Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2008384

The Wealth of Data: Mining Patterns behind Online Customer Reviews

Summary

We build models and offer analysis to help the online business of Sunshine company. Specifically, we provide models to help the director analyse product purchases and reputations, and give replies to the director's questions.

We first use Naive Bayes model to do sentiment analysis to text-based reviews to help determine the purchase actions of customers reviewed online. The model is helpful in mining sentiments of reviews of customers and predict selling situation. We use K-Means Clustering to construct model on star ratings, and use the model to predict purchase actions of customers. Relying on the two models, we provide ways to analyse purchase separately using reviews and star ratings.

Then we use a Conditional random field model to combine review-based Naive Bayes model and rating-based K-Means model. We also introduce time-based factors to the CRF model by using former outputs of the CRF model itself to help determine new outputs. The combined model achieves better results in predicting purchases than the two models separately. The introduction of time-based model also solves the problem of lack of data, for the combined model achieves far better results in small dataset.

We construct random walking model to analyse the reputation change of products using the reviews of customers on the product. We use improved Markov chain to model the probability change of every step of the random walking. The simulation results show success after compared to the changes of customer star ratings.

After constructing the 4 models, we use them to analyse questions the director raised. We compare results of text-based model and rating-based model and conclude that text-based measure is more informative. We use the random walking model as the time-based measure to suggest the reputation change. We combine text and rating measures using CRF models. We analyse purchase actions after impressive reviews and conclude that users tend to have similar impression after seeing them. Finally, after analysis we find that most positive quality descriptors tend to appear more in positive ratings and negative ones more in negative ratings.

Keywords: Online Review Data Mining Purchase Product Reputation

Contents

1 Letter to the Marketing Director

Dear Director,

I write to you on behalf of the MCM team 2008384. We did careful analysis on the questions you concern about. Using the data you provided, we constructed models to mine patterns and information about customers and products. We used the models to analyse your concerns and came with conclusions to all of them. Now we give formal replies to your questions.

We first built Naive Bayes model for sentiment analysis on the reviews. We also used K-Means clustering on star rating data. After the experiment, we found that the review-based model achieves far better results than the rating-based model on predicting customers' purchase behaviours. After analysing the test result, We conclude that the amount of information reflected in the reviews is far greater than the star ratings. Customers express more detailed and real feelings about the products, which is crucial to their final purchase decision.

We then built a random walking model with improved Markov chain to simulate the time-based process of reputation of certain products. We noticed the relation between people's star ratings and the tendency of reputation change, and used the relation to model the product's reputation. Our simulation result showed that with the change of time, customers' average ratings within certain amount of time changes, and reputation increases or decreases with the same tendency of average ratings. Our simulation model can help you determine reputation situation of products in the future.

Combining text-based measures and rating-based measures to predict purchase behaviours of customers is not only possible, but achieves better results than using them alone. We used Conditional Random Field to combine the Naive Bayes Model and the K-Means Clustering model. Besides, we introduced time-based measures into the CRF model to solve small dataset problem. We achieved over 90 percent of precision on all 3 datasets. We strongly recommend our CRF model to you to predict the success or failure of your products based on reviews of customers.

We have determined the strong positive correlation between the sentiment of customers and their final purchase actions. Therefore, we analysed the later purchase rate of products with extreme star ratings and high useful votes(We call them "key reviews") in the early stage. We found products with positive key reviews tend to have far higher purchase rate and those with negative key reviews tend to have far lower purchase rate later for a while. Therefore, we conclude that customers tend to write positive reviews after seeing positive key reviews, and negative reviews after seeing negative key reviews.

Finally, we analysed relation between certain quality descriptors and extreme star ratings. We separately analysed 5 star and 1 star rating reviews in one dataset. We did POS tagging to reviews and extract adjectives. We then collected their fre-

quencies in 5 star and 1 star rating reviews. We found that the gap between frequencies of quality descriptors such as "great" and "disappointed" in two rating reviews is extremely wide. We conclude that positive quality descriptors tend to appear more in positive reviews, while negative quality descriptors tend to appear more in negative reviews.

After analysis, we provide answers to your concerns and offer prediction models and simulation models to help you make decisions in online business strategy and improve user experience of your products. We hope our help is useful to you and good luck in business!

Sincerely yours,

MCM 2020 Team 2008384

2 Introduction

2.1 Problem Statement

With the development of Internet, more people tend to accept shopping online. As a result, a great business opportunity emerges in the field of online shopping. Platforms like Amazon provides customers with opportunity to rate and review purchases. Star rating allows purchasers to express level of satisfaction with a product using a scale. Customers also can submit text-based reviews to express further opinions and information about the product. Besides, other customers can submit helpfulness ratings on these reviews. These ratings and reviews are of great value on business. Companies use these data to gain insights into the market. These data helps them decide where to participate in the market and the time of that participation. and the economic potential of product design features.

But the potential patterns of ratings, reviews and other data parameters are hard to identify. The relationships and meaningful patterns hide behind sophisticated time line, complex customer relations and emotions. Finding these patterns needs many data mining methods and machine learning techniques. We use multiple methods to find out relationships of purchase and other product related patterns within reviews and ratings.

2.2 Literature Review

E-commerce has accumulated a large number of data, which has a natural advantage in forecasting market changes. T. Gottfried[1] pointed out that big data helps many companies make great achievements in advertising marketing, business intelligence, etc. D.Judy[2] explained that the Internet has made great changes in marketing, which has promoted the emergence of new business models. J. Leon Zhao[3] thinks that data analysis will fundamentally change the traditional way of marketing based on market research, consumer behavior and product design. Big data analysis will drive e-commerce companies to change from business intelligence to big data intelligence, and explore the business opportunities brought by big data analysis.

Our team will study further on the basis of previous studies, and discuss the impact of customer feedbacks on product sales. We try to find out relationships of different rating ways and combine our models. We also analyse relationships of different dimension of reviews such as time scale to help recognize patterns. Based on our model and analysis, we answer the questions presented by the director and list practical suggestions to Sunshine Company.

