For office use only	Team Control Number	For office use only
T1 _____	**2022107**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**C**	F4 _____

2020
MCM/ICM
Summary Sheet

Analysis of Amazon Customer Reviews and Implication of Sales Strategies

Summary

The goal of our analysis is to identify key patterns of past customer reviews to identify the most effective sale strategies.

First, we perform exploratory data analysis to obtain basic statistics of the given data sets and set the foundation for the further analysis. We construct correlation plot between $star\_rating$, $vine$, $verified\_purchase$, and $helpful\_votes$, and use inferential tests to determine the significance of each variables.

Second, we apply sentiment analysis in NLP to measure the positive or negative level of a customer's review. Then we use bigram and network analysis to identify the most frequently mentioned or most associated words in the review body, calculating keywords frequency in each comment. With the above quantified criteria, we apply SVM to train the data to identify a relationship between a review's helpfulness and its sentiment, word count, and keywords. In addition, we discussed the special cases with updated reviews.

Third, we construct time-based models to identify the trend of a product's popularity and reputation over time. The result suggests that several consecutive high-star ratings tend to be followed by lower-star ratings. We then forecast future trend using ARIMA model and Exponential Smoothing, with basic assumption being that the past information should be weighted less than the latest events.

Finally, we adopt unsupervised learning. We employ PCA to find the best linear combination of variables which capture the most variability of the data, which is used to classify whether a product is successful or failing. We also apply ordinal logistic regression to predict the potential star ratings of reviews based on their sentiment scores.

Keywords: Unsupervised Learning; Principal Component Analysis; Time Series Analysis; Support Vector Machine; Data Visualization

Analysis of Amazon Customer Reviews and Implication of Sales Strategies

March 9, 2020

Summary

The goal of our analysis is to identify key patterns of past customer reviews to identify the most effective sale strategies.

First, we perform exploratory data analysis to obtain basic statistics of the given data sets and set the foundation for the further analysis. We construct correlation plot between $star\_rating$, $vine$, $verified\_purchase$, and $helpful\_votes$, and use inferential tests to determine the significance of each variables.

Second, we apply sentiment analysis in NLP to measure the positive or negative level of a customer's review. Then we use bigram and network analysis to identify the most frequently mentioned or most associated words in the review body, calculating keywords frequency in each comment. With the above quantified criteria, we apply SVM to train the data to identify a relationship between a review's helpfulness and its sentiment, word count, and keywords. In addition, we discussed the special cases with updated reviews.

Third, we construct time-based models to identify the trend of a product's popularity and reputation over time. The result suggests that several consecutive high-star ratings tend to be followed by lower-star ratings. We then forecast future trend using ARIMA model and Exponential Smoothing, with basic assumption being that the past information should be weighted less than the latest events.

Finally, we adopt unsupervised learning. We employ PCA to find the best linear combination of variables which capture the most variability of the data, which is used to classify whether a product is successful or failing. We also apply ordinal logistic regression to predict the potential star ratings of reviews based on their sentiment scores.

Keywords: Unsupervised Learning; Principal Component Analysis; Time Series Analysis; Support Vector Machine; Data Visualization

Contents

1 Introduction

1.1 Background

With the increasing popularity of e-commerce, online marketing becomes a significant factor in competition between companies. Ratings and reviews serve as valuable references and criterions for potential costumers. they also function as a critical source for companies to improve existing products and gain competitive advantage, and potentially, strengthen market power.

In our case, Sunshine Company is particularly interested in the reviews of three products that they are planning to launch. By analyzing their competitive products, they can gain more insights about how to improve their product features to distinguish themselves and attract more customers. Amazon's online review is a mixture of rating levels and text-based reviews, Sunshine Company can gain useful information from both the numerical and text variables.

1.2 Overview

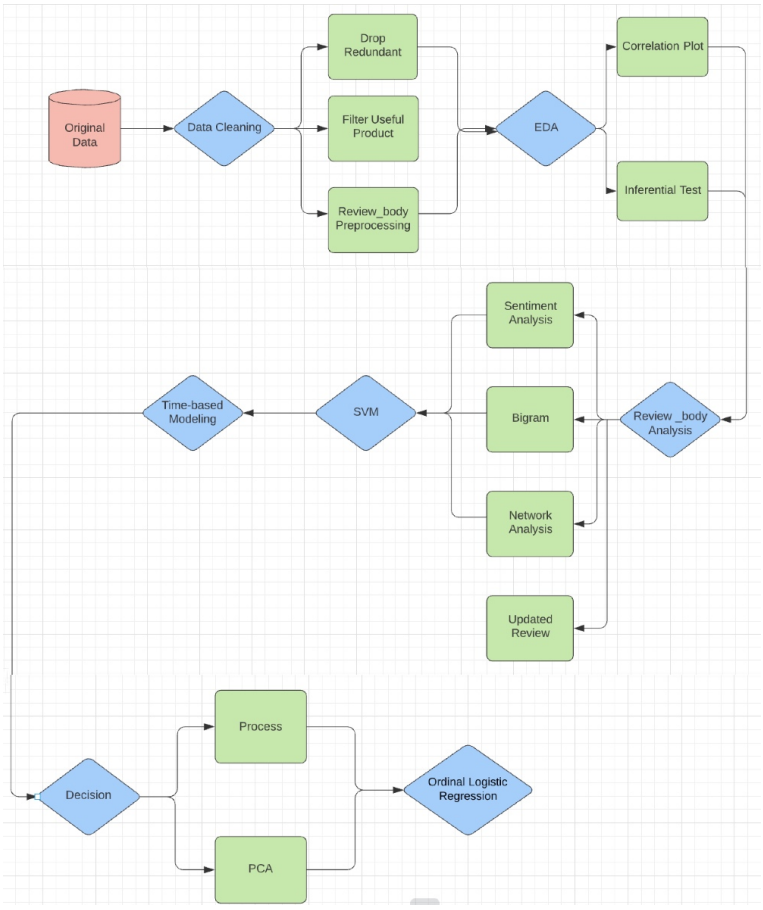


Figure 1: Workflow

2 Data Cleaning

To facilitate data analysis, we first drop redundant variables such as $marketplace$ and $review\_id$, and change important categorical features into numerical ones. To specify, $verified\_purchase$ and $vine$ are now numerical predictors with Y=1 and N=0. We also quantify helpfulness by introducing a new covariate called $help\_ratio$, which is the normalized difference between helpful and unhelpful votes.

