For office use only	Team Control Number	For office use only
T1 ________	**2020042**	F1 ________
T2 ________		F2 ________
T3 ________	Problem Chosen	F3 ________
T4 ________	**C**	F4 ________

2020
MCM/ICM
Summary Sheet

Commercial Opportunities beneath Reviews

Summary

In order to choose the most appropriate online sales strategy and find potentially important design features, our team makes an analysis on the data and find the hidden relationship between several measures, especially based on time. We completed the following works to provide Sunshine Company a comprehensive profile to make business decisions.

Firstly, we preprocess the data, including missing value imputation, data cleaning, abnormal characters transformation. Based on our dataset partition and review separation, we separately use Naïve Bayes Classifier and LSTM to achieve the sentiment analysis, which provide us a relatively objective sentiment evaluation based on reviews.

Next, we construct the AHP-FCE model to reflect the reputation of product in a period of time. We classify different indicators and divide them into three levels. By using FCE model, we obtain a preliminary evaluation of products reputation.

Moreover, we further propose the Multivariate Nonlinear Regression Model based on market factors. We target the number of reviews and construct a function related with several factors, based on the assumption that the number of reviews is positively correlated with the purchase.

We consider the potential influence of history reviews on future reviews and use the simplified time series. In order to fit the data better, we also introduce the number of effective fans and the occurrence of market events to the model and obtain the relationship between several factors.

According to this, we can provide a more appropriate online sales strategy for the company.

Finally, we analyze potentially important features of specific products, and find the most influential words to evaluate the quality of a product.

Keywords: Sentiment Analysis; Market Factor

Contents

1 Introduction

1.1 Problem Restatement

With the rapid development of technology and logistics, there is an increasing number of people choosing shopping online. While the customers open the parcel and enjoy the shopping, the sellers receive feedbacks based on their reviews and star ratings, which express the customers level of satisfaction. Moreover, other customers can also submit ratings on these reviews as being helpful or not.

Firstly, we programmed data processing and made a comprehensive analysis on the patterns of reviews. After that, we obtained a relatively objective rating of sentiment based on the reviews.

Secondly, we synthesized several elements, including sentiment, star ratings, proportion of helpful votes, and constructed the reputation point to reflect publics synthetic preference of specific products.

Finally, based on the reputation point we gave, we analyzed the hidden relationship between peoples expectation, the potential of specific products, and time.

2 Assumption

1. **Balance Supply and Demand Hypothesis**
   We assumed that supply and demand on American market are balanced on the whole.

2. **Consumer Preference Hypothesis**
   Consumer Preference Hypothesis. Consumer preference of a specific product does not change in a period of time.

3. **Authentic Data Hypothesis**
   We assume all the data provided is authentic and original.

3 Data Preprocessing

1. Select the target products. We only need to focus on hair dryer, pacifier and microwave, but the data provided includes several products other than our targets, for example, infant care kit. So we first selected our target products from the data set.

2. Transform the abnormal characters. We find there are many abnormal characters in the data set. Emoji, foreign language and HTML tags will have a negative impact on further data programming. We used appropriate regular expressions to filter and translate some of them.

3. Missing value interpolation. Some of the reviews contents are missing. If this kind of reviews have a title, we use the title to replace its contents, otherwise we removed this review from the data set.

4. Data nationality test. We checked the nationality of the data. For example, if the number of helpful votes is smaller than that of total votes.

4 Sentiment Analysis

We use two methods to do our sentiment analysis: Naive Bayes Classifier and LSTM. Both have their advantages and disadvantages.

The advantage of Naive Bayes Classifier is that it can extract features from the text, which enables us to do a wordcloud. The disadvantage of Naive Bayes Classifier is that it doesn't consider the relationship of words, which is so-called context.

The advantage of LSTM is that it consider the relationship of words, which leads to a better accuracy. The disadvantage of LSTM is that it is an uninterpretable blackbox, which is a common defect of neural networks.

4.1 Dataset Partition

We select out reviews that are rated one star and five stars as our train-test set, where reviews with one star are tagged as "pos" and reviews with five stars are tagged as "neg". We use 70% of the train-test set as our train set and use remaining 30% as our test set.

