Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2017225

PAPE: Processing, Analyzing, Predicting and Evaluating a Data-based Sales Strategy

Summary

Currently, as data of ratings and reviews are available, there is a synergy between consumers and companies in that consumers obtain purchase reference and companies tailor products to consumers' feedback. In this paper, we formulate a sales strategy with the key words of the product which has high total sales and high applause rate by detecting reliable reviews data for Sunshine Company's online business.

Firstly, we process the data provided, incorporating missing data imputation ,data synthesis and screening. In the process of data screening, we utilize the helpfulness ratings with **Laplace Correction** to preliminarily screen the data. To explore the intimate relationship between reviews and ratings, we analyze the data to select several parameters and set the definition of success. Hence we identify their qualitative patterns from a study on consistency and polarity. After that, the **TIS-SGD Model** is constucted to detect the consistency of review polarity and star rating. We also compare the accuracy of the **TIS-SGD** with **XGBoost** and trained TextBlob module by the **consistency**.

Secondly, we further define the success as a overall rank, which is composed with scores that measured by sales and applause rate.We bulid **a Polarity Quantification Model** to calculate the value of polarity about the headlines and reviews, to make a correlation analysis of reviews and overall rank. Based on the analysis above, wo introduce a regression model to study the Reputation changes over time. Then, through the statistics of the overall review of the text data about the relationship of star rating, we sum up the recommended product design characteristics.

Additionally, we gather the statistics of two distributions of word length and polarity of comments in the presence of each star separately and continuously. With **NLP(Natural Language Processing) theory**, we collect the statistics of the distribution of adjective word frequency and polarity corresponding to their star ratings and verify the strong correlation between star and polarity, thereby detecting the specific adjective word that appears the most in the different star ratings.

We finally conduct sensitivity analysis, dissect pros and cons of our model and present a memo of our work to Sunshine Company.

Keywords: TIS-SGD Model; XGBoost; TextBlob; polarity;consistency;NLP

Contents

1 Overview

1.1 Background

With the proliferation of the Internet, massive online platforms have been facilitating people' life. Particularly, the past decades have witnessed the unprecedented boom of the online market. In this accelerating trend, Amazon constuct a platform for customers to give their multiple ratings and diverse reviews of purchases. As for the former, which is called as "star ratings", customers can express their gratification by rating a product on a 5-star scale, with the lowest rating corresponding to the lowest gratification and the highest rating to the highest gratification. For the review, it allows customers to submit text-based messages in that they are able to voice their specific usage experiences and detailed products information, which is called as "the helpfulness rating". On grounds of these multiple data, customers have further and more authentice understanding of the product, thereby contributing to their options.

Additionally, such a review and ratings mechanism is also pivotal for corporations to gear their online sales strategy to the demands of consumers. As consumer's purchases and prefereces are documented by the platform, the company are able to craft the product with its features desired by the market and probably project the blueprint of future success.

Realizing that this particular combination and type of the data is win-win for both the customers and the manufactures, Sunshine Company collects the history data of the products they intend to introduce and sell: a microwave oven, a baby pacifier and a hair dryer. Then they hired us to help their sales project so that we are going to process, analyze, forecast and evaluate the sales data-based strategy.

1.2 Problem Restatement

Initially, we need to process the data of ratings and reviews and make a comprehensive analysis. The analysis is supposed to be based on the data and parameters we identify so that we can determine the quantitative or qulitative patterns, relationship and measures of them. With the preliminary understanding of the data, we need determine the definition of successful sales.

Additionally, with the provided data and identified parameters, we are required to identify the data measure that contribute to the track of sales. Then we should base the pattern on time factor so that explore how a product's reputation varies by that. Next, combining both reviews and ratings, we should make predictions of a successful or failing product.

Finally, we are supposed to explore more concrete relation of reviews and ratings.That is to say, on the one hand, detect future trend of reviews influenced by the specific star ratings. On the other hand, we need to identify how the review incite some rating level with some specific quality descriptor.

1.3 Our Goals

Based on the comprehensive understanding of the problem, we intend to achieve the followinng goals:

In the process of the data processing, determine the objectives of the success(total sales and applause rate) for Sunshine Company's three new online products. Among the star ranting, review and helpfulness rating, analyze their respectful influence on the objectives and the relationships between them. Then research and identify the qualitative patterns, measures and parameters of these three types of data. Demonstrate their impact on success to make preparations for a concrete sales strategy.