$$\begin{aligned} X_{unhelpful} &= X_{total} - X_{helpful} \\ X_{net} &= X_{helpful} - X_{unhelpful} \\ X_{sd} &= \frac{X_{net} - \mu}{\sigma} \end{aligned} \tag{1}$$

As we further review with the data, it turns out that all purchases made by Amazon Vine members are not verified. This parallels with the real-world scenario that Amazon Vine members will receive free sample products from vendors or manufacturers so they can try the samples and write reviews. Some of them may even receive pre-release products as a part of the product's market testing before official launch to the public. Based on their reviews, the manufacturers and vendors may make certain adjustment of product features so that they can obtain more potential customers and market share. Since these Amazon Vine members do not necessarily need to buy the products themselves, their purchases will be classified as unverified yet they are still trusted voice selected by Amazon based on the number of their past helpful votes. However, for other unverified purchases, we believe that they may be the cases that the customers did not make purchases at all or they bought the products at with a discount which was offered because of some innate defects of the products. So, we exclude the data entries of non-Vine members with unverified purchases.

We then filter the products based on their number of reviews to obtain the most significant data. We decide to only use products with reviews above a certain threshold, the median of reviews among all products, since products with low number of reviews are not particularly suitable for generalizing hypothesis and models.

Table 1: Basic Statistics of Reviews

Product	Median	Mean	Standard Deviation
Hairdryer	44.00	24.25	121.73
Microwave	22.00	29.36	69.39
Pacifier	3.00	3.48	37.52

We also clean the $review\_body$, $review\_headline$, and $review\_date$ to facilitate the word frequency analysis and time series analysis which will be discussed in the following part. By deleting the extra space and punctuations in review

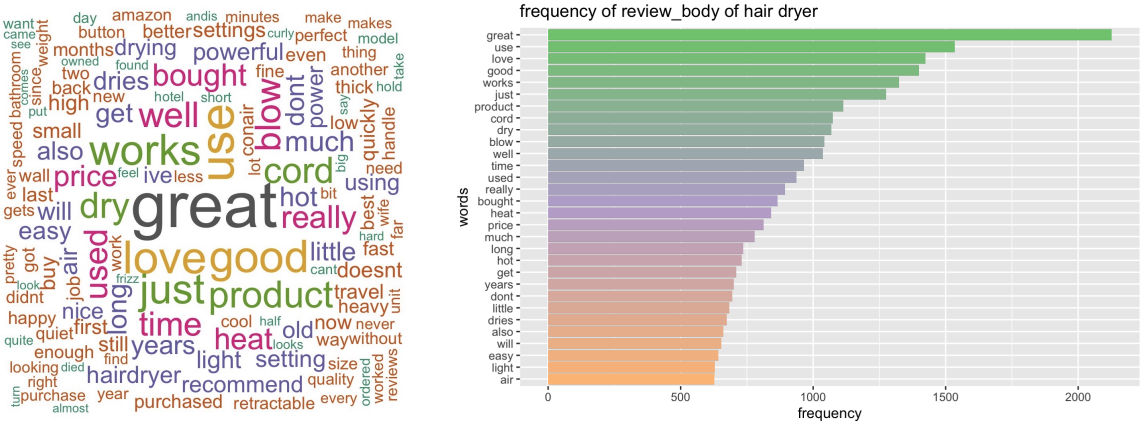
body and headline, we can obtain the word count of each review and used for further analysis, similarly, we change the format of review date to help plot the time-based graph. The following picture is an overview of our cleaned data set:

customer_id	product_parent	product_title	star_rating	vine	verified
34678741	732252283	remington ac2015 t studio	5	0	1
2282190	16483457	andis micro turbo hair dryer	5	0	1
...	...	...	...	...	...

review_body	sentiment_body	sentiment_headline	wordc	review_date	helpful_ratio
works great	0.5303300859	0.53033009	2	2015-08-31	-0.12130619
love this dryer	0.2638280940	0.00000000	3	2015-08-31	-0.19144853
...	...	...	...	...	...

Figure 2: Overview of Cleaned Data Set

Besides, we want to obtain an overview of the customers reviews which is the core factor in our analysis. By generating to the Word Frequency and Word Cloud plot, we can easily visualize the words frequently mentioned in the review body which provides a general idea of what customers care more about the product. For instance, Sunshine Company's Research and Development team should work on improving drying efficiency instead of beautifying the appearance of the hair dryer since it was rarely mentioned in the review.



(a) Word Cloud (b) Word Frequency

Figure 3: Review Overview

3 Exploratory Data Analysis

In this part, we aim to explore the interrelationship between certain variables and identify patterns in the given data sets which could provide instructive im-

plication.

3.1 Correlation Analysis

In order to identify the correlation between variables, we use the correlation plot which indicates a positive correlation between $star\_rating$ and upvote. Five-star rating has significantly larger upvote in 80 to 100 percent range which suggests that customers tend to find reviews with higher $star\_rating$ more useful. When promoting their products and doing market research, the company should focus more on good reviews to improve or characterize their products.