We train models respectively for hair dryer, pacifier and microwave. The test accuracy are all above 0.90

4.2 Naive Bayes Classifier

Naive bayes classifier is a probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.

We use the Naive Bayes Classifier from TextBlob. We have extracted featrues from both postive reviews and negative reviews. Below are what we found in these reviews.

Fig.1 and Fig.2 are the wordclouds based on the featrues of postive reviews and negative reviews of hair dryers. We can note that in the negative wordcloud, the words "spark", "dangerous", "fire" appear frequently, which indicates that the company needs to pay more attention to safety issues if it wants to make an achievement in the hair dryer's market.

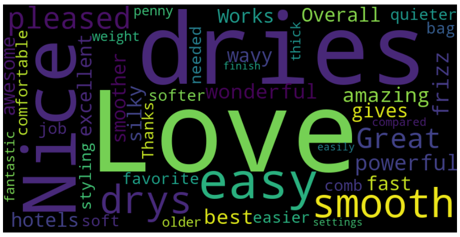

Figure 1: Hair dryer postive wordcloud

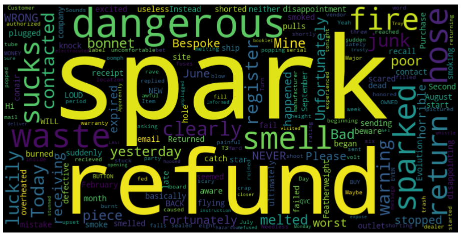

Figure 2: Hair dryer negative wordcloud

Fig.3 and Fig.4 are the wordclouds based on the featrues of postive reviews and negative reviews of pacifiers. As we can note, "cute" is cited a lot in the postive reviews. Hence a pacifier which is cute and lovely may attract babies' attention.

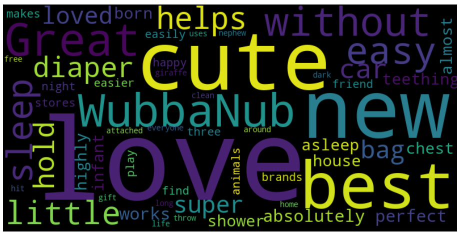

Figure 3: Pacifier postive wordcloud

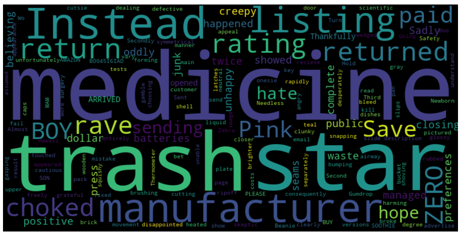

Figure 4: Pacifier negative wordcloud

Fig.5 and Fig.6 are the wordclouds based on the featrues of postive reviews and negative reviews of microwaves. In the negative wordcloud of microwave, "repair" and "warranty" stand out, which reveals that consumers mostly complain about the malfunction of the machines and unsatisfactory warranty. Companies that want make a success in microwave market should consider better after-sale service. Also they need to improve their products' quality.

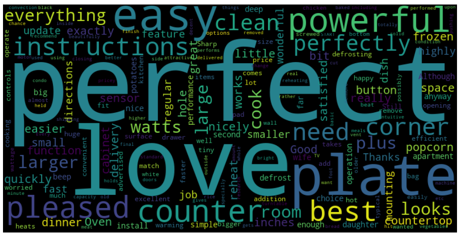

Figure 5: Microwave postive wordcloud

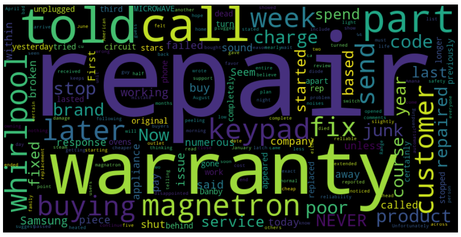

Figure 6: Microwave negative wordcloud

4.3 LSTM

LSTM is so-called long short-term memory. It is a kind of recurrent neural network(RNN). It differs from standard feedforward neural networks in that it has feedback connections. It is built for sequences of data instead of single data points, which makes it a good choice to do sentiment analysis since reviews also have contexts. We need these contexts to better predict the sentiment based on the text.