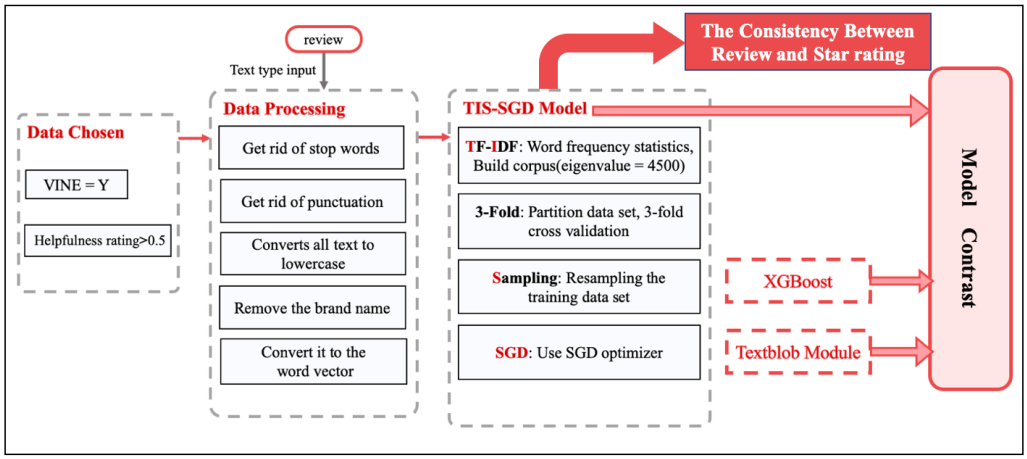

Figure 1: The Framework of Our Goals

Next, we update the definition of successful sales to offer more elaborate and comprehensive strategy.In order to identify the quantification of the paraters we determine before, such as polarity and confidence level, we creat and calculate the quantification model. We also use TextBlob to make a correlation analysis of polarity and comprehensive score. Combining the parameters and statistics method, we should determine the time-based patterns and make predictions of sales and applause rate in the near future by Bayes Regression Model.

2 Assumptions and Notations

2.1 Assumptions

In our model, we set the following assummptions:

- **We assumpt that the criteria reflecting the objectives of success are the total sales and applause rates and are merely influenced by their own quality and feature design.** This assumption ignore other subjective and objective factors (such as advertising effect and competitive relationship). Since the prodect prices are unknown, we only consider the total sales and applause rate based on the data set in our model.
- **We assumpt that the keyword of product titles and reviews are the product's feature design.**Since product title summarizes its features and reviews express the use feelings of its features, so we can identify the successful features from those keywords.

- **We assumpt that the ratings and reviews are authentic and reliable to a certain degree.** Despite a few missing values, we will make a reasonable screening on the unreliable data to guarantee our valid solution.
- **We assumpt that Amazon Vine members who are invited to become AVV (Amazon Vine Voices) are able to accurately and genuinely point out the merits and demerits of the products. And they can give unbiased star ratings that are consistent with their comments.**
- **We assumpt that the text-based measures of the review accuracy will not alter by the time.**
- **We assumpt that the relative quantitative relation of three products don't mirror the success of sales.** Since the demands and wastage differ as to different types of products. So the relation doesn't reflect whether it is successful.

2.2 Notations

Here are the notations and their meanings in our paper:

symbols	definitions
i	the identifier of the product
j	the identifier of the consumer
k	the identifier of the review
TS	total sales
r_{A_i}	applause rate
S_{ij}	the score of the star rating,
N_t	the number of star ratings if $t = S_{ij}, t = 1, 2.., 5$
r_{H_k}	the ratio of the helpfulness rating
V_{H_k}	number of helpful votes.
V_{T_k}	number of total votes the review received
$r\tilde{_{H_k}}$	the ratio of the helpfulness rating after Laplace correction(coonfidence level).
TF_m	term frequency of word vector x_m
IDF_m	inverse document frequency of word vector x_m
x_m	word vector
N_{mk}	the times the word vector x_m appears in the review that is signed as k
N_x	the time the word vector x_m appears in the entire data set
w_m	the out put of the word vector x_m
θ_m	the weight of the word vector x_m
y	the label

3 Data Processing

3.1 Missing Data Imputation

Given that a couple of missing values only exist in data files: $pacifier.tsv$ and $hair_dryer.tsv$, we make the data imputation as follows: for a few missing part values of data, we fill them with ""; for a few missing part values of reviews, we substitue them with "".

3.2 Data Synthesis and Screening

As for $microwave.tsv$, $pacifier.tsv$ and $hair_dryer.tsv$, there are 1615, 18939 and 11470 data in sum respectively. Each type of products has 80, 6482 and 538 products respectively. The ranges of the times each product is rated rspectively are $[1, 117]$,$[1, 515]$ and $[1, 587]$. Judging from the reality, Not all the comments and star ratings are completely impartial and correct. Therefore, it is necessary to screen the needed data, thus resorting them into three new data set. Before determining the screening standard, wo are supposed to identify the definition of three types of data.

- **Star rating(S_{ij}):** The number of the stars is its score.
- **Review:**The text-based data that reflects the quality and feature design of the product. Also, the data it covers express the emotive information of customers.
- **Helpfulness rating(r_{H_k}):**The indication that reflects the confidence level of the review.

$$r_{H_k} = \frac{V_{H_k}}{V_{T_k}} \tag{1}$$

The reason of screening data: considering the reality, not all the reviews and star ratings are objective and accurate(such as the malicious evaluation and submission errors), we should remove some data failing to embody quality and characteristic of the product correctly.

Screening according to the following criteria:

(i) $Vine(string)$: Keep all the data whose $Vine(string) =$ "Y".Since all the reviews and star ratings are regarded as objective and accurate, they are the indispensable factors to embody product characteristics and quality.