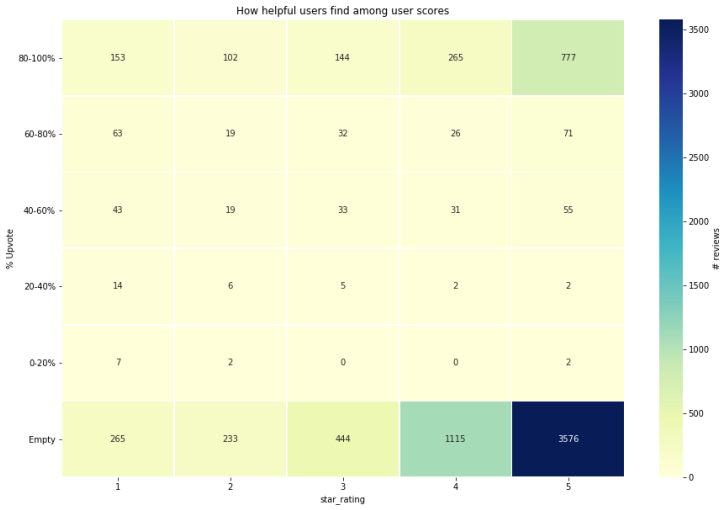


Figure 4: Correlation between star rating and upvote

$$Upvote = \frac{Helpful\ Votes}{Total\ Votes} \tag{2}$$

Moreover, we obtain the correlation plot among five numerical covariates: $verified\_purchase$, $vine$, $star\_rating$, and $help\_ratio$. The plot indicates a positive correlation between $verified\_purchase$-$vine$, and a negative correlation between $verified\_purchase$-$star\_rating$, $vine$-$star\_rating$. This relationship reveals important information about the impact of non-verified purchases, which may be the minor cases that the customers did not make purchases at all or they bought the products at with a discount which was offered because of some innate defects of the products. Therefore, these data entries would not be representative and may introduce bias and noise to our analysis.

$$Corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-EX)(Y-EY)^T]}{\sigma_X \sigma_Y} \tag{3}$$

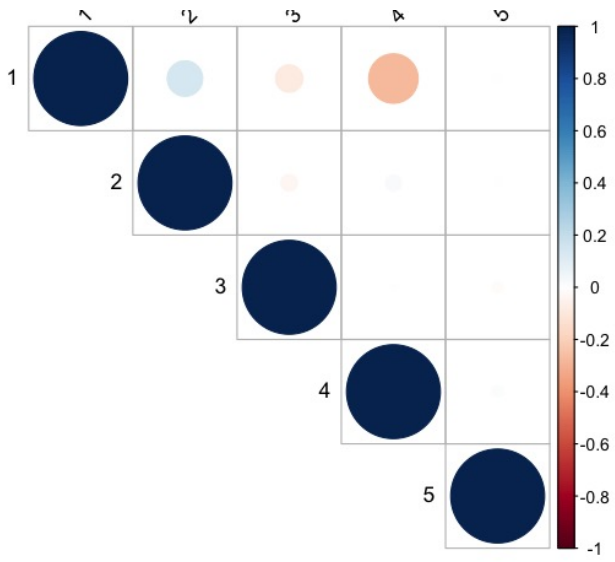


Figure 5: Correlation Plot among Numerical Covariates

3.2 Inferential Test

In addition, to find the significance of the key categorical variables on some numerical ones, we use t-test to determine whether there is a significant difference between the mean of two groups.

In our hair dryer example, we apply t-test to see the relationship between vine and non-vine, verified and non-verified customers' $star\_rating$ and $helpful\_ratio$. Between vine and non-vine customers, there exist yet not significant differences in $help\_ratio$ (p-value=0.7077); whereas the p-values for verified and non-verified groups are 2.2e-16 and 5.214e-09, which indicate that there exist significant differences in both $star\_rating$ and $helpful\_ratio$, with both covariates being lower in non-verified purchase. This result corresponds to our assumption that non-verified purchases tend to give bad reviews for the defected or less popular products with deep discount. Similar results can be obtained in the other two datasets of microwaves and pacifiers

$$\hat{\sigma}^2 = \frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}, \quad T = \frac{\bar{X}-\bar{Y}}{\hat{\sigma}\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \overset{d}{\sim} t_{n_X+n_Y-2} \tag{4}$$
$$\text{under } H_0: \ \mu_X = \mu_Y, \quad H_\alpha: \ \mu_X \neq \mu_Y$$

4 Quantified Customer Reviews and Implications

In this part, our main goal is to quantify customer reviews using different criteria to further explore instructive implication for Sunshine Company.

4.1 Sentiment Score

First, we create a new variable(column) by counting the number of words in each review, functioning as a basic criterion of the helpfulness of the review and its connection with star rating. Our main assumption is that the more words a review contains, the more relevant information it will provide, such as attractive features and potential defects.

Secondly, we employ sentiment analysis of the review which is a measurement of positive or negative level in the customer's review. Here we quantify the $review\_body$, $review\_headline$ by calculating their polarity. The unbounded polarity score δ is calculated as follows:

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,jn}}} \tag{5}$$

where $c'_{i,j}$ is the sum of weighted context clusters, $w_{i,jn}$ is the word count. The score 0 indicates neutral attitude, positive scores indicate relatively positive tone, and negative scores indicate relatively negative tone.

Table 2: Positive Ratio of Different Star Ratings

Star	Negative	Positive	Positive Ratio
1-star	1314	1211	0.4796
2-star	2716	1129	0.6119
3-star	990	2029	0.6721
4-star	1577	5171	0.7663
5-star	3257	15237	0.8239

Other than calculating the sentiment score of the review body, we also try to find the connection between of review's sentiment scores and its star rating. In the box plot of sentiment scores and star ratings, it is clear that the lower-star ratings tend to have more negative tone while the higher-star ratings have more positive tone which adheres to our general assumption.