Our model structure is revealed as Fig.7. And the LSTM's structure is revealed as Fig.8

Some predicted results are showed in Table.1

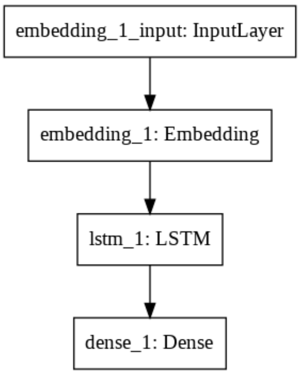

Figure 7: Model structure

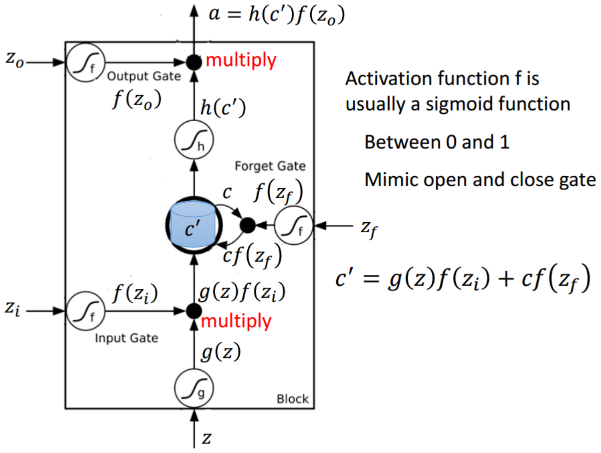

Figure 8: LSTM structure

Table 1: Some predicted results of LSTM

Reviews	Results
i really like this hairdryer i havent had it very long but its working well for me	0.9990048
was a great dryer loved it however it burned out within 2 months	0.06360153
clumsy apparatis i wanted more mobility like multitacking and this is not set up to run with the flow hose waned to stay on box and i wanted it to travel with me my own problem it was meant to be stable	0.5436161

5 AHP-FCE model

We used AHP-FCEM (Analytic Hierarchy Process-Fuzzy comprehensive evaluation model) to synthesize and conclude the characteristic of customers opinions.

Firstly, we conducted Correlation Test of several elements provided. According to the scatterplot and the correlation matrix, we found a strong linear correlation between the number of helpful votes and that of total votes, which is coordinated with common sense. Therefore, we combined this two elements into the proportion of helpful votes for further use.

Meanwhile, in order to reflect the representativeness of data, we calculated every weeks average value of each element.

Moreover, we also transformed N into 0 and Y into 1 to express the value of vine and verified purchase. We set three levels to evaluate the degree of verified purchase and two levels to evaluate that of vine.

Table 2: Main Symbols used in AHP-FCE repution Model

Symbols	Definition
TR_t	Total reviews in week t
TV_t	Total votes in week t
APH_t	Average proportion of helpful votes in week t
ASP_t	Average sentiment point in week t
ASR_t	Average star ratings in week t
AV_t	Average number of vine in week t
AVP_t	Average number of verified purchase in week t

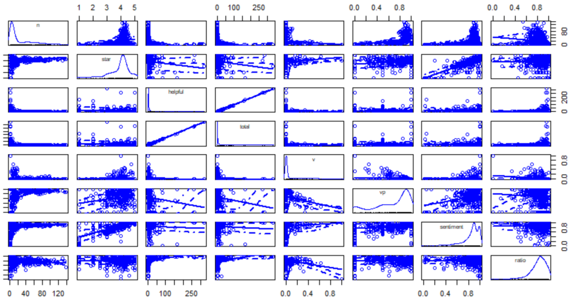

Figure 9: Correlation matrix

$$VL = \begin{cases} 2, & \text{if } AV_t \geq 0.05 \\ 1, & \text{if } AV_t < 0.05 \end{cases} \tag{1}$$

$$VPL = \begin{cases} 3, & \text{if } AVP_t \in (0.8, 1] \\ 2, & \text{if } AVP_t \in (0.5, 0.8] \\ 1, & \text{if } AVP_t \in [0, 0.5] \end{cases} \tag{2}$$

According to the elements provided in the data set, we conducted the following classification. The number of total reviews and number of total votes reflect the popularity of a product in a week. The satisfaction degree is combined with the proportion of helpful votes, reviews sentiment points and the star ratings customers gave. We also measured if a review is convincing by using the average number of vine and that of verified purchase.

