Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2022849

In the age of e-commerce, textual data has become an ever rich of source of information for firms because there are now cheaper and easier ways to get sentiments of consumers on a product. Thus, over the years natural language processing has become more important as firms use both reviews and other forms of information such as ratings to improve their products and their brand image. More importantly with so many important variables we need to determine the important data measures that firms follow that impact their priorities which in this problem are sales.

In our paper, we want to important features per product for consumers. First, we use a term frequency model that chooses the important on the basis of words that appear the most in all reviews disregarding stop words such as “the”, “I”. We also use a textual analysis method known as the Latent Dirichlet Allocation which is a generative probabilistic model that determines the distribution of words over a collection of reviews. The Latent Dirichlet Allocation model assumes that every review is a distribution latent topic and each topic is represented as a distribution of words, with certain words having a higher probability of appearing under certain topics.

Another important aspect in our problem is to determine the important measures and patterns among data variables. For this problem we first identify that Sunshine wants to maximize their sales, and that do so would be a success for the them. To establish the measures that determine the success of their products and identifying the important measure we multiple linear regression where we identify the star ratings and quality or impact of reviews and average sentiment as the independent variables. We measure quality of reviews simply as ratio between the sum of helpful votes and total votes. Further, we look at the impact of who is making the review since one can be either a vine member or not a vine member.

Key words: Sentiment analysis, textual analysis, topic modelling, term frequency, Latent Dirichlet Allocation, logistic regression, multiple linear regression

MCM Problem C: A Wealth of Data

Team 2022849

March 10, 2020

Contents

1 Introduction

1.1 Overview

Data analysis is very important for firms in that it helps them gain extra market data for product development and marketing that could boost their brand, and improve their profit level. A significant amount of data that firms have to interpret comes is in text form such as product reviews, blogs or tweets which computers cannot draw information from at first hand. On the other hand, although we humans can understand the meaning behind a text, this process of drawing meaning from text becomes much more difficult for us when we have a massive amount of data to deal with. Despite the structural issues that plague textual data, we cannot deny its significance. Textual data provide crucial information such as how consumers perceive our product and what features they deem to be important on a product. Thus, both product reviews and star ratings provide a great deal of information to firms that they can use to position their brand and product for success.

1.2 Problem statement

Our paper deals with a number of problems that are all aimed at making Sunshine make more informed decisions that will help launching successful microwave, hairdryer and pacifier products into the market. Further, we have to find important data measures and relationships that exists in among helpful votes, ratings and reviews to help measure the popularity and success of Sunshine in the market.

1.3 Assumptions

- Total votes represent potential customers and helpful votes highlight whether the consumer will purchase our product which is determined by the rating and sentiment of the review
- The quality of review is determined by the ratio of people who vote it to be helpful to all the votes.
- A Vine review, written by a trusted member of the Amazon review community for their history of insightful reviews, has much more influence to any other review.
- The longer the review the more information it contains about the product, thus, more likely to influence the next purchase.
- Popular products have higher visibility to the potential customers, by having higher ranking in Amazon recommender systems. The current overall rating influences when the particular product appears in a search, and products with better ranking appear first. This ranking is determined by both the quantity and quality of star ratings and reviews for the product.
- A successful online sales strategy constitutes maximized effectiveness of the descriptions of their product features to attract the most potential customers possible, as well as good quality of the products