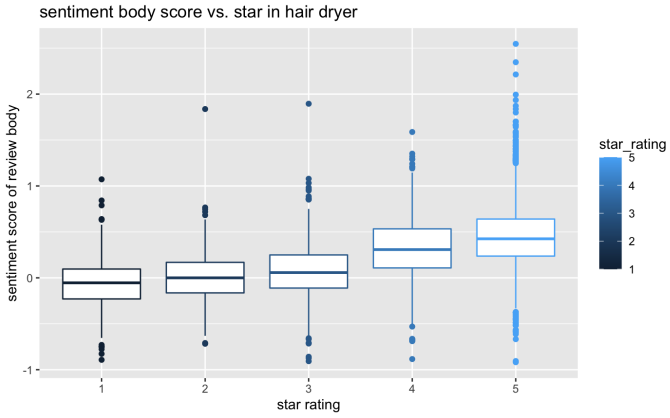


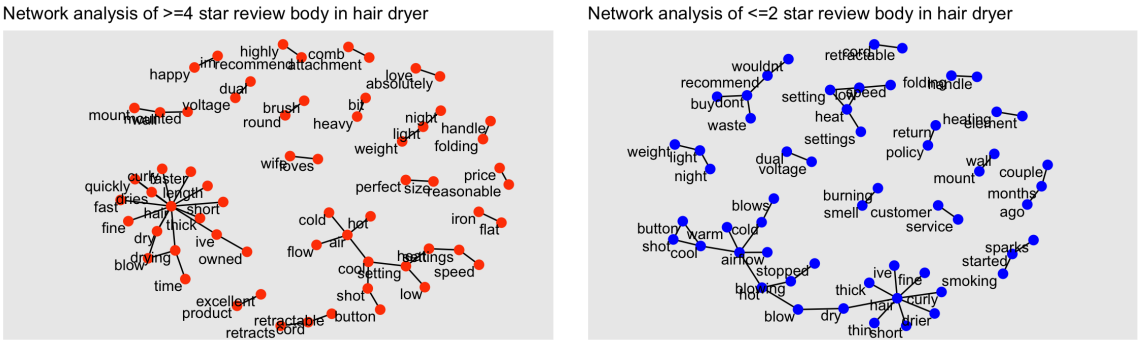
Figure 6: Sentiment Scores VS Star Ratings

4.2 Keywords Frequency

We also quantify *review_body* by counting the frequency of specific keywords that appears in the review body. The new criterion is used for measuring whether the review can be classified as an informative description of the product rather than simple, generalized comments. For example, reviews complaining about the retractable cord or the weight of a hairdryer are more informative and solid than comments like "awful product". After choosing keywords, the algorithm will count the number of occurrences of each keyword in the review body, and return the total number of occurrences for all keywords as the output. We notice that more than 50 percent of the reviews received 0 point, which makes sense as lots of reviews just express their temporary feelings without giving specific evidence as support. Thus, the reviews with 0 point are not useful for the Sunshine Company as a source of helpful feedback.

4.3 Network Analysis

Applying Network Analysis and Bigrams, we are able to identify the most associated keywords, which often reveals the most important product features that potential customers would look for. For example, we can see that such words include “blow”, “fast”, “airflow”, “battery”, “heat”, “retractable cord”, “light weight”, “thick hair”, and “curly hair” for hair dryer reviews. This would inform the Sunshine Company about hair dryers’ desirable features which could help the company increase their sales by embedding such features into their product design or highlighting the corresponding features on their product titles. At the same time, they can target customers’ major complains of their products to make corresponding improvement and make proactive upgrade of their customers service system by preparing to answers such FAQs instead of spending additional expense to do feedback surveys.



(a) Positive Feedback

(b) Negative Feedback

Figure 7: Network Analysis

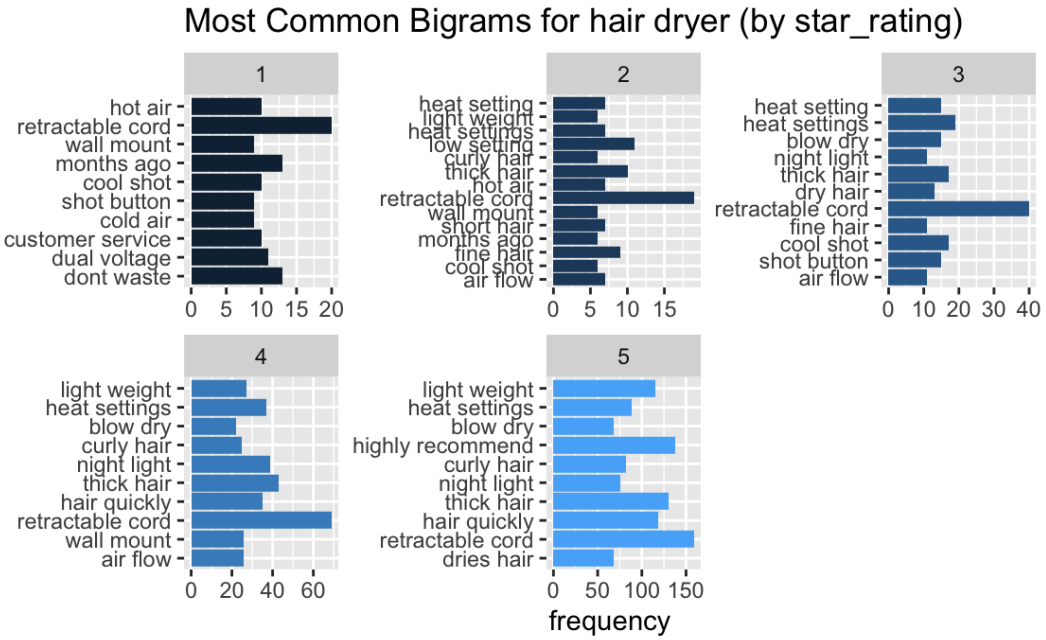


Figure 8: Bigrams of Customer Reviews

4.4 Support Vector Machine

Based on the three previous quantified criteria- word count, sentiment score, and keywords -that already provide information about the helpfulness of the review, we want to further explore whether we can apply these criteria to classify a review as helpful or not helpful without other customers' giving their votes for helpfulness. We employ Support Vector Machine(SVM) which is a supervised machine learning model that uses classification algorithms for two-group classification. The basic algorithm for SVM is to maximize the distance to the closest points from either class. Let γ denote the distance between two classes, we want to maximize γ while classifying individual point in the correct category.

$$\begin{aligned} &\max_{w,b,\|w\|=1} \gamma \\ &y_i(b + x_i^T w) \geq \gamma \end{aligned} \tag{6}$$

In our example, we want to use number of keywords, word count, sentiment score of review body, and sentiment score of review headline to determine whether a review is helpful or not. Using 10-fold cross-validation on our training set, we see that the best model is a SVM with radial kernel and parameters with cost=1, gamma=15, and cross-validation error=0.188. This model successfully classifies most individual data entries into the correct class in the test set.

The following picture shows the SVM classification for helpfulness using word count and sentiment scores on the test set. The light grey area is classified as helpful, and green as unhelpful. It is clear that for both classes, most

individuals are classified correctly (For example, orange data points represent helpful reviews, most of which are classified in the light grey area).