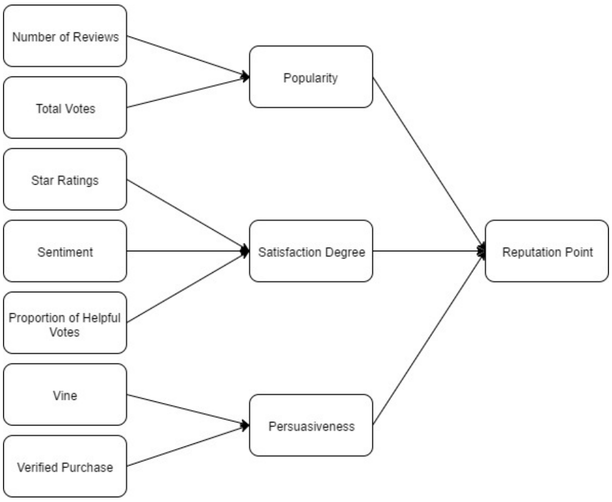

Figure 10: Classifier

$$\begin{aligned}\text{popularity}: POP_t &= \frac{1}{6}\log(10TR_t)\cdot\log(TV_t+10)\\ \text{Perservasiveness}: PER_t &= VL\cdot VPL\\ \text{Satisfaction Degree}: SD_t &= \frac{1}{45}APH_t\cdot ASP_t\cdot ASR_t\end{aligned}\tag{3}$$

Based on this, we made the following FCEM model. R_t represents the common reputation of specific product in a week.

$$R_t = PER_t(0.4POP_t + 0.6PER_t)\tag{4}$$

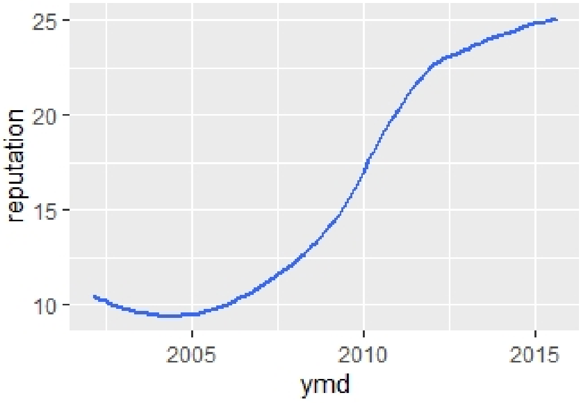

Figure 11: Reputation to time 2002-2015

6 Multivariate nonlinear regression model based on simplified time series

Table 3: Main Symbols used in MLR Model

Symbols	Definition
a	parameters
b	parameters
c	parameters
msr_t	The mean of star rating at time t
ms_t	The mean of sentiment at time t
mr_t	The mean of ratio at time t
α	The percentage of buyers willing to comment
β	Scope of information dissemination
γ	The rate of decline in the proportion of information
$R1_t$	Proportion of information recipients converted to buyers at time t
$R2_t$	The amount of information propagated at time t
Irp_t	The downward pressure on the proportion of information at time t
pb_t	Effective number of fans at time t
mrp_t	The pressure of the total market expansion leads to a decline in market influence at time t
ss_t	Simplified random event effects at time t

Parameter description:

$$\begin{aligned}
Reputation &= \alpha * R1_t * R2_t * Irp_t + pb_t + mrp_t * ss_t \\
\alpha &= 0.9 \\
\beta &= 1.4 \\
\gamma &= -0.01 \\
R1_t &= a * msr_t + b * ms_t + c * mr_t \\
R2_t &= \exp{(\beta * t^{\frac{1}{4}})} \\
Irp_t &= \exp{(\gamma * t^{\frac{1}{2}})} \\
pb_t &= 1/(a * msr_t + b * ms_t + c * mr_t) \\
mrp_t &= \exp{(-t^{\frac{1}{8}})} \\
ss_t &= (-t^{\frac{1}{4}})^3
\end{aligned} \tag{5}$$

Explanation:

Under the assumption that the number of reviews is positively correlated with the purchase.