(ii) $Helpfulness\_rating$: Since great amounts of total votes in the helpfulness rating(V_{H_k}) has a zero value. We make a Laplace correction on it.

$$r_{\tilde{H_k}} = \frac{V_{H_k} + 1}{V_{T_k} + N}, (N = 2) \tag{2}$$

Then we keep the data whose $r_{\tilde{H_k}} > 0.5$

4 Analysis of Data and Basic Parameters

4.1 Definition and Analysis of Success and Paramters

After the data processing, we select and detect some variables to creat our model and make further analysis of them. These variables are from the three types of data and called as basic parameters.

- **Total sales(TS):** the total amount of the product has been sold, which is calculated by the amount of reviews it has received. **It is one of the definitions of success.**

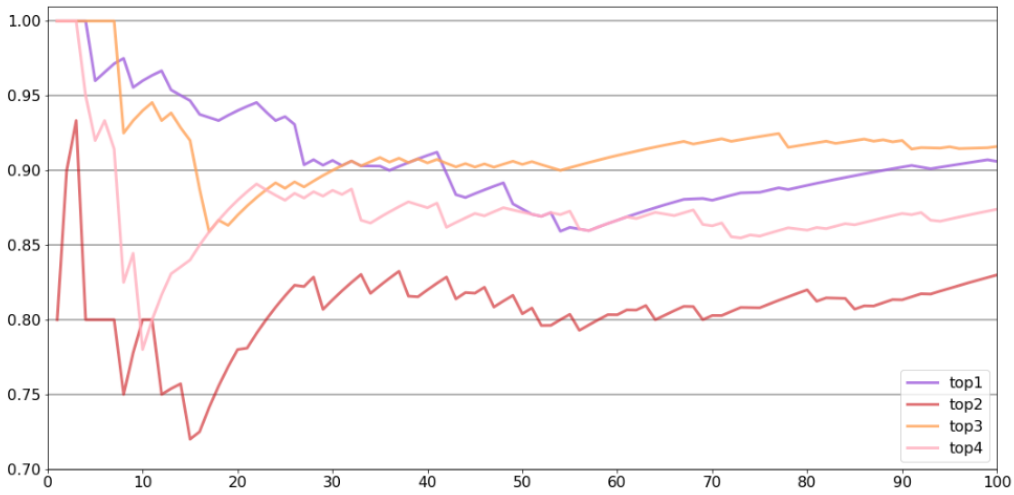

Figure 2: The Framework of Our Goals

- **Applause rate(r_{A_i}):** the average rating of each product over the entire time period.

$$r_A = \frac{\sum_{j=1}^{N} S_{S_{ij}}}{5 \sum_{t=1}^{5} N_t}, \tag{3}$$

  We set N as the amount of reviews the product has received. **It is the other definition of success.**

- **Key word:** the text of it can embody the traits of the product.
- **Confidence level ($r_{\tilde{H}_k}$):** the ratio of helpfulness rating after Laplace correction.
- **Polarity:** criteria for binary classification of comments based on textual information (a qualitative analysis standard). The sign of value (positive and negative) can embody consumers' attitudes towards products and the absolute value reflects the consumers' intensity degree of attitude towards the product. [**?**]
- **Text characteristic:** numbers of noun, adjective, verb, the length of a sentence [**?**].

After the identification of the definitions above, we have a general analysis about the historical market atmosphere about the total sales and applause rate of three products. The trends are shown as follows: As the yellow lines show, apparently, the total sales keep presenting an ascending trend as the

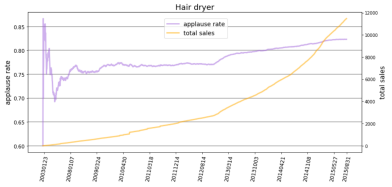

Figure 3: Hair dryer in the market atmosphere

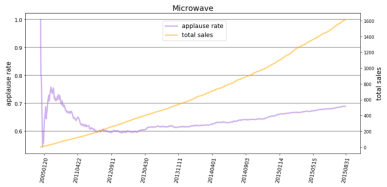

Figure 4: Microwave in the market atmosphere

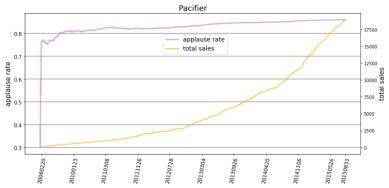

Figure 5: Pacifier in the market atmosphere

time passes. However, observing the pruple lines, the gneral tendency is rising while every product

presents differen degree of fluctuations. Hair dryer has a drastic fluctuations in early years, but now it is in a steady increase; microwave, conversely, after a sudden decrease, its ascending trend is slow and unconspicous; pacifier has a drastic increases in early years, and since then, it has kept its high sales.

4.2 The Qualitative Patterns of Reviews and Ratings

4.2.1 The Consistency of Reviews and Star Ratings

Since reviewers are likely to have different understanding of star rating and their objectivty are various, the scores of star rating maybe inconsistent with the reviews (such as a positive review with a one-star score). To further screen the data, we intend to determine the established relationship (consistency) between the review and star rating. Hence ,we define the consistency as follows: First, according to the assumption about the Amazon Vine members above, we classify the Vine members with consumer whose confidence level is high ($r_{\tilde{H}_k} > 0.5$) as a sample group. Then we deem their relationship between review and star rating as objective and accurate. Next, we regard such a relationship as consisten. In this way, we can judge the accuracy and objectivity of the relationship by comparing with the consistency.

