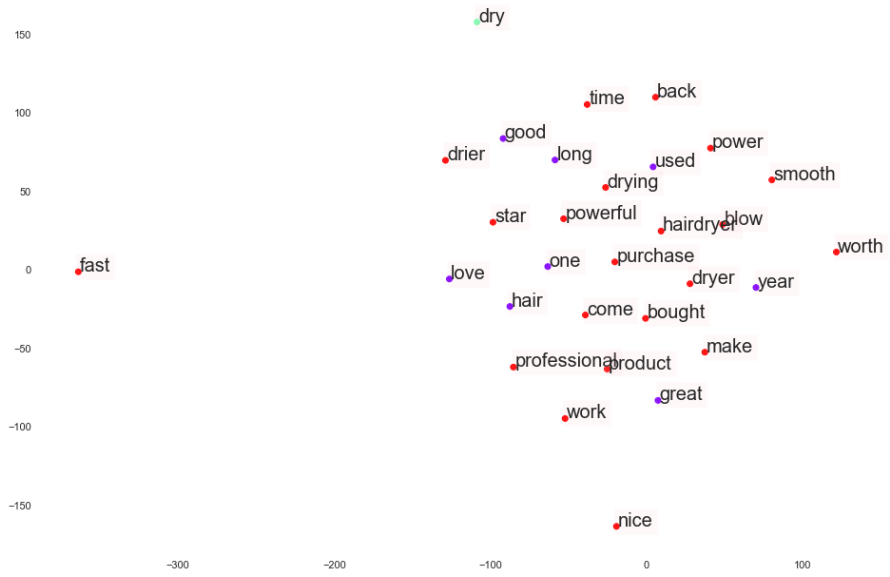

Fig.16 The mapping of words

The color in the picture is the division of the types of words. From the observation, we can see that the words can be divided into three groups. We choose a group of words with more emotional colors. That is the set of words represented by the purple in the Fig.16.

To portray rating levels, we think of emotional scores as rating levels. To be specific, after selecting the words, a simple rule-based emotion analysis model is selected. The emotion score of these words is calculated and regarded as rating levels. The results are shown in the following table.

Tab.9 Sentiment score of word

Word	Sentiment score
hair	0
great	0.6249
love	0.6369
year	0
used	0
one	0
long	0
good	0.4404

From the table, it can be seen that the emotional words "great", "love", "good" and other words with positive emotional colors also correspond to high emotional scores, which proves that the emotional tendency of the emotional words is closely related to the rating level.

7.2 Method 2: Function Descriptors-based Model

For hair dryer products, we first select the products according to the number of reviews on the product category. Then, according to the size of CPS, the product with the most potential for success is selected, 959834931. Word frequency statistics are conducted on its review content. The results are shown in the following table.

Tab.10 Word frequency statistics

word	frequency	word	frequency	word	frequency
dryer	67.08%	great	53.51%	love	52.01%
hair	47.07%	best	35.66%	good	27.58%
amazing	15.58%	hairdryer	15.49%	awesome	13.36%
price	12.34%	blow	12.24%	works	12.07%
nice	11.54%	product	10.48%	worth	10.11%
buy	7.48%	excellent	6.59%	better	5.8%

It can be seen from the table that the keywords “powerful“ and “works” exist, and these keywords reflect the specific performance of the product to some extent. The appearance of these properties makes the product have more development potential. If we take feature words as quality descriptors and rate levels as to whether a product has potential for success. In the data set of this product, the extracted content is characteristic function words, which have obvious performance tendency, and this product has the most potential for success. Therefore, it verifies whether the performance tendency of function words is closely related to the rating level.

9 Strengths and Weaknesses

9.1 Strengths

- Low Memory and Evaluation Complexity: We use pretrain model and only using the higher CPS can stand for the better prospect.
- Run Fast and Deploy Easy: We run as the speed of the traditional machine learning the Deep Learning Model and programs are easy.
- High Generalizability: We use three metircs to evaluate the model and we build the model in the three datasets.
- High Accuracy and Adaptive Prediction:We use word-vector and deep-network,what is more,we train the model with a lot data.
- Quantified and Rational Goals:We set quantified goals based on text-analysis.

9.2 Weaknesses

- No Verification of Raw Data:We have no guarantee of the accuracy of given data.
- No Complex Time Series:Our time-series only predict in short time .
- No Cosideration of product Brand:We have not many the models about different brand prod uct.

10 Summary

In this paper, we build different models to analyze the market situation.

- **T**he FastText model based on evaluation index is the most reliable. The CPS is a key metric to synthesize the quantitative review score and the star rating.
- **T**he public praise of brands such as hair dryers, microwave ovens and pacifiers are on the rise and pacifiers have the best reputation.
- 4-or 5-star rating with the words in the reviews, such as “great”, “love” “, “blow”, “good”, “like”, “best”, “powerful” and more than 46 reviews are the symbol of potential success when referring to a hair dryer.
- **W**hen customers see a 1-star or 5-star rating, they are more likely to write a review.
- **S**pecific quality descriptors like sentiment descriptors and function descriptors from text-based reviews strongly associated with rating levels.