Amount, we believe that the reputation of a product will be affected by the following quantities: the number of reviews, the number of effective fans and the occurrence of market events.

- The number of reviews:

  1. With the help of the Internet, the spread of commodity information will increase exponentially with time, that is, $R2_t$.
  2. However, as the amount of information in the whole Internet increases over time, the information proportion of this commodity decreases exponentially, and the rate of decline is measured by γ, Irp_t represents the impact of the decline(the percentage of product information that can be seen by people)The true number of information receivers at time t is: $R2_t * Irp_t$.
  3. $R1_t$ is not only the ratio of information receiver to buyer, but also the influence degree of comment emotion

  So the number of comments is equal to $\alpha * R1_t * R2_t * Irp_t$

- The number of effective fans:

  1. We divided the fans of merchandise into two categories: effective fans and invisible fans, Effective fans will help the sales and reputation of the products by increasing the purchase volume and scanning comments. Due to a series of difficulties, invisible fans will not take action in normal times, and only when the product is impacted, they will buy it. The number of effective fans means the number of potential purchases.
  2. When the comment sentiment level $R1_t$ is relatively low, it indicates that the product is hit by negative comments. At this time, fans will maintain the image of the product by increasing the amount of purchase and comments.Therefore, the reduction of $R1_t$ can promote the increase of effective fans

- The occurrence of market events:

  1. We assume that the market attention of a commodity is positively related to its age in the market. At the same time, positive and negative events occur randomly. To simplify the random factor, we alternate positive and negative events. The combination of event nature and market attention of the commodity constitutes ss_t, which will affect the reputation of the commodity.

Through multiple nonlinear regression, We take α=0.9, β=1.4, γ=-0.01, a=0.2, b=0.5, c=0.3, and obtained Fig.12:

Among them, the blue curve is the total market sales volume (cumulative volume), and the yellow curve is reputation. As can be seen, reputation better reflects the changing trend of total market sales. Due to the mathematical definition of reputation, it is bounded. At the same time, as t approaches infinity, reputation will approach zero. These two properties indicate that the influence of goods is limited and it will be eliminated from the market over time, which is consistent with our common sense.

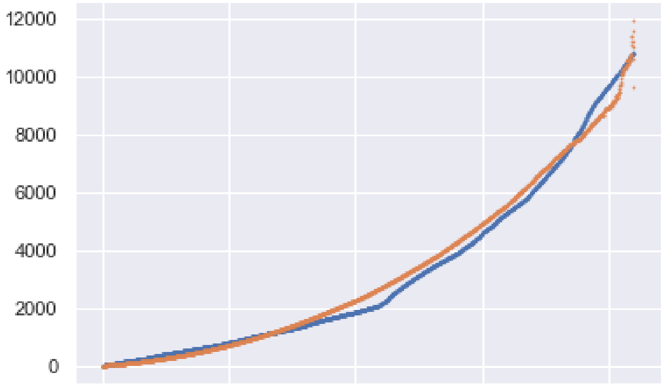

Figure 12: Simulation Curve of Reputation

7 Correlation Test

In terms of individual products, the number of stars in year n will affect the number of reviews in year n+1, and this influence will gradually increase.

From 2008 to 2015, the average star number of each product in the previous year and the total number of reviews in the following year were calculated, and the following distribution relationship was obtained. It can be seen that from 2008 to 2013, the impact of different star rating on the number of reviews in the next year is gradually obvious. Compared with low star rating, products with high star rating are more likely to receive more reviews in the second year. In 2014 and 2015, the average star rating of most products was between 3 and 5, and there was still a positive correlation between star rating and the number of reviews.