4.2.2 Polarity: Binary Classification

We utilize the polarity to implement a binary classification according to the definition of polarity. The review with a 4-star and 5-star score is positive while the one with a 1-star and 2-star score is negative. The reason why we avoid the consideration of 3-star review is that we consider three aspect of it. First, we are inclined to conclude the influential ingredient of the success and failure while a 3-star level theoretically has the same contribution to the success and failure. Second, in the **NLP (Natural Language Processing)**, the neutrality has a certain degree of ambiguity in definition. To guarantee the accuracy, we don't take it into account.

4.3 The Construction of The TIS-SGD Model

Since we have a preliminary understanding of consistency and polarity. in our model, we intend to make an explicit judgment on the data of reviews about whether they are positive and negative. This model will benefit follow-up work on the text-based data analysis.

4.3.1 The Precondition for the TIS-SGD Model

First, we operate on the text-based data. We preprocess all the texts of reviews by removing stop counts, punctuations, "hair dryer","pacifier" and "microwave"(as they can not reflect the polarity). We also convert the text toall upper case or to all lower case, cutting the sentences into words. Also, we convert every word into a word vector so that each sentence is a 4500 dimensional vector. Count the number of times each word appears and make all the word vectors into a review lexicon

Second, We divide the modeling samples (the sample group mentioned in the **The Consistency of Reviews and Star Ratings**) into training sets and test sets. We selected 20% samples from different star levels as the test set of the model.

In the process of inputting vectors, we find that the magnitude of the 1-2 star reviews is quite different from that of the 4-5 star ones. Therefore we opt to resample the data of 1-2 star with SMOTE , which is used to offset the impact of imbalanced data. During data resampling, comments are mapped to word vectors. We sign the imbalanced vector as x, pick one of its neighbors at random, sign it x_n. Then we obtain a new sample vector x_{new} according to the following formula:

$$x_{new} = x + rand(0,1) * |x - x_n|, \tag{4}$$

x_n is the semantic closest word in the review lexicon. As is shownabove, some noise is added to fit the new data to get the same number of negative comments and positive comments.

Next, we use TF-IDF to operate on data. TF-IDF is a statistical method to evaluate the importance of the word vector in the comment lexicon. The importance of the word vector increases in proportion to the number of times it appears in the comment data. We select the top 4500 words with the highest frequency in the lexicon, where the words have been converted into word vectors. The training set is fed into the model for training, and the training set is divided into positive and negative.

$$TF_{x_m} = \frac{N_{km}}{N_x}, \tag{5}$$

$$IDF_{x_m} = \log \frac{k}{k_m + 1}, \tag{6}$$

(k is the number of the reviews in the data set,k_m is the number of the reviews that incorporate the word vector x_m)

$$w_m = TF_{x_m} * IDF_{x_m}, \tag{7}$$

In this way, we can obtain the weight (w_m) of the word vector, which demonstrate the importance of the word vector.

4.3.2 The Construction of the TIS-SGD Model

To further explore the relationship between other reviews and ratings whose consistency is difficult to determine, we develop a **SGD (Stochastic Gradient Descent)** which is based on the optimisation technique. In our model, the input is the word vectors above, the objective function is $h(x)$, the loss function is hinge loss $l(y)$

$$h(x) = \sum_{m=0}^{n} \theta_m x_m, \tag{8}$$

$$l(y) = \max(0, \frac{1}{4} - 2 * [y - \frac{1}{2}) * (h(x) - \frac{1}{2})]), \tag{9}$$

Since the sentence is a high-dimensional sparse matrix, we use L2 regularization to reduce the complexity of the objective function.

$$\begin{aligned} J(x) &= \tfrac{1}{2}||w||^2 + C \textstyle\sum_m max(0, 1 - y_m w^T x_m) && (10)\\ &= \tfrac{1}{2}||w||^2 + C \textstyle\sum_m max(0, m_m(w)) && (11)\\ &= \tfrac{1}{2}||w||^2 + C \textstyle\sum_m L_{Hinge}(m_m)) && (12) \end{aligned}$$

After that, we can use the test set to test its accuracy and recall rate.

4.4 The result of the SGD Model

We construct the TIS-SGD Model to understand the relationship between star rating and reviews. Also, by the polarity we calculate from the model, we can further screen the review that is inconsistent with the star ratings.The result is shown as follows by demonstrating the confusion matrix: From the

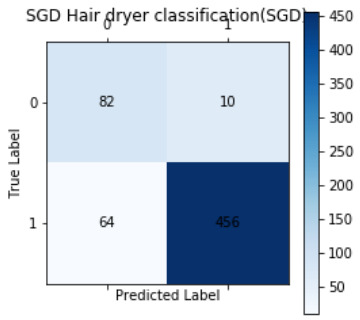

Figure 6: SGD-Hair dryer reviews classification result

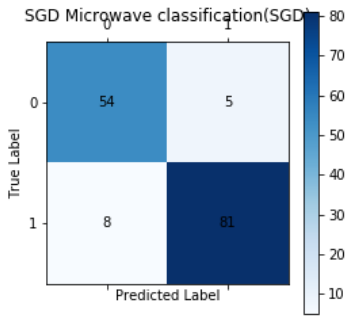

Figure 7: SGD-Microwave reviews classification result

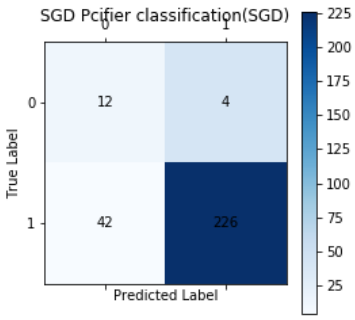

Figure 8: SGD-Pacifier reviews classification result

3 figures, we can see that the percentage of accuracy of the model can attain at least 83.10%. Also, by these data, we can classify most reviews into 2 distinct group: the positive group and the negtive one.