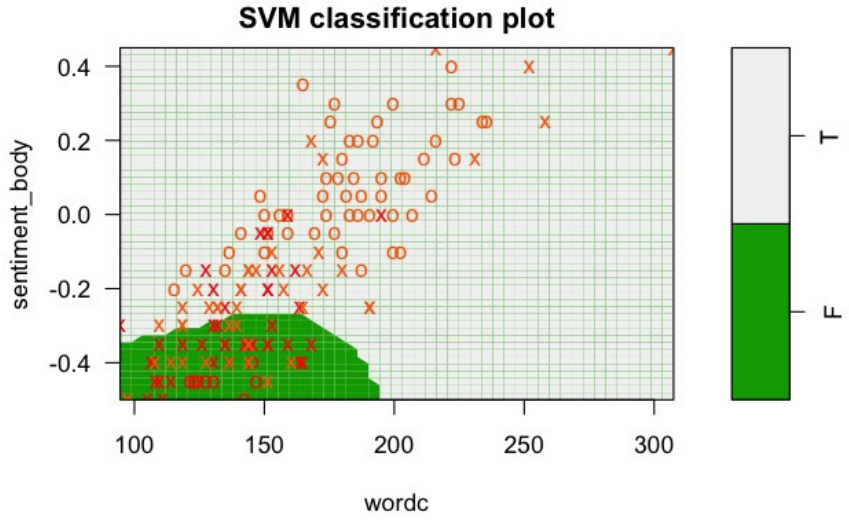


Figure 9: Support Vector Machine

With the help of this classifier, it suffices that the Sunshine Company can look at the important features of the review (sentiment score, word count, and keywords) to determine its helpfulness right after the review is posted. Instead of waiting for other customers to vote for the helpfulness of certain reviews, the Sunshine Company can make timely adjustments of their product based on the helpful reviews. Since the competitive edge of e-commerce lies in the prompt responses to customers buying decisions and feedback, this classifier can increase the chance of the Sunshine Company's success when launching their products to the online market.

4.5 Additional Implication

Furthermore, we notice some customers updated their initial reviews after using the products for a certain period. So, we filter out all items with updated reviews. Employing sentiment analysis to compare the overall sentiment on initial reviews and updated reviews, we find that the comments' positive-word ratio decreases, which indicates that the durability of these product might be terrible, or the defects of the products showed up later. Thus, Sunshine Company can pay more attention to these problems mentioned in updated reviews which can help them distinguish their products with competitors' by improving the durability and making certain advertisements focusing on the detailed yet important advantages of their products.

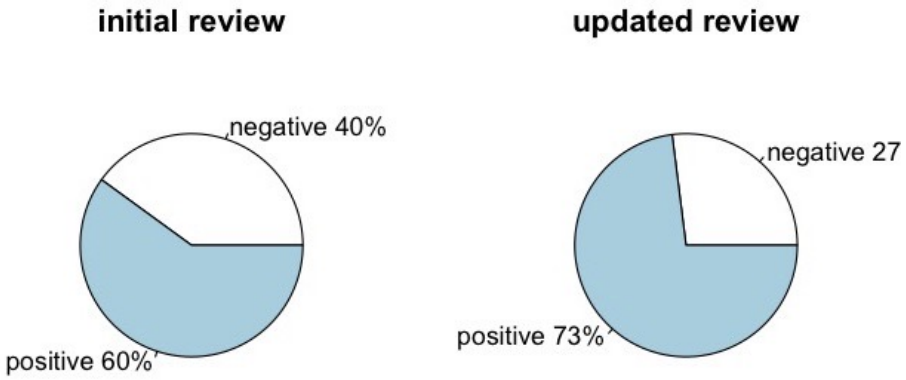


Figure 10: Sentiment Ratio of Initial and Updated Reviews

5 Time-based Modeling and Analysis

As an attempt to find the underlying time-based patterns in the star ratings of a certain product, we use the product with the highest sales and plot the corresponding time series. Since there is no obvious trend or seasonality that can be easily captured by observing the time series, we use the auto fit function in $forecast$ package to obtain the best model ARIMA(3,0,3) which is AutoRegressive Integrated Moving Average model consists of order 3 autoregressive model, 0 degree differencing, and order 3 moving average model. This function searches for a range of the ordered pair (p, q) values after fixing d by Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and chooses the model with the lowest AIC score.

$$y_t = \delta + \sum_{i=1}^{3} \phi_i y_{t-1} + \sum_{j=1}^{3} \theta_j \epsilon_{t-j} + \epsilon_t \tag{7}$$

The model also provides the prediction of the following 30 star ratings and corresponding confidence intervals. After repeating the process for the most popular hair dryers, most of them show similar trend that the overall tendency is downward which suggests the slight decreasing public reputation of these product although accompanying with frequent fluctuations among lower-star ratings and higher-star ratings. This adheres to the real-world scenarios that the constant competitions between similar products are fierce and customers have a higher requirement for a "good" product as the companies are all improving their products to gain market shares.

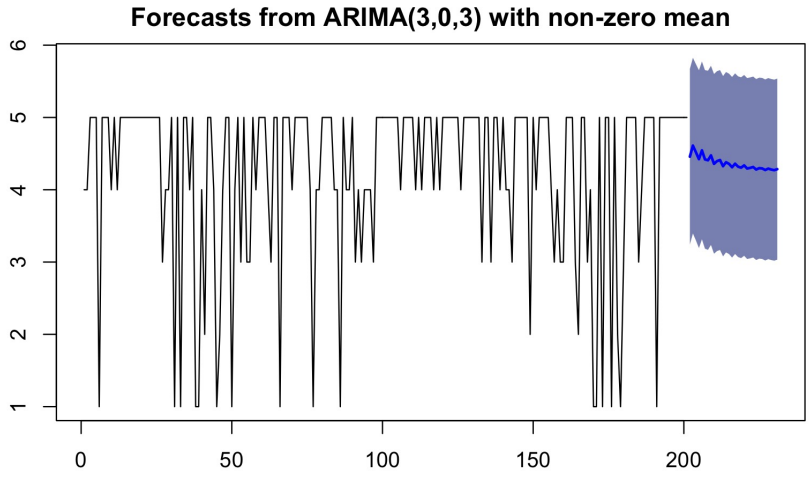


Figure 11: Time Series and Prediction

In addition, based on the plot of monthly average of star ratings, the three data sets show similar trends that helpful reviews have an overall higher average star rating than unhelpful ones, and the helpful reviews' ratings have less fluctuations over time. As we use hair dryer's plot as representative, it shows that the unhelpful votes have their corresponding star ratings vary over a time period. Thus, it's hard for the Sunshine Company to get useful information from the reviews whose star ratings are highly fluctuated, instead, they should focus on reviews with relatively stable star ratings in the long term.