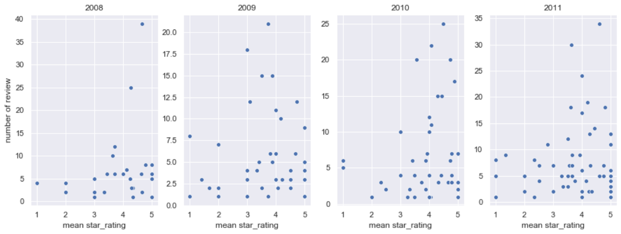

Figure 13: Mean star rating to number of review 2008-2011

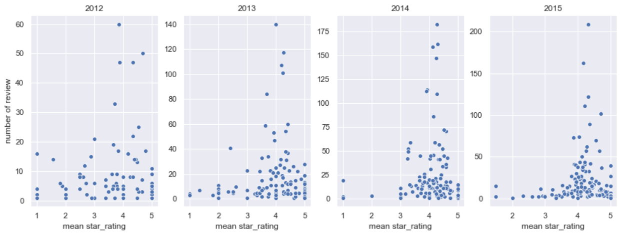

Figure 14: Mean star rating to number of review 2012-2015

In terms of the whole market, the distribution relationship between stars and comments is relatively stable. 1,2 stars will bring less comments, while 4,5 stars will bring more comments.

In each year, the number of stars and comments conforms to the quadratic relation, and satisfies: the number of comments in the second year corresponding to 1 and 2 stars is relatively low, while the number of comments in the second year corresponding to 4 and 5 stars is relatively high.After observing the data set, it is found that since most consumers have distinct emotions, the polarization degree of starmaking is high, so the number of comments in the middle of three stars is lower than that in the low stars

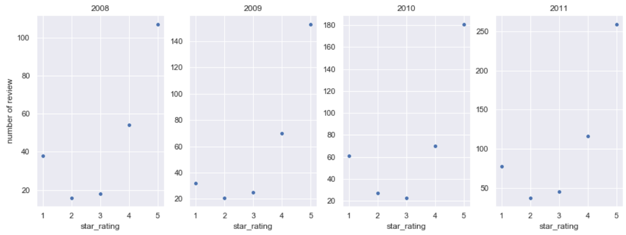

Figure 15: Star rating to number of review 2008-2011

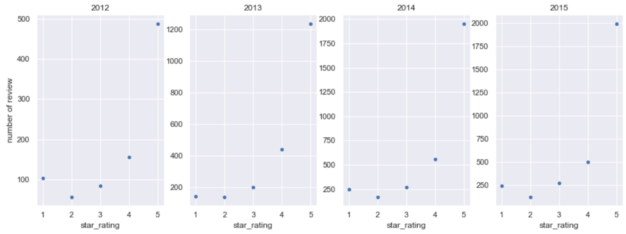

Figure 16: Star rating to number of review 2012-2015

In terms of the increment of the number of review, the influence of star number on the increment of comment number is increasing with the passing of time. Compared with low star rating, high star rating is more likely to bring the increment of comment number. At the same time, the distribution dispersion of comment number increment corresponding to low star rating is relatively small, while the variance of influence of high star rating on comment number increment is larger, which has higher uncertainty.

delta= Number of reviews in year n of a product - Number of reviews in year n-1 of a product Mean star rating= The average star rating of a product in year n-1

After removing some abnormal points, from 2008 to 2012, the distribution of mean star rating on delta is symmetric with respect to the line delta =0. Therefore, we believe that mean star rating disturbs delta with an average value of 0. Between 2013 and 2014,there is a greater probability that a product with larger number of mean star rating will have larger increment of reviews. Available data for 2015 show that as of August 31, the increment in comments for the year was still marked with a delta=0 symmetry, possibly due to incomplete data.

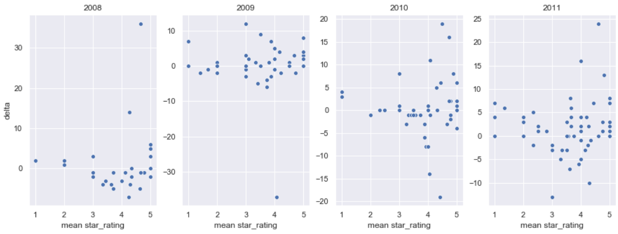

Figure 17: Star rating to delta 2008-2011

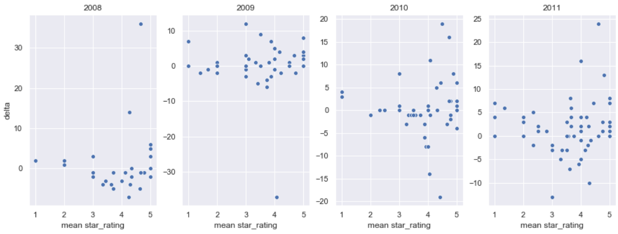

Figure 18: Star rating to delta 2008-2011

8 Conclusion

We use two models to do sentiment analysis: Naive Bayes Classifier and LSTM. There two models help us to select out some common featrues in postive and negative reviews, which is valuable for decision making.