4.5 Accuracy of SGD, XGBoost and Trained TextBlob Model

Compared with TIS-SGD Model, we also use the XGBoost and trained TextBlob model.

- **XGBoost** is a tree learning algorithm that can deal with the sparse confounding matrices we study. Likewise, we take the input and output as above,we can calculate the confusion matrix according to the formula as follows:

$$Obj^{(t)} \simeq \sum_{m=1}^{n} [l(y_m, y_m^{(t-1)}) + g_m f_t(x_m) + \frac{1}{2} h_m f_t^2(x_m)] + \Omega(f_t) + bias \tag{13}$$

$$L(\psi) = \sum_{m} l(y_m^t - y_m) + \sum_{k} \omega(f_t) \tag{14}$$

  And, the result is shown as follows: We can see that the percentage of accuracy of the model can transcend 80%, but is still lower than the percentage of the TIS-SGD Model.

- **TextBlob**:The naive bayesian classifier based on NLTK TextBlob module is used to analyze the polarity of emotion. When its polarity score is greater than 0, it is positive; if it is less than 0, it is negative.We can also compare its result likewise. In this way, we will find that the percentage of accuracy of it is also lower than our model. Consequently, we select the TIS-SGD Model as the basement of our subsequent study.

After the above data calculation and analysis, we use the confidence level and TIS-SGD model to screen the data twice. We retain the data that we approve, that is, the comments of the data can

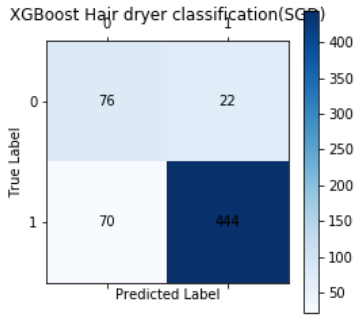

Figure 9: XGBoost-Hair dryer reviews classification result

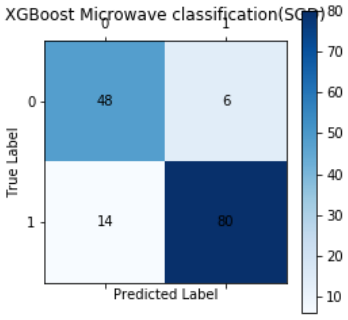

Figure 10: XGB-Microwave reviews classification result

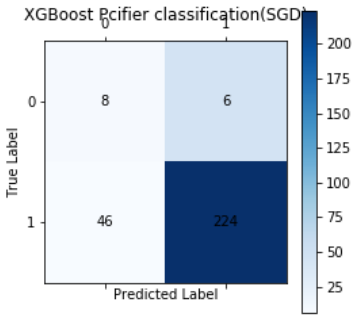

Figure 11: XGBoost-Pacifier reviews classification result

be considered to be consistent with the star rating. Therefore, for all of the following modeling processes, without special instructions, we maintain utilizing the data set that are filtered twice.

5 A Polarity Quantification Model

This model base on the identification of the new definition of success, the quantification of polarity by TextBlob and the correlation analysis of polarity and comprehensive score.

5.1 Track the Dynamics of the Applause rate with Sales

s we have stated above, the ultimate goal of manufacturers and companies is the comprehensive pursuit of high sales and high reputation.Excellence in a single indicator does not represent a company's sales success.Therefore, based on the above, we further define successful sales.We believe that the criterion of success is a comprehensive score that considers both sales and favorable comments. First of all, we set up a sales ranking, according to the sales of the ranking, according to the ranking.After this treatment, each item receives a sales score.Similarly, we can also get a separate ranking score for positive ratings.We take the sum of the two scores as the criterion for the comprehensive ranking, and the following are the top four for analysis. To obtain comprehensive score: map sales volume and applause rating to an interval of (1,5) for addition by Min-Max Scaling.The linear function converts the original data linearization method to$[0,1]$.

$$X_{norm} = \frac{X - Xmin}{Xmax - Xmin} \tag{15}$$

After that, we take four times the value as the score, that is, the score range is$[0,4]$, which corresponds to the star rating. What we can see is that the early changes in all three types of products are quite drastic. This is for two reasons. First, the early online shopping market is not mature. Second, there are too little early data and the results are one-sided. So we want to find a threshold for sales when positive reviews tend to stabilize. We take the hair dryer as the prime example to analyze. According to the figure on the left, we first selected the product change chart with the sales volume greater than the average sales volume.We found that from the sales volume above 50, the fluctuation of the favorable