LETTER

DEAR Sunshine Company，

It is my pleasure to give policy recommendation to Sunshine Company. We have deeply analysis the data that you provided by building models, and what we what to tell you now is some online sales strategies that we have found in our investigation.

Due to the virtual nature of online shopping, it is difficult for users to have a comprehensive and objective understanding of products, so other users' description of goods will have a greater impact on the sales of goods. Through the data analysis of existing brands, we found that, on the whole, the current users' rating of goods is much higher than the negative rating, but often the negative rating will have a greater impact on users' buying behavior. Through box chart analysis, we found that users tend to write comments after seeing comments with low or high scores, and comments with low scores have a greater impact on users' comments. Through cluster analysis, we find that the characteristic keywords in user comments are often related to certain star ratings. Besides, There are some combinations of text-based measures and ratings-based measures that best indicate a potentially successful or failing product.

Currently, Amazon's official overall rating for a product only considers users' ratings. But in practice, users tend to pay more attention to the content of user reviews. Therefore, we quantified the comment content with FastText and bulided the OPI index in combination with the rating. The value of this index ranges from -100 to 100, and the bigger the value is, the higher the overall evaluation level of the product will be. Based on this index, we used ARIMA time series analysis to predict the changes in the reputation of hair dryers, pacifiers and microwave ovens in the next two years. The results showed that the reputation of all three products in the next three years was in a good direction, and the reputation of pacifiers was the highest.

In accordance with the above research idea, we'd like to give you some suggestions referring to your online sales.

- You can use the OPS to gain insight into the changes of the market as a whole as well as other brands, and it can also a good way to predict the future sales trend.
- In the early days of online sales, you need to pay attention to the combination of ratings, comment keywords, and comments, because we found that these combinations could indicate potential success and failure by modeling portfolio analysis. The combination of the three products you are interested in is shown in the following table.

Product	Potential Type	Keywords in Reviews	Star rating	Number of the review
Microwave Oven	Success	"paint", "product", "worked", "great", "like"	5	>73
	Failure	"just", "repair", "warranty", "poor", "bad", "quality", "broke", "terrible", "worst", "died", "damaged"	1,2	<32
Pacifier	Success	"baby", "loves", "love", "cute", "great", "son"	5	>87

	Failure	“medicine”, “hard”, “son”, “just”, “mouth”, “nipple”, “bad”, “disappointing”, “wrong”, “poor”	1,2,3	<39
Hair dryer	Success	“great”, “love”, “blow”, “good”, “product”, “like”, “works”, “best”, “powerful”	5	>46
	Failure	“cord”, “stopped”, “retractable”, “dry”, “dangerous”, “hazard”, “mistake”, “poor”, “defective”	1,2,3	<32

- Through the analysis of the more successful types of existing brands, we find that there are some potentially important design features that would enhance product desirability. For a hair dryer, these features are “turbo power”, “coldmatic hair dryer” and “2000 watts.” For a microwave, they are “cavity paint 98qbp0302”.And for a pacifier, these features are “wubbanub” and “animal”.

We sincerely hope everything goes well with your company. We are very eager to hear your opinion on our recommendation and to have more communication about it. Thank you for taking the time to read our letter.

Best Regards,

Sincerely Team 2012604.

Reference

[1]Tang, Duyu, Bing Qin, and Ting Liu. "Document modeling with gated recurrent neural netwo rk for sentiment classification." Proceedings of the 2015 conference on empirical methods in nat ural language processing. 2015.
[2] Joulin, Armand, et al. "Bag of tricks for efficient text classification." arXiv preprint arXiv:16 07.01759 (2016).
[3]Agnew, William, and Pedro Domingos. "Self-Supervised Object-
Level Deep Reinforcement Learning." arXiv preprint arXiv:2003.01384 (2020).
[4]Straka, Milan, Jana Straková, and Jan Hajič. "Evaluating Contextualized Embeddings on 54 L anguages in POS Tagging, Lemmatization and Dependency Parsing." arXiv preprint arXiv:1908. 07448 (2019).
[5] https://www.netpromoter.com/know/
[6] Alastair Hall (1994) Testing for a Unit Root in Time Series With Pretest Data-Based Model Selection, Journal of Business & Economic Statistics, 12:4, 461-470
[7] J. Contreras, R. Espinola, F. J. Nogales and A. J. Conejo, "ARIMA models to predict next-day electricity prices," in *IEEE Transactions on Power Systems*, vol. 18, no. 3, pp. 1014-1020, Aug. 2003.
[8] https://www.nltk.org/
[9]Morton, Craig, Jillian Anable, and John D. Nelson. "Consumer structure in the emerging mark et for electric vehicles: Identifying market segments using cluster analysis." International Journal of Sustainable Transportation 11.6 (2017): 443-459.
[10]Bongiorno, Christian, Salvatore Miccichè, and Rosario N. Mantegna. "Nested partitions fro m hierarchical clustering statistical validation." arXiv preprint arXiv:1906.06908 (2019).
[11] Liu, An-An, et al. "Hierarchical clustering multi- task learning for joint human action grouping and recognition." IEEE transactions on pattern analysis and machine intelligence 39.1 (2016): 102-114.