To figure out how the reputation of product vary in a period of time, we construct the AHP-FCE model, through which we obtain a preliminary evaluation of products reputation.

Also, we study the market factors by constructing the Multivariate Nonlinear Regression Model.

In a nutshell, we find some valuable information for market descision making, which will be revealed in the memo.

9 Strength and Weakness

9.1 Strength

1. We used LSTM and naive bayesian models for emotion analysis respectively, and the results of these two methods were mutually verified with high accuracy.

2. It has a strong theoretical and practical basis. The model we set is consistent with the real life situation of the change of commodity sales over time. The reputation we constructed is based on the reality and reflects the future trend of product sales. .And after parameter modification, the model can be applied to other markets.

3. Universality..Our model can be widely applied to sentiment analysis of comment statements..

4. Creativity We creatively simplifies the random factors in the time series model that we constructed, and from the result, this simplification process has little impact on results. In the future model construction, the random terms in the model can be replaced and the solving process of the model can be simplified on the premise of less influence on the results.

9.2 Weakness

1. The parameters of the current model are only effective for the hair dryer market, and the other two markets need to find their own parameters.
2. We use future comments as a metric, but this is not necessarily true. We lack convincing indicators to measure the potential sales of goods, and the existing reputation is only based on common sense and market theory.
3. There is still some error in the best parameter we get, and maybe there is a better way to reduce it.

10 Memo

From: MCM/ICM 2020 Team #2020042

To: the Marketing Director of Sunshine Company

Date: March 10, 2020

Subject: Analysis and Suggestions on Online Marketing

From the data you provided, here we introduce our sentiment analysis model, AHP-FCE model and MNRM-MF(Multivariate nonlinear regression model based on market factors). Details and main results are as below:

- Potentially important design features

  1. hair dryer
     Most of the positive reviews mentioned quiet, powerful, smooth and easy, while most of the negative reviews focused on dangerous, spark, fire, smell. In conclusion, our design of hair dryer should be quiet, powerful, and easy to use. Safety is also a significant issue to pay attention to.
  2. Microwave
     large, quickly, easy are features praised most frequently in positive reviews, and some customers mentioned the power of the product. Meanwhile, many customers complain about their microwave using words fix, repair. To conclude, we should enhance the power of our microwave and make it easy to use. Also, our microwave ought to have a long service life.
  3. Pacifier
     "cute" is cited a lot in the positive reviews. Hence a pacifier which is cute and lovely may attract babies attention.

- Suggested online sales strategy:

  1. We should speed up our express delivery. Products delivered quickly will earn more reputation.
  2. After-sale service should be paid attention to. Many customers made complaints about products after-sale service, which may have a negative impact on our reputation.
  3. We can employ online ghostwriters to improve the online sentiment of our products. Customers tend to write more reviews and give more feedbacks based on the contents of past reviews.
  4. Company can take actions to promote their products during specific holidays, for example, Christmas Day.

References

[1] S. Hochreiter, and J. Schmidhuber, Long short-term memory, Neural computation 9 (8): 1735–1780 (1997).

[2] Zijian Zheng and Geoffrey I. Webb . , Lazy Learning of Bayesian Rules,Machine Learning volume 41, pages5384(2000)

[3] Freud, S., The Standard Edition of the Complete Works of Sigmund Freud, Vol. VII, J. Strachey (Ed.), London , Hogarth Press, 1958.

[4] H. Zhang., N., The Optimality of Naive BayesPsychother. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004), AAAI Press, (2004)

[5] comScore/the Kelsey group, "Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior" 2007.