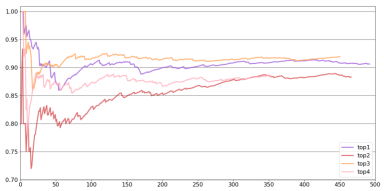

Figure 12: Hair dryer: applause rate changing by the sales

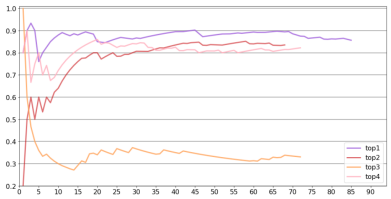

Figure 13: Microwave: applause rate changing by the sales

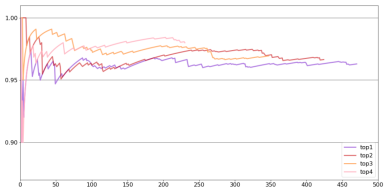

Figure 14: Pacifier: applause rate changing by the sales

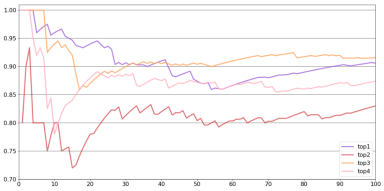

Figure 15: 1-Hair dryer: applause rate changing by the sales

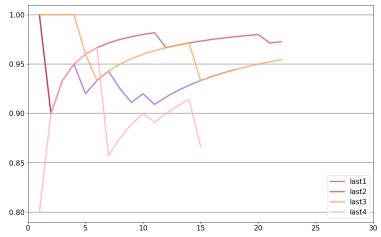

Figure 16: 2-hair dryer: applause rate changing by the sales

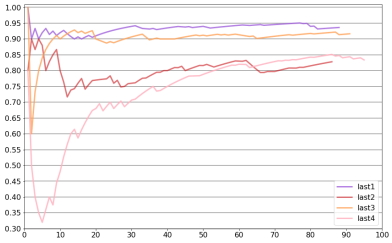

Figure 17: 3-hair dryer: applause rate changing by the sales

rating of hair dryer began to stabilize.In order to further determine the threshold, we selected the product changes ranked in the bottom four.In the middle picture, the fluctuation is still large.Finally, we select the right figure and observe that the threshold of stabilization can be selected as 50. Similarly, microwave ovens and pacifiers were analyzed.Finally we determine the three thresholds of 50,30 and 50.

5.2 Correlation Analysis of Polarity and Comprehensive Score

We take $nltk.corpus.movie_reviews$ as the review lexicon. With the use of the previous TextBlob, we calculate the polarity of the headline and body in the review. Take these two variables as the result of our quantization of polarity. Then we make the correlation analysis of polarity and score. We obtain the result as follows:

6 Regressive Prediction Model

6.1 Linear Regression Model

To explore the tendency of reputation over time, we introduce the linear kernel function is combined with the regularized linear regression model.

$$f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{16}$$

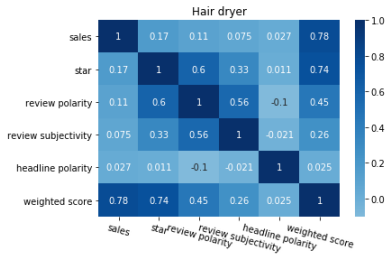

Figure 18: Hair dryer: The thermal map of polarity with respect to the coefficient of success

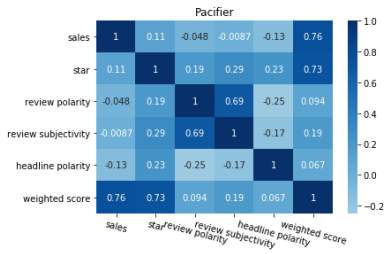

Figure 19: Pacifier:The thermal map of polarity with respect to the coefficient of success

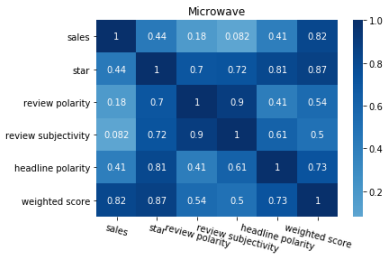

Figure 20: Microwave:The thermal map of polarity with respect to the coefficient of success

First, we calculate the applause rate of top four products every year. We make an analysis on the sales of the total products and each product. We can detect than the curve has a similar trend. In the figure, we can see the future tendency from 2016 to 2017.

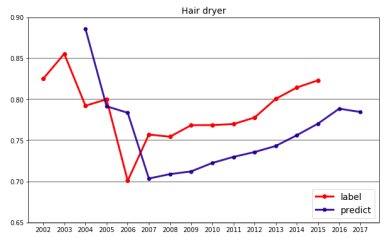

Figure 21: Hair dryer: The market atmosphere forecast 2016-2017

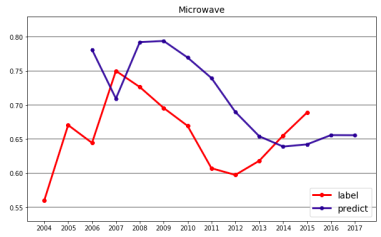

Figure 22: Pacifier:The market atmosphere forecast 2016-2017

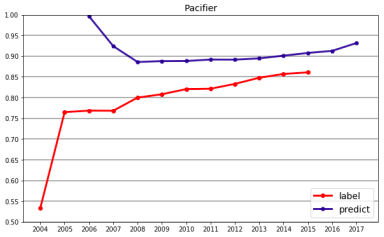

Figure 23: Microwave:The market atmosphere forecast 2016-2017

6.2 Bayes Regression Model

Now we creat the improved Bayes Regression Model.