3 Assumptions and Notations

3.1 Assumptions

- The change probability of product reputation obeys Markov property, the probability of next movement is only related to present probability and present situation.
- Sentiment of reviews is good or bad, which corresponds the verified purchase of yes and no.
- The reputation of products only moves next to its present level in a short period of time, which means it obeys the property of random walking.
- Customers' purchase actions are only related to their own sentiments on the product and other purchasers' reviews and ratings.

3.2 Notations

Symbol	Definition
$O_{t-N,t-1}$	Rate of purchase of reviewers on certain product.
μ_j	The cluster centroids of a K-Means cluster.
X_n	A process of random walking starting from position x.
S_t	The star rating of record t.
θ_t	The helpful vote rate of product t.
d_i	Distance from a star to cluster centroids i.
c_j	Sentiment classification of Naive Bayes model.
Precision	The precisoin value of cetain test.
Recall	The recall value of cetain test.
F1	The F1 measure of cetain test.

4 Data Preprocessing

Though structural data is available, we make necessary preprocessing to fit the need of use during the training of the models. We analysed the Yes and No ratio of verified purchase in the dataset. Note that customers records of positive verified purchase are positive and the rest are negative. The following is the statistical result:

	Positive	Negative	Number of Records
Hair dryer	0.855	0.145	11470
Pacifier	0.859	0.141	18937
Microwave	0.678	0.322	1615

According to the result, the positive records and negative records are highly unbalanced. Random dividend can lead to severely imbalanced dataset, which is a disaster to model training. To solve the problem, we split dataset into positive part and negative part, and extract records separately but randomly according to the ratio above. After the split process, positive and negative ratio in training and test set is the same. After the split process, we shuffle both training set and test set.

We also extract data from time dimension. We extract records that belong to same products and category them in small sets. We use these extracted data to analyse reputation changes with time from reviews and star ratings. We also use these data to analyse previous reviews and ratings' potential impact on later customers and their purchases.

For the Microwave dataset, though the record balance is better, the data size is rather small. Therefore, we consider using cross validation during the training process. We break the data into 6 folds, and in each fold we maintain the same ratio of positive and negative records like in figure 1 as the original dataset.

Figure 1: The process of cross validation

5 The Models

We build models to analyse and simulate the rating and review system, and hope to find patterns assisting future sell of products. We know the final goal of market analysis is to sell more products, which in our case means higher purchase rate(verified purchase rate of all reviewers). Therefore we build Naive Bayes, K-Means Clustering and CRF models to find the pattern in verified purchase. To model time based reputation pattern, we use random walking and improved Markov chain to simulate the changes.

5.1 Naive Bayes Sentiment Analysis Model

We first analyse the quantitative and qualitative patterns within the reviews on verified purchases. The purchase behaviours reflects in the dataset as binary action. Customers in the record either purchase or not. We consider the relation between review and purchases an sentiment analysis process. Positive emotion means purchase action of customers, while negative emotion means no purchase action. The datasets are relatively small, which means methods like deep learning are not to have great results. Therefore we consider naive bayes method for sentiment analysis.

5.1.1 Naive Bayes Classification

Naive Bayes Classification (NBC) is a method based on Bayes theorem and assumes the independence of feature conditions. First, through the given training set, the joint probability distribution from input to output is learnt based on the assumption that the feature words are independent. Then, based on the learnt model, the output with the maximum posterior probability is obtained by input. The following Bayes formula is the basis of the method:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$

We estimate prior probability P(X | Y) and classification probability P(Y) from training set. We suppose the words are independent and maximise the posterior probability P(Y | X) through the following formula:

$$\arg\max_{c_j} P(c_j) \prod_{i \in positions} P(w_i|c_j)$$

The following is the main process of the algorithm:

- 1 Data processing. Extract review headline and review body from training set. Directly combine review headline and body as training data. Map the purchase condition to 0, 1
- 2 Build dictionary for training set. Extract words from training data and change words into lowercase. Remove irregular words, digits and other identifier form the dictionary.
- 3 Calculate prior probability P(X | Y) through training data. Each feature word calculates and stores 2 weights, one for positive records, the other for negative records.
- 4 Calculate classification probability using training data.
- 5 Test on test sets. Calculate precision, recall and F1 measure.

5.1.2 Cross validation

For Microwave dataset, we have divided the data into 6 folds during data preprocessing. During the training process, we orderly choose a fold as temporary test fold and compute performance on the set. The rest as training folds. Each fold has one time to be tested. In each epoch, new learnt weights are stored and tested on the chosen fold. Finally, we average the performance of 6 runs as the final result.

5.1.3 Test result

We calculate Precision, Recall, and F1 Measure evaluation on the test set. The following are the calculation methods of the four measurement:

We have the following definitions:

Prediction/label	Positive Sample	Negative Sample
Positive Sample	TP	FN
Negative Sample	FP	TN

Formulas:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Consider the importance of some adjectives, and adverbs in sentiment analysis, we don't use regular stop words removal. But we remove all emojis and irregular words, and we only keep punctuations full stop, exclamation mark and question mark, which we consider more important in sentiment analysis. We test on Pacifier, hair dryer and Microwave dataset. On Microwave dataset we use cross validation. The following is the test result:

	Precision	Recall	F1 measure
Pacifier	0.937	0.842	0.887
Hair dryer	0.935	0.767	0.843
Microwave	0.832	0.737	0.782

5.2 K-Means Clustering on Star Rating

5.2.1 Correlation Factor

Customers' purchase intention is often closely related to their impression of products.According to the preprocessing results of the data, the data does not conform to normal distribution. Because the rating data is a kind of ordered data, and quality data only reflects the order relationship of the observation object, we use Spearman correlation coefficient to do correlation classification. We do correlation analysis to ratings and verified purchase. The result shows that in the case of 99 percent confidence interval, it's likely that there is a strong correlation between rating and purchase. Therefore, consider that people's impression on a product is strongly related to purchase situation, we can use clustering to further mine their relationship.