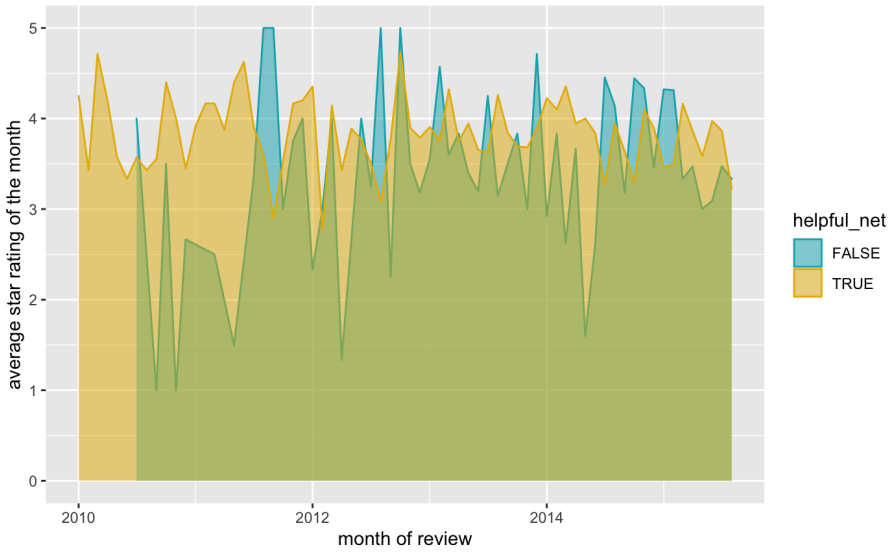
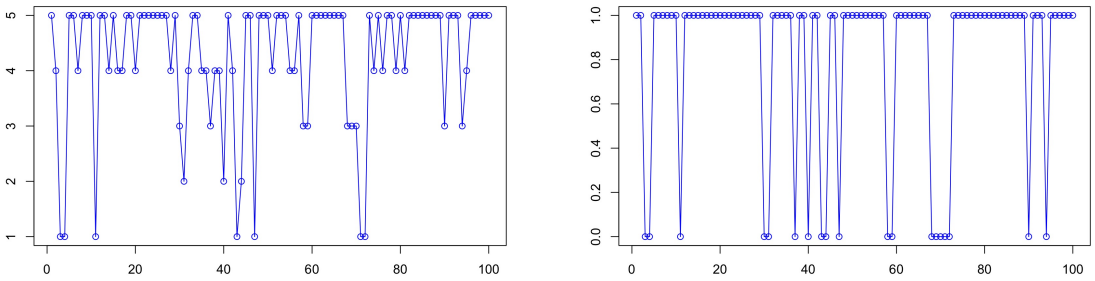


Figure 12: Star Ratings and Helpfulness

To see the interaction between ensuing reviews, we pick one of the bestsellers in all hair dryers as the representative since they share similar time-based pattern

in star ratings. It shows that a lower-star would follow after several consecutive higher-star ratings, while there rarely exist the cases with sequential lower-star ratings. In order to make the pattern clearer, we classify the star ratings less than or equal to 3 as 0, and classify the star ratings greater than 3 as 1, turning the $star\_rating$ into a categorical variable. From the second graph, the pattern is more obvious that a series of high-star ratings will be followed by a low-star rating.



(a) Original Time Series (b) Simplified Time Series

Figure 13: Time-based patterns in Rating

Since the basic assumption of ARIMA model is that the past information should be weighted less than the latest events, it suggests that the Sunshine Company should pay more attention to the up-to-date reviews and ratings that would contain more instructive information for their business decision-making. Also, the pattern that a series of higher-star ratings is usually followed by a lower-star rating implies that the Sunshine Company should arrange their products review properly by mixing reviews with high and low star ratings so that the customers would not leave negative reviews due to psychological factors.

We also use Holt exponential smoothing extended the simple exponential smoothing so that it can forecast with a trend which is helpful since we want to know whether the reputation of a product is increasing or decreasing. The prediction line goes slightly downward that adheres to the forecast based on our ARIMA(3,0,3) model.

$$\begin{aligned} \hat{y}_{t+h|t} &= \ell_t + hb_t \\ \ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \end{aligned} \tag{8}$$