Initially, we implement the data cleaning without dividing the text-based data into word vectors. Then we take the top 50 products in total sales as our statistics data set. Based on this data set, we build lexicon by collecting the applause rate of the noun and adjective words in the data of $product_title$. The applause rate of some word is similarly calculated by the $equation(3)$. We calculate the average applause rate of all the words, setting it as the prediction score of this product. After that, we do correlation analysis between the prediction value and real one. Then we get the result of hair dryer, microwave and pacifier respectively at 0.85, 0.99 and 0.85.

Next, because the data of microwave oven is less, so the statistical accuracy is higher. With the improvement of vocabulary, the prediction accuracy of the model tends to be normal.

We place the top parameter values in the constructed product feature dictionary at the bottom for the Sunshine company to consider when designing the product.

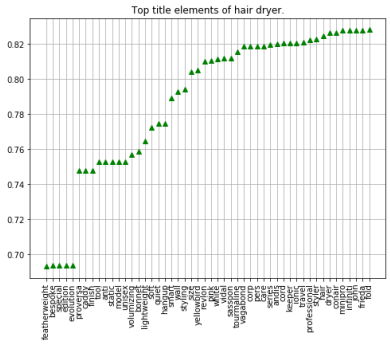

Figure 24: Top title key word-Hair dryer

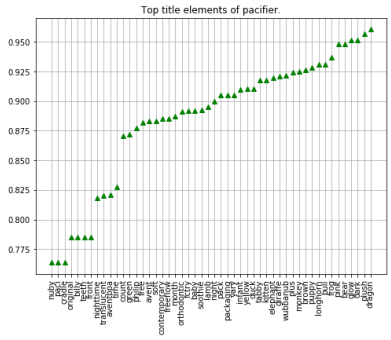

Figure 25: Top title key word-Pacifier

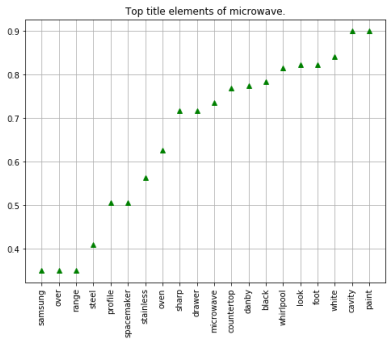

Figure 26: Top title key word-Microwave

7 Specifi Association of the rating level with review

7.1 Specific Star Rating Impact

To find the relationship between specific star rating and reviews, we traverse the text-based data of review from various products.If two consecutive reveiws of the same star rating level appear, we record the length and polarity of the first different review, thus comparing it with the average values of them.

We make an analysis on the length of reviews that come from differnt star levels with different polarity. We take pacifier as a prime example, detecting that the length of consecutive positive reviews is 79.71, greater than 47.25, the average value of the positive one-star reviews.Conversely, the length of consecutive negative 5-star reviews is 54.64, greater than 62.49, the average value of the positive reviews. Other values calculated similarly turn out the the similar consequence.

7.2 The Grounds of Specific Quality Descriptor from Reviews

Now we explore the text-based data of the adjective words who has a strong correlation with star ratings, we implement a further analysis on each review. We classify reviews into five distinct group by the criterion of star rating. Neglecting the 3-star level as before, we make a statistics research on the rest of reviews. The result is demonstrated as follows:

8 Strengths and Weaknesses

The TIS-SGD model we use can fill the unbalanced data with better data.This model is more predictive than the XGBoost and TextBlob modules in terms of accuracy.Considering the fuzzy boundary of the multilevel classification of Chinese text emotion in natural language, the model combines the binary classification of text emotion with the objective of this topic.The TIS-SGD Model can guarantee The accuracy of prediction without affecting The operation of subsequent data mining, and can reflect The consistency between comments and stars in The current data set.

However, due to the small training data set, the actual accuracy of the label is unknown.Some errors still exist in the prediction of the model.And this will affect the subsequent analysis to some

star	product name	word	time	star	product name	word	time
1	hair dryer	disappointed	53	4	pacifier	little	558
1	hair dryer	dangerous	31	4	pacifier	good	465
1	pacifier	hard	46	4	microwave	good	96
1	pacifier	disappointed	45	4	microwave	great	81
1	pacifier	poor	20	5	hair dryer	great	2334
1	microwave	sharp	66	5	hair dryer	good	1131
1	microwave	poor	23	5	hair dryer	powerful	605
2	hair dryer	retractable	25	5	hair dryer	quiet	358
2	hair dryer	loud	14	5	pacifier	great	3193
2	pacifier	small	18	5	pacifier	little	1886
2	pacifier	disappointed	18	5	pacifier	easy	1707
2	pacifier	hard	16	5	pacifier	good	1230
4	hair dryer	great	618	5	pacifier	cute	879
4	hair dryer	good	606	5	microwave	great	272
4	hair dryer	pwerful	175	5	microwave	small	141
4	hair dryer	cool	123	5	microwave	easy	123
4	hair dryer	quiet	102	5	microwave	good	122
4	pacifier	great	686	5	microwave	little	95

Figure 27: specific quality descriptors of text-based reviews

extent.Furthermore, the predictive power of the model is related to the dictionary library.When new words appear, the model has difficulty capturing their emotional polarity.In other words, although the model is more targeted, the generalization ability is weak.