5.2.2 K-Means Clustering

K-means clustering is a vector quantization method. K-Means randomly selects K objects as the initial clustering center, then calculate the distance between each object and seed clustering center, and assign the object to the nearest clustering center. Each clustering center and objects assigned to them represent a cluster. The following is the main process of K-Means clustering algorithm:

- 1 Randomly select k clustering centroids for $\mu_1, \mu_2, ..., \mu_k \in R^n$
- 2 Repeat the following procedures until convergence:

  For each sample i, determine the cluster it belongs to. Minimise the following formula:

$$c^{(i)} := \arg\min_j ||x^{(j)} - \mu_j||^2$$

  For each cluster j, recalculate its cluster centroids. The following is the formula:

$$\mu_j := \frac{\Sigma_{i=1}^m 1(c^{(j)} = j)x^{(i)}}{\Sigma_{i=1}^m 1(c^{(j)} = j)}$$

5.2.3 Clustering Result

We cluster on all three datasets. We have two purchase result, therefore we set K = 2 to cluster customers into 2 categories. We get the following cluster center results:

Pacifier	1	2	Hair dryer	1	2
Rating center	4.832	2.063	Rating center	4.772	1.981

Analysing the result, we can roughly divide the customer's ratings into two categories: Level 4 and level 5 represent that the customer has higher purchase intention for the product. level 1, level 2 and level 3 represent that the customer's purchase intention for the product is lower. The following is the final results on test set:

	Precision	Recall	F1 measure
Pacifier	0.901	0.812	0.855
Hair dryer	0.894	0.767	0.826
Microwave	0.858	0.756	0.805

5.3 Combine Random Walking and Improved Markov Chain

In the models we construct above, we don't consider time dimension of reviews and star ratings. However, time plays important role in reputation tendency of products. Past reviews can influence future reviews and lead to increasing or decreasing of product reputation. We use random walking to model changes of reputation of certain products. However, it's clear that the probability of increasing and decreasing each time the reputation change is not a constant, but changes with the ratings of customers. Therefore, we use improved Markov chain to model the variation of transition probability. Inspired by Markov process changing with time, we improve Markov chain by changing the constant transition probability.

5.3.1 Random Walking

Random walking is an random model formed by a series of random steps. Random walking can be in line, plane, surface, high-dimensional space, graphics or user group and other random processes, and it's often changing with time parameters. In our case, We consider random walking in discrete time. 1-dimensional random walking is usually parts and sequences of Bernoulli test structures. If Y_k is a Bernoulli series and $Y_k \in \{1, -1\}$ we suppose

$$P(Y_k = 1) = p, P(Y_k = -1) = 1 - p$$

then $X_n = x + \Sigma^n Y_k$ is a 1-dimensional random walking from x. In our case, p changes with time.

5.3.2 Markov Chain

Markov chain is a set of discrete random variables with Markov properties. If the values of random variables are all in the countable set, and the conditional probability of random variable satisfies:

$$p(X_{t+1}|X_t, ..., X_1) = p(X_{t+1}|X_t)$$

then X is a 1-Order Markov Chain. At each step of Markov chain, the system can change from one state to another or keep the current state according to the probability distribution. The change of state is called transition, and the probability related to different state changes is called transition probability. Random walk is an example of Markov chain. The state of each step in random walk is a point in the graph. Each step can move to any adjacent point.

5.3.3 Model Construction

We discretize reputations of products and set reputation levels from -30 to 30. We set original reputation level to 0. reputation level below 0 means bad reputation, and reputation level above 0 means good reputation. Compared to reviews, star ratings more directly reflect satisfaction or disappointment of customers, and it directly lead to changes of reputation of products. Therefore, we model changes of reputation by star ratings.

We set time unit for random walking to 1 month. Considering changes of reputation is gradual and unlikely to fall or increase rapidly, we use 1-dimensional random walking and set step to 1. Each month after modifying transition weight, we simulate the reputation of the product to increase by 1 level or decrease by 1 level. Random numbers subject to uniform distribution in scale 0 to 1 are used to simulate the process.

We set time unit for Markov Chain to each time there is new ratings on the product, so we skip real time intervals. This means each time a new ratings is posted, we modify transition weights, until end of the month, when modified weights are used to renew random walking. After that, we restore to original weights to start transition modifying of next month.

We have the following original distribution at the start of each month:

$$\begin{Bmatrix} 0.5 & 0.5 \end{Bmatrix}$$

The distribution means at first probability of increasing and decreasing are the same. If S_t denotes the star rating of record t, according to experiments and experience, we define the following transition matrix:

$$\begin{Bmatrix} 0.6 + 0.05 * S_t & 0.4 - 0.05 * S_t \\ 0.3 + 0.05 * S_t & 0.7 - 0.05 * S_t \end{Bmatrix}$$

This definition enables higher star ratings to improve the probability of increasing levels, and lower star ratings to decrease the probability. In each month, original probability goes through all different transition matrix, one matrix per record.

We also notice the helpful votes and total votes of records. For $record \in \{s|s[totalvotes] > 0, s \in records\}$, we define helpful vote rate θ_t:

$$\theta_t = \frac{records[helpfulvotes]}{records[totalvotes]}$$

Through observation we notice that records t with higher θ_t value tend to have higher influence on the reputation of the product. Therefore, we define the following formula to take votes into consideration. We modify star ratings according to helpful vote rate θ_t:

For $S_t \in \{1, 2\}$, we have

$$S_t = \begin{cases} S_t - 0.05 * \theta_t & \theta_t > 0 \\ S_t + 0.5 & \theta_t = 0 \\ S_t & else \end{cases}$$

For $S_t \in \{3, 4\}$, we have

$$S_t = \begin{cases} S_t + 0.05 * \theta_t & \theta_t > 0 \\ S_t - 0.5 & \theta_t = 0 \\ S_t & else \end{cases}$$

Star ratings with high θ_t will have more impact on the modifying of transition weights, and low θ_t ratings will have less effect on the modifying process.

5.3.4 Simulation

We use time-based data processed in the preprocessing process. We calculate numbers of records for every product and find the product with the most records from all three datasets. We extract these records and form three datasets. And we simulate in the following process:

- 1 Initialize random walking model and Improved Markov chain.

- 2 While there is unused data, repeat steps 3-6
- 3 Extract the earliest unused month's data, form a subset U. Mark data in U used.
- 4 Use dataset U to iterate Markov chain.
- 5 Use iterated probability of Markov chain to simulate and record a new step of random walking model.
- 6 Restore Markov chain to original mode.