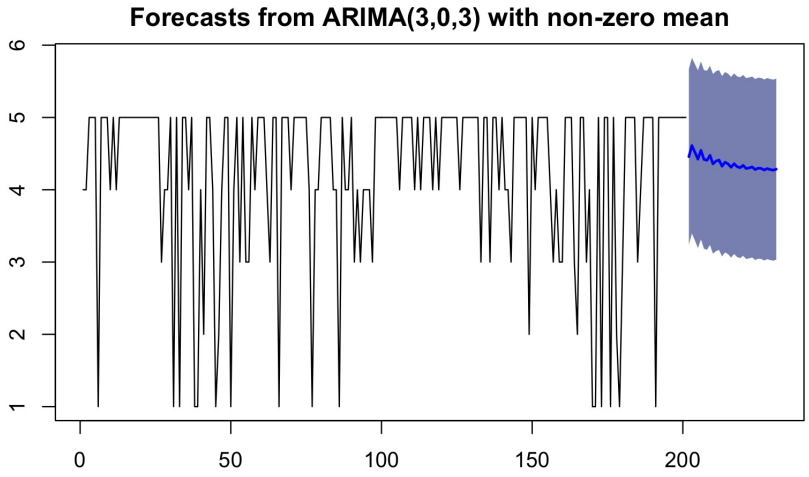


Figure 14: Holt exponential smoothing

6 Analysis of Significant Variables

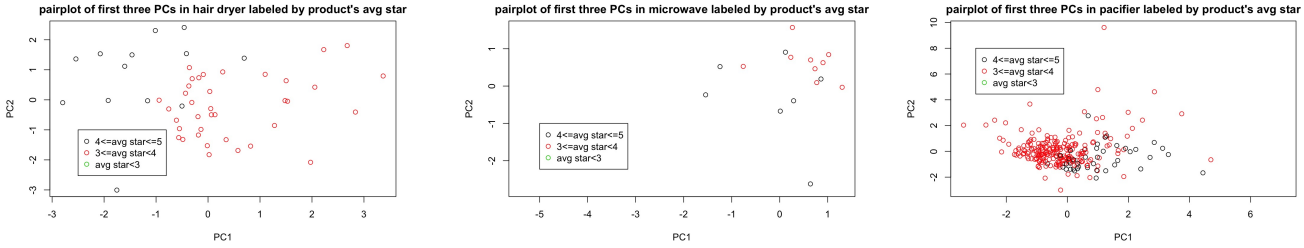
6.1 Principal Component Analysis

To combine different text-based and rating-based measures, we used unsupervised learning to distinguish different products.The first approach is Principle Component Analysis (PCA), where we aim to find a subspace V consisting of orthonormal vectors that minimize:

$$\sum_{i=1}^{n} \|\mathbf{x}_i - proj_V(\mathbf{x}_i)\|^2 \tag{9}$$

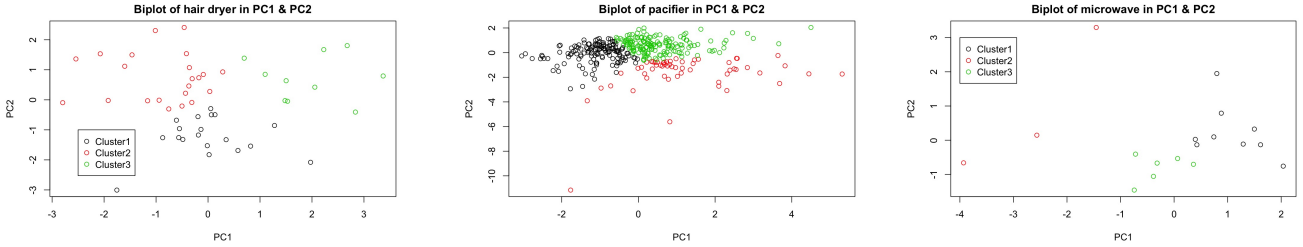
where $\mathbf{x}_i$ is the ith sample.

The vectors in the subspace are linear combinations of the original vectors. According to Courant Fischer Theorem, the ith principal component (PC) is the ith largest eigenvector of the column-centered covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$. To identify whether a product is successful or not, we group products by their parent ID, and average their sentiment scores, word count, specific keyword score, and helpful votes. The plots are labeled by average star.



(a) PCs of Hair Dryer (b) PCs of Microwave (c) PCs of Pacifier

Figure 15: Principal Component Analysis



(a) Clustering of Hair Dryer (b) Clustering of Microwave (c) Clustering of Pacifier

Figure 16: Clustering Plots

For hair dryers and microwaves, we can see that there are no popular products that receive average star ratings lower than three. The first two PCs identified the two subsets well. Thus, we used the following linear combinations of the measurements. However, for pacifiers, PCA does not work well in separating different groups.

Table 3 Linear combinations of measurements

	KW score	head sen	body sen	wordc	help votes
hair dryer	0.54	0.62	0.51	-0.12	0.22
microwave	-0.62	0.23	-0.29	-0.33	-0.61
pacifier	0.25	-0.34	-0.36	0.64	0.53

6.2 Ordinal Logistic Regression

First, we will introduce some notation and review the concepts involved in ordinal logistic regression. Let be Y an ordinal outcome with J categories. Then $P(Y \leq j)$is the cumulative probability of Y less than or equal to a specific category $j = 1, ..., J - 1$. The odds of being less than or equal a particular category can be defined as

$$\frac{P(Y \leq j)}{P(Y > j)}$$

for $j = 1, ..., J-1$ since $P(Y > J) = 0$ and dividing by zero is undefined. The log odds is known as the logit, so that

$$log\frac{P(Y \leq j)}{P(Y > j)} = logit(P(Y \leq j))$$

In ordinal logistic regression, we are trying to find a best model for

$$logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - ... - \eta_p x_p$$

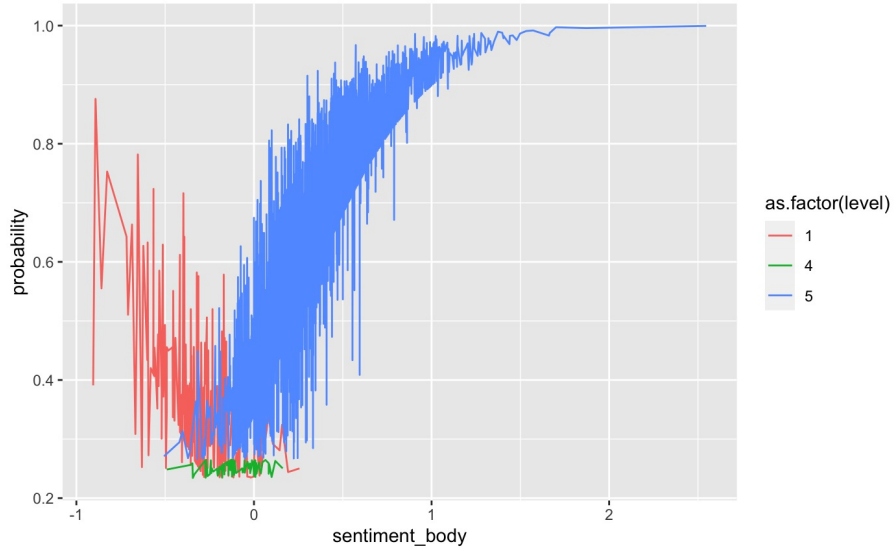


Figure 17: Ordinal Logistic Regression

For problem 2.e, we use the ordinal logistic regression as our model and $sentiment\_body$ and $sentiment\_headline$ as our predictors to predict the probability of every star. We used half of the dataset for training and the remaining part for testing. For the hair dryer dataset, we get the following plot for the ordinal logistic regression. We can see there are no cases being predicted as 3 or 2 star, and the majority of case are predicted as 5 star, resulting in an error rate of 0.401. The graph is noisy, but we can still see that sentiment is a valid predictor for star rating. For microwave and pacifier, their results are similar to the first one. We can say a more positive, or enthusiastic review will result in a higher rating level.