In the process of quantification of success, considering sales and received the makes the success rate of more comprehensive consideration, the model makes full use of all the data, and the noise value point was clear analysis, to find in the chaos of the market data to a smooth area, and to the subsequent analysis of the stability of the cement strength.The linear regression model can accurately express the trend of market data changes, but there is still room for improvement in model accuracy compared with LSTM model in the popular neural network in time series prediction.

In addition, the improved bayesian model based on product characteristics has a very high correlation between the predicted value and the real value, with low model time complexity and high accuracy, but it is also limited by the small generalization ability of the data set.With more data, the accuracy may fall to about 80, which is normal.

9 Conclusion

In this paper, we first determine the definition of successful sales, that is, high sales, high praise. Then, after a pre-processing of the data, we created a TIS-SGD model. This is the basic model for all of our parameters, data analysis, post prediction, and also the core model .Then we build a quantitative model of polarity, and a correlation analysis of successful goals.We effectively took advantage of our model.The accuracy of the model under repeated testing also helps us gain a lot in dealing with the relationship between text and numerical data.After making clear the advantages and disadvantages of the core model, we can study the generality of the model.And the derivation of other auxiliary models under this model.

Letter

To: The Marketing Director of Sunshine Company

From: Team 2017225

Date: March 8, 2018

Subject: The Data-based Online Sales Strategy

Honorable The Marketing Director of Sunshine Company,

Currently, your company intend to introduce and sell three new products online, incorporating a microwave oven, a baby pacifier and a hair dryer. Our team implement a comprehensive study on the ratings and reviews based on the historical data for decades, calculate and predict the time-based pattern and dynamic relationship of them. With optimisation technique and regression model, we ultimately acquire the successful online sales strategy and detect the optimal product design to boost your company's online business!

Data Measures: We identify the polarity and consistency as the most two pivotal data measures. Since the reviews are too diverse that we cannot judge by mere text-based data, polarity can convert the complicated text into explict and valid numeric data, thereby simultaneously contributing to the qualitative and quantitative research. The polarity can classify the text into two distinct group, the positive word and the nagtive one. In this way, it turns a complicated emotive information into binary patterns. Additionally, the parameter consistency reflect the relationship between the review and the star rating, which originally is inclined to be influenced by consumers' objectivity and understanding.

Predictions: We predict a near future of three products'sales tendency. We detect that both pacifiers and hair dryer might have a more impromising future. Because their ascending inclination is obvious. Nevertheless, the sales of microwave are optimisitic. Microwave is a kind of electrical appliance, whose sales and applause rate on average and in time-based pattern are all rising slowly and even be inclined to be steady.

Product Feature Design: A hair dryer:smart quiet,powerful,low watt. A microwave:white,small,with cavity, with whirlpoor pacifier:dark/pink dragon/bear/frog shape little cute

The above is the summary of our analysis and recommendations. Sincerely hope that our work will boost Sunshine Company's business!

Thanks!

References

A Appendix A:TIS-SGD Model Python code

```python
import pandas as pd
import joblib
from sklearn import metrics

from textblob import TextBlob
FDir = 'C:/Users/SkyTu/Desktop/数模竞赛/Problem_C_Data/'
Dir = 'D:/python_enviroments/PythonCodes/MathCopetition/解答/第二题/used product id/'
MDir = 'D:/python_enviroments/PythonCodes/MathCopetition/解答/第一题/SGD模型/'
hair_dryer = pd.read_csv(FDir+'hair_dryer2.csv')
microwave = pd.read_csv(FDir+'microwave2.csv')
pacifier = pd.read_csv(FDir+'pacifier2.csv')
hair_dryer_product_id = joblib.load(FDir+'sorted_hair_dryer_product_id_above_50.pkl')
microwave_product_id = joblib.load(FDir+'sorted_microwave_product_id_above_50.pkl')
pacifier_product_id = joblib.load(FDir+'sorted_pacifier_product_id_above_50.pkl')

#%%

def get_above_average(dataset, useful_list):
    useful_var = list()
    useful_dict_var = dict()
    for line in dataset.values:
        if line[3] in useful_list:
            useful_var.append(line)
            if line[3] not in useful_dict_var:
                useful_dict_var[line[3]] = 0
            else:
                useful_dict_var[line[3]] += 1
    useful_var = pd.DataFrame(useful_var,columns=dataset.columns)
    return useful_var, useful_dict_var

#%%
hair_dryer_var,hair_dryer_dict = get_above_average(hair_dryer,hair_dryer_product_id)
microwave_var,microwave_dict = get_above_average(microwave,microwave_product_id)
pacifier_var,pacifier_dict = get_above_average(pacifier,pacifier_product_id)
```

Figure 28: Python code