Figure 2 is a line chart of one time of simulation on all three datasets. This is the result of product 423421857(Parent number) in microwave dataset. Two other simulation results are in appendix.

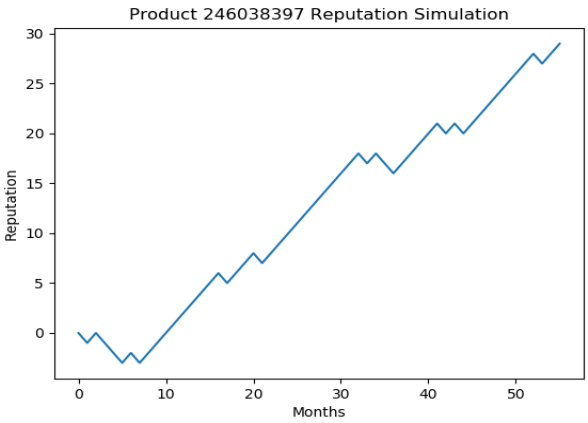

Figure 2: Simulation result of one of the Pacifier products.

5.4 Multidimensional Conditional Random Fields

We constructed Naive Bayes model and K-Means model to find patterns in purchase. However, the models are built on all products, which is not fit for analysis of single product's potential success or failure. Besides, we modelled time based patterns by random walking. We consider building Naive Bayes and K-Means model on single product's reviews, and combine reviews, star ratings and time-based features to predict purchase situations of single product. Therefore, we build Conditional Random Field to model features on instantaneous and timely scales.

5.4.1 Conditional Random Field

Conditional Random Field(CRF) is a kind of discriminant probability model, which is often used to label or analyse sequence data, such as natural language texts or biological sequences. CRF is whole-sequence conditional model rather

than a chaining of local models. It considers both sequence features such as previous output and present features. If we have present features d, and previous output c, we have

$$P(c|d,\lambda) = \frac{exp\Sigma_i \lambda_i f_i(c,d)}{\Sigma_{c'} exp\Sigma_i f_i(c',d)}$$

λs are learnt weights. The space of cs is the space of sequences. In special cases, if the features f_i remain local, the conditional sequence likelihood can be calculated exactly using dynamic programming.

5.4.2 Model Construction

We construct CRF model to predict purchase actions of certain products. We use time based data processed in data preprocessing section. We use training data to train the model on the time line of one product, and test on the time line after the training process.

To construct the model, we first choose analysis features. Our intention is to combine review based model and rating based model, and introduce time based features to the model. Therefore, we choose the following features for CRF:

- 1 Use the training data to train a Naive Bayes model. The model achieves over 90 percent of accuracy on the test set. We use the Bayes sentiment analysis model to map review sentiment to 0,1, and use the result as the first feature. We design the following function to map sentiment results to reasonable features. If sen_t denotes sentiment analysis result for review t, $sen_t \in \{0,1\}$, we have

$$f_1(sen_t) = \begin{cases} sen_t & Positive \quad result \\ 1 - sen_t & Negative \quad result \end{cases}$$

- 2 Use the training data to do K-Means clustering to divide data into two clusters(purchase or not).Then calculate the clustering center. If S_t is the star rating of record t, and clustering center is $\{c_1, c_2\}$, we have distance $\{d_1, d_2\}$ satisfy:

$$d_i = |S_t - c_i|$$

We use the distance towards clustering center as the second feature. After clustering, we have the following clustering center:

Center/Dataset	Pacifier	Hair Dryer	Microwave
c_1	2.152	2.172	1.917
c_2	4.861	4.763	4.611

- 3 We also consider time based features. If O_t denotes the output of record No.t, we define purchase rate $O_{t-N,t-1}$, which satisfies:

$$O_{t-N,t-1} = \frac{\Sigma_{i=t-N}^{t-1} O_i}{N}$$

N is a hyper parameter. Purchase rate denotes in a range of reviews, how many reviewers finally purchase the product. In this case, if purchase rate is high in the near past of reviewer t, the reviewer tends to purchase the product. The higher purchase rate is, the higher the tendency. Therefore, we use N nearest purchase rate $O_{t-N,t-1}$ as the final feature. Obviously, $O_{t-N,t-1} \in [0, 1]$. Same with Naive Bayes Model, we use the same function to map purchase rate to reasonable feature. we have

$$f_3(O_{t-N,t-1}) = \begin{cases} O_{t-N,t-1} & Positive \quad result \\ 1 - O_{t-N,t-1} & Negative \quad result \end{cases}$$

The main structure of the CRF model is in figure 3.

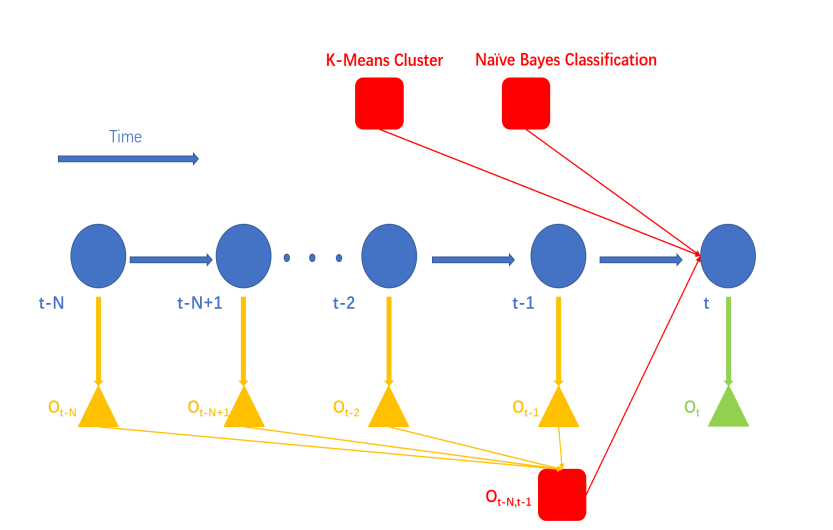

Figure 3: Main structure of the CRF model.

We combine Naive Bayes based sentiment analysis model, K-Means clustering model and N nearest purchase rate model. We use gradient descent to learn parameters λ_i.

5.4.3 Test result

We choose the product that has most reviews in all three divided datasets. We train a new Naive Bayes model on them. We use K-Means clustering on them and find cluster center. Then we build CRF on the models. We set N to 10.

We separately train on 3 dataset, and the following is the learnt weight:

Dataset/Parameters	λ_1	λ_2	λ_3
Microwave	1.00	0.34	0.90
Pacifier	1.20	0.32	0.89
hair dryer	1.05	0.31	0.87

The following is the test result of three datasets.

Dataset/Measurement	Precision	Recall	F1 measure
Microwave	0.893	0.903	0.895
Pacifier	0.946	0.935	0.940
hair dryer	0.926	0.975	0.920

6 Model Analysis

6.1 Informative Comparison of Reviews and Ratings

We calculate Precision, Recall and F1 Measure to compare Naive Bayes sentiment analysis model and K-Means Clustering model on three datasets. After the calculation of the evaluations, the results of the three datasets are summarized in figure 4. We have results on Pacifier set here and the other two results in appendix.

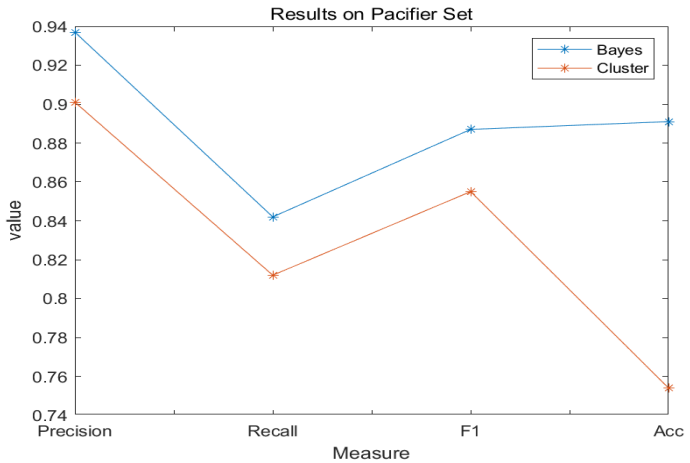

Figure 4: Result of evaluations of both methods on Pacifier dataset

Through the analysis of evaluations, we can tell from the figure that most evaluation scores of Naive Bayes sentiment analysis is higher than those of K-Means clustering. Therefore, we conclude that measures based on text reviews is more informative and complete than measures based on star ratings, and reviews are more accurate and effective in tracking customers' attitudes and purchase actions on products. We analyse main reasons for this advantage and following is the result:

- The amount of information reflected by reviews is far greater than that of the star rating, and customers express more detailed and true feelings and emotions about the product, which is crucial to purchase.
- Through the analysis of the data, we found that in some records reviews and star ratings don't match well(high review, low star,for example), while purchase is more related to reviews of the customers. A possible reason is that star rating is too simplified, and they may be interfered by other customer's ratings in some cases. However, when using words to describe it, they express their own true feelings. Therefore, reviews are more suitable to identify customers' true attitude to the product.

We notice that in Microwave dataset, the advantage of sentiment analysis is not obvious. A simulation result is in figure 5.

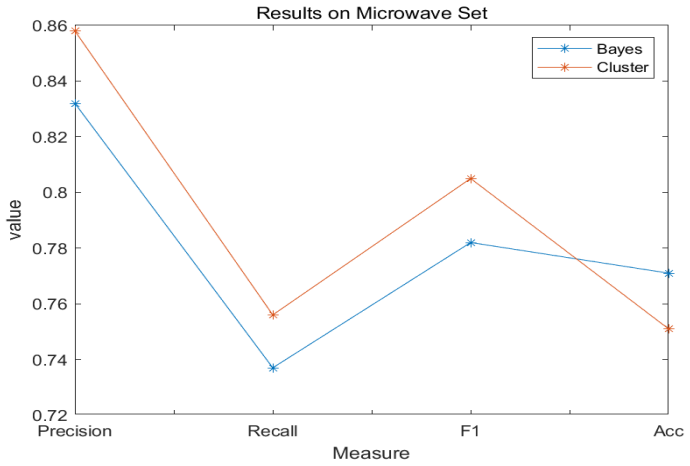

Figure 5: Result of evaluations of both methods on Microwave dataset

We can see that on Microwave dataset Naive Bayes model doesn't outperform K-Means model as on the other two datasets. We analyse the reason and notice that the size of Microwave dataset is smaller than the other two sets. Therefore, we consider the dataset may have imbalance problems when it is divided into 6 folds in cross validation. Small datasets are also not enough for Naive Bayes model to mine sentiment patterns. Considering the situation, the exception in Microwave dataset is not convincing to deny the advantage of reviews over star ratings on analysing purchase patterns.

6.2 Time Based Analysis on Product Reputation

We combine 1-dimension random walking with improved Markov chain to model time based measures and patterns in last section. We use probability transition to model the increasing and decreasing tendency of reputation. According to our simulation result, most of the time the reputation tend to increase. We recorded star ratings of the products, and up to 70 percent of the rating are above 3 stars. Therefore, the simulation result fits our intuition. However, we need more specific proof to prove the validity of our model.

We record all star ratings in each month and average them. We suppose that the average result denotes upcoming tendency of reputation. If average star ratings is low, we expect upcoming tendency to decrease. Otherwise the tendency is expected to increase. Note that the change of reputation is expected to delay, following the change of average star ratings. If the test matches our expectation, the model is considered successful. Also, maintaining high/low level of star ratings is expected to come with steady growth/decrease of reputation of products.

We test our model and assumptions on all three datasets. And figure 6 is one of the results on microwave dataset. We choose result on the product of microwave dataset because the dataset is smaller, and the time range is only about 25 months, which is easier to analyse. We can see from the figure that in fourth month there is an apparent decrease in average star rating, and after that there is a long term decrease in reputation of the product. After 13th month, the average star ratings maintain high level, and the reputation continues increasing, and the growth rate remains steady. Therefore, the simulation result matches our assumption.

We use random walking and Markov chain to find time-based patterns in changes of product reputation. The test suggests that our model is reasonable and successful.

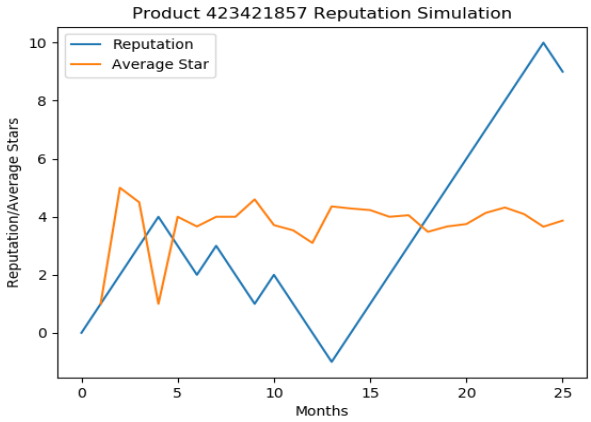

Figure 6: Simulation and average star tendency diagram on Microwave dataset.

6.3 Combination of Reviews and Ratings

In The Models section, we combined our model of text-based reviews(Naive Bayes sentiment analysis) and model of rating based reviews(K-Means Clustering). We used a CRF to combine both of them, and additionally combined time-based features. We use the CRF model to predict the purchases of reviewers, which is crucial in product selling, and reflects the potential success or failure of the product. We compare valuations of the 3 models. The comparison result of Pacifier dataset is in figure 7.

We can see from the figure, after combined with time based measure, Preci-

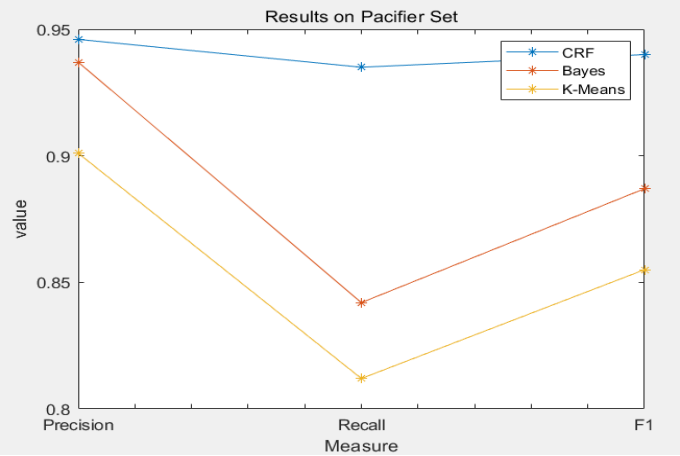

Figure 7: Result on Pacifier set.

sion increases and Recall measure shows great improvement. The result proves that the combination of reviews and ratings works well. The CRF model can help determine future situation of the product, then help the company make adjustments on time.

What's more, we compare results on small dataset Microwave, and the result is in figure 8.

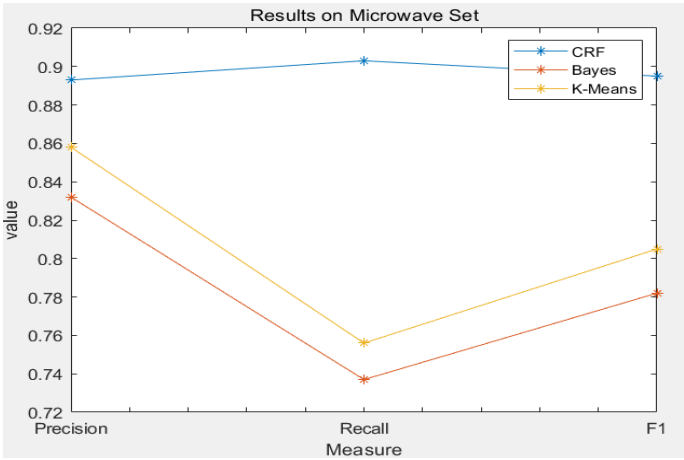

Figure 8: Result on Microwave set.

We can see that precision and recall value both shows great improvement. The combination of reviews and ratings and the introduction of time-based method works well on small dataset, which is a problem to Naive Bayes model. The improvement is greater than on the other two dataset, therefore we speculate the introduction of time-based model is the key to solve the problem of small dataset.

6.4 Customers Mutual Influences Analysis

We analyse mutual influences between customers. When we analyse time based models, we find customer rating levels are often clustered, which mean-

s high star ratings and lower star ratings do not often appear near each other. Therefore we speculate mutual influences in reviews.

We do sentiment analysis using Naive Bayes in former section. According to the result, we find strong relations between review sentiment and final purchase situation. The test result shows we have 0.875 of accuracy on hair dryer dataset and higher on Pacifier set, indicating that purchase is in high positive correlation with sentiment of reviews. Therefore, instead of directly analysing review sentiment of customers, we analyse early stage reviews' influence on later purchase rate.

We analyse hair dryer dataset. We choose 14 products that have over 100 sells and 90 percent of purchase rate(reviewers purchased the product),and analyse them. We find that 12 of them received high ratings with high helpful votes in early stage of selling. These hot products generally observe the law that the more great reviews with high helpful votes they possess, the higher the purchase rate of customers. For example, we collect data about best selling product B00132ZG3U and B0009XH6TG:

Product	sell number	purchase rate
B00132ZG3U	506	94.58%
B0009XH6TG	510	91.89%

We find they both possess high rating reviews with over 400 helpful votes, which supports our analysis.

We also analyse the worst selling products. We find products with sell number over 10 but with purchase rate below 0.6, and we find they mostly possess many negative reviews with high helpful votes in the early stage. For example, product B000LQB5YS possesses one negative review with 304 helpful votes, and has purchase rate of 59.37%. Product B000BFJJ7E has almost all negative review in the early stage, and it only has purchase rate of 38.77%. We extract some products with positive reviews and high helpful votes, and products with negative reviews and high helpful votes. Then we use scatter diagram to model relations between purchase rate and early stage reviews. The results are in figure 9.

From the diagram we can see that most products with low early stage review evaluation level tend to have lower purchase rate. And most products with high early stage review evaluation level tend to have higher purchase rate. Therefore, we can conclude that early stage reviews can influence reviews and attitudes of later customers, and influence purchase rate of the products.

6.5 Quality Descriptors Analysis

Reviews are evaluations and impressions on products of customers. Therefore quality descriptors are important parts of reviews. Extracting quality descriptors and analyse their relations with products and ratings is essential to understand

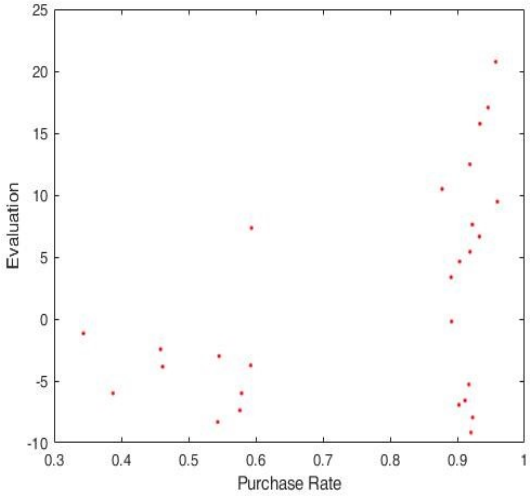

Figure 9: Relations between reviews and purchase rate.

customers' feelings and reactions. We analyse different kinds of quality descriptors and their relations to star ratings.

To analyse quality descriptors strongly associated with ratings, we extract reviews with star rating 5 and 1, which is obviously distinct in ratings and more representative. We call them polar reviews. We also notice that quality descriptors are mostly adjectives. Therefore we use NLTK(To know more about NLTK tool kit, please visit website: http://www.nltk.org/) tool to do word segmentation task and POS tagging task to both review headlines and review bodies of polar reviews. We use POS tagging sequences to extract adjectives from polar reviews. To avoid influence of different product categories, we separately analyse three datasets.In this analysis, we present the analysis process of Pacifier dataset.

We record top 20 frequency adjectives from both 5 star polar reviews and 1 star polar reviews. Then we choose quality descriptors such as "Great", "horrible", and compare their frequency in 5 star polar reviews and 1 star polar reviews. We select 6 words in top 20 adjectives from both results in two polars. Note that the size of 5 star reviews and 1 star reviews are different, we reasonably improve the frequency of quality descriptors in 1 star reviews. The result is in figure 10.

We can see from the histogram that positive quality descriptors such as 'Great', 'Perfect' appears far more frequent in 5 star reviews than in 1 star reviews. On the other hand, negative descriptors such as 'horrible', 'disappointed' appears far more frequent in 1 star reviews.

According to the analysis, we conclude that for specific quality descriptors, especially descriptions of extent in satisfaction or disappointment, their frequencies in reviews are strongly related to rating levels. Positive descriptors appear more in high rating reviews, and negative descriptors appear more in low rating reviews.

However, quality descriptors related to specific user experience of certain product, such as "hot", "wet" for hair dryers, are not strongly related to rating

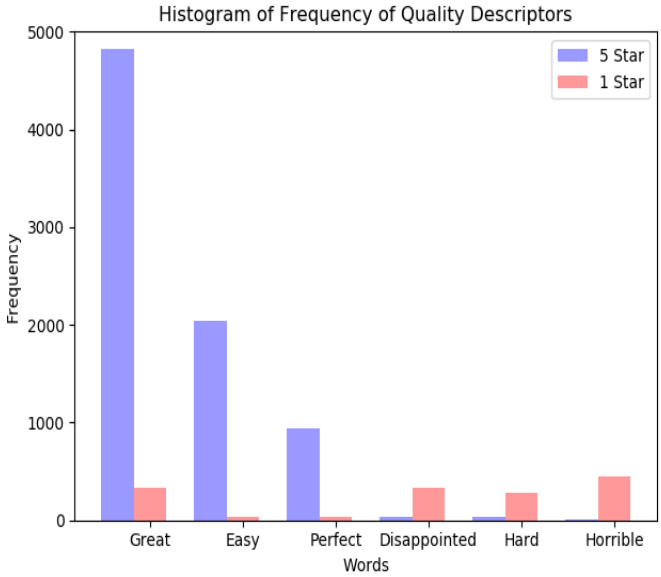

Figure 10: Frequency of quality descriptors in polar reviews.

levels. We analyse the reason is that users with both high and low ratings tend to describe specific experience. Following is an example from pacifier dataset:

- It's small but the kids don't care... Star:5
- ...with a small brush in it, which drives me crazy... Star:1

We can see that the Review with 5 stars and the Review of 1 star both used the descriptor "small", but they describe completely different sentiments. Therefore, the frequencies of this kind of quality descriptors don't have much difference in high rating and low rating reviews.

7 Strengths and Weaknesses

7.1 Strengths

- **effectiveness.** The CRF model achieves over 90 percent of precision and F1 measure on Pacifier and Hair Dryer dataset, and on small dataset Microwave it achieve over 85 percent of precision and F1 measure. High precision on both big and small datasets make the CRF model reliable in product purchase prediction.
- **Efficiency.** The training of CRF model, including the training of Naive Bayes model and K-Means model on Pacifier dataset completes within 5 minutes. Besides, the simulation of reputation by random walking and Markov chain on Pacifier dataset over 50 months completes within 1 minutes. The models are very efficient,and easy to use in business analysis.

7.2 Weaknesses

- **Over simplification.** We map the sentiment of reviews into 0, 1, which is over simplified. Sentiment of customers is far more complicated than binary dividend. Therefore, practical application may not be effective as we are in the idealized environment.
- **Data Lacking Dilemma** When the dataset is small, the Naive Bayes model and k-Means Clustering model don't work well as they do with big dataset. When time data of reviews is sparse, the random walking model doesn't work well.

8 Conclusion

Our team has completed the tasks given by the sunshine company marketing director. we used Naive Bayes Model to deal with sentiment of reviews and K-Means Clustering to cluster and analyse star rating data.We used conditional random fields to combine ratings and reviews information. We also introduced time-based information to CRF, which improved the result, especially on the small Microwave dataset. To model product reputation, we combine random walking and improved Markov chain to simulate the reputation changing process. We proved the validity of our models by comparing the simulation with the star rating changes of the product. Using these models, we analysed questions of the director, and made proper answers to them. Finally, we analysed texts of reviews, and found out relations between quality descriptors and star ratings. With all these analysis and model construction, we made reasonable suggestions to the director and offer tools of product analysis.

References

[1] big data: new driving force in business and other intelligent fields, Elisa T. Gottfried, Oscar T., 2005(18):105-134.

[2] Internet marketing, D. Judy, 2010.

[3] Research and analysis of challenges and management faced by enterprises in the era of big data, J. Leon Zhao.

[4] Minimax Entropy Principle and Its Applications to Texture Modeling. S. C. Zhu, Y.N. Wu and D.B. Mumford, Neural Computation Vol. 9, no 8, pp 1627-1660, Nov. 1997.

[5] Discriminative Random Fields, Sanjiv Kumar. Martial Hebert. International Journal of Computer Vision. June 2006, Volume 68, Issue 2

[6] Sentiment analysis: A combined approach, Rudy Prabowo, Mike Thelwall, Journal of Informetrics

[7] Sentiment analysis algorithms and applications: A survey, Walaa Medhat, Ahmed Hassan, Ain Shams Engineering Journal

[8] Markov Chain Sampling Methods for Dirichlet Process Mixture Models, R.M Neal, Journal of Computational and Graphical Statistics

[9] Algorithms for Random Generation and Counting: A Markov Chain Approach, Sincliar

[10] Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier, Sanjay Chakraborty, Lopamudra Dey

9 Appendix

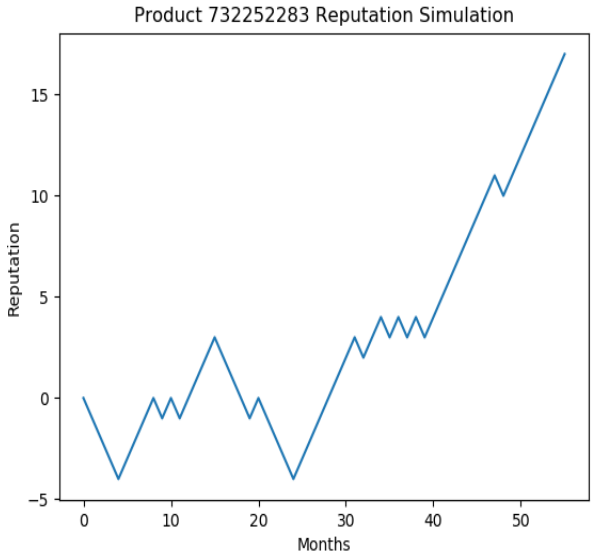

Figure 11: Random walking Simulation Result on Hair dryer dataset.

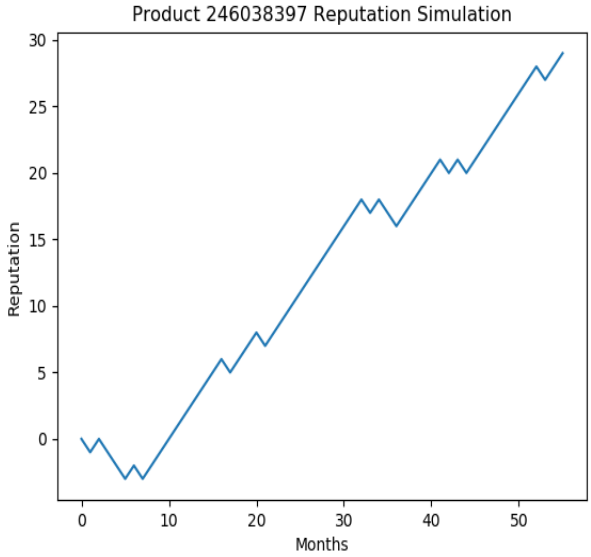

Figure 12: Random walking Simulation Result on Pacifier dataset.

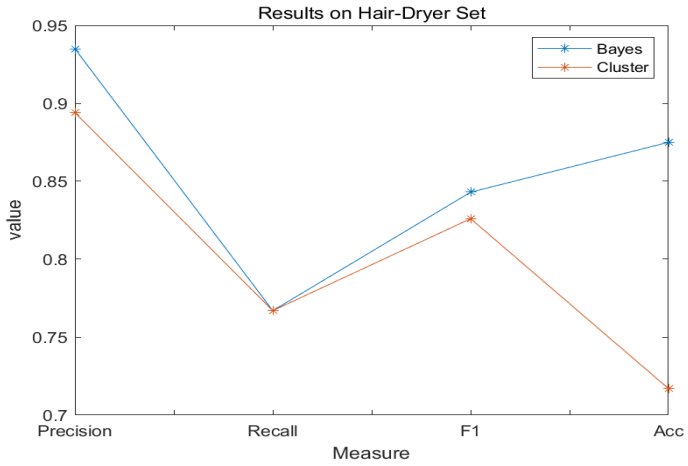

Figure 13: Result of evaluations of both methods on Hair Dryer dataset

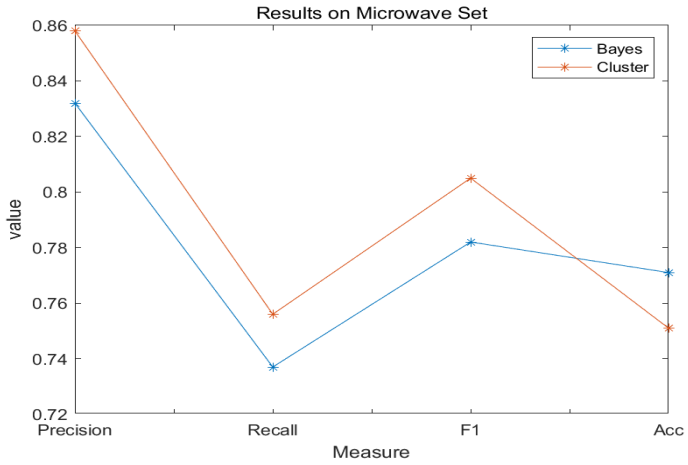

Figure 14: Result of evaluations of both methods on Microwave dataset