7 Letter to Marketing Director of Sunshine Company

Dear Marketing Director of Sunshine Company:

Our team used text mining on the customer reviews' body to extract desirable features for each product, constructed SVM classifier and identified that

sentiment scores, word count, and keywords that are good at predicting a review's performance, applied time-series ARIMA model to predict the potential success or failure of a certain product, and determined the relationship between ratings and reviews. The following is our suggestions for you to focus on when designing and promoting your products:

- Desirable Feature

For hair dryer, the most desirable features include retractable cord, dual voltage (perfect for travel), light weight, not easily overheated, strong and consistent airflow, suitable for certain type of hair (for example, thick or curly hair). For pacifier, the most desirable features include durability, texture, and the look of the pacifier, to be more specific, customers really love pacifiers with bright colors and cute animals. For microwave, the most desirable include size, heating function, material (for example, stainless), and most importantly, capacity (how many types of modes, including defrost, popcorn, grill, bake, etc.) We suggest that you focus on these features when developing your products, and list them on your product title or description to attract more potential customers. Also, make sure that your IT team make certain changes of your website using these keywords to achieve search engine optimization.

- Helpfulness of a Review

Customers tend to find longer and more emotional reviews helpful. Therefore, we suggest you to look at these features right after the reviews are posted instead of waiting for other customers to vote for its helpfulness. Also, you can identify the helpful ratings and reviews using our classifier to help you save time and cost. In this fashion, you will be able to make timely adjustments of your products and make corresponding change of your sales strategies.

- Deal with Received Reviews

Based on the time-series model assumption, past information should be weighted less than the latest events. We suggest that you should focus more on the most up-to-date reviews for product design adjustment. You will also find that a series of consecutive high-star ratings tend to be followed by some low-ratings. It makes sense in that customers tend to have higher expectations of the product after seeing a series of high ratings. Therefore, our team suggest that you to arrange product reviews by mixing high-star and low-star ratings so that the customers would not leave more negative reviews due to psychological factors.

- Memo

One other notable point is that different from hair dryers and microwaves which are durable goods, baby's pacifiers need to be replaced frequently. We notice that it is common for a customer to buy several different pacifiers at one time or sequentially within a short period so that they can try each of them and compare to determine which one they want to further purchase. In order to maintain loyal customers, Sunshine Company may provide free samples to potential customers and get their feedback to secure the stable need.

References

[1] https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

[2] http://course1.winona.edu/bdeppa/FIN%20335/Handouts/Exponential_Smoothing%20(part%202).html

[3] http://www.sthda.com/english/articles/32-r-graphics-essentials/128-plot-time-series-data-using-ggplot/

Appendices

Here are some programs we used in our model

R code for extracting keyword score:

```
KWScore <- function(text_col = hair_dryer_ratio$review_body, keywords =
c("cord", "cold", "quickly","settings","price")) {
  text_length = length(text_col)
  scores <- rep(0, text_length)
  for (i in 1:text_length){
    word_counter <- rep(0, length(keywords))
    # count words
    for (j in 1:length(keywords) ){
      # num of a keyword occurence in the text
      word_counter[j] = sum(str_count(text_col[i], keywords[j]))
    }
    scores[i] = sum(word_counter)
  }
  return (scores)
}
```

R code for generating word cloud and word frequency:

```
CloudFreq<- function(text_col, removedwords = c("hair", "dryer","can","however","drye
                     ggtitle = "frequency of review_body of hair dryer"){
  # pass in the text column that need to generate a word cloud and frequency plot
  library("wordcloud")
```

```
  corpus <- Corpus(VectorSource(text_col))
  corpus <- tm_map(corpus,removeWords,stopwords("english"))
  corpus=tm_map(corpus,removeWords,removedwords)


  # plot word cloud
  dtm <- TermDocumentMatrix(corpus)
  m <- as.matrix(dtm)
  v <- sort(rowSums(m),decreasing=TRUE)
  d <- data.frame(word = names(v),freq=v)
  wordcloud(words = d$word, freq = d$freq, min.freq = 1,
            max.words=200, random.order=FALSE, rot.per=0.35,
            colors=brewer.pal(8, "Dark2"))

  getPalette = colorRampPalette(brewer.pal(3, "Accent"))

  # plot freq barplot
  p<-ggplot(data=d[c(1:30),], aes(x=reorder(word, freq), y=freq)) +
    geom_bar(stat="identity")+coord_flip() + labs(title = ggtitle, y="frequency", x =
    geom_col(fill = getPalette(30))

  p
}
```

R code for getting sentiment:

```
helper <- function(corpus){
  # take in a col of corpus, in MCM, is the review body or headline
  tokens <- data_frame(text = corpus) %>% unnest_tokens(word, text)
  sentiment <- tokens %>%
    inner_join(get_sentiments("bing")) %>% # pull out only sentiment words
    count(sentiment) %>% # count the # of positive & negative words
    spread(sentiment, n, fill = 0) %>% # made data wide rather than narrow
    mutate(positive_ratio = (positive)/(positive+negative) )# # of positive words - #

  return(sentiment)
}

GetSentiments <- function(corpus) {
# accept df like hair_dryer_ratio
  df <- NaN
  for (i in c(1:5)){
    star = corpus$review_body[corpus$star_rating == i]
    r <- helper(star)
    df <- rbind(df, r)
  }
  df = df[-1,]
  return (df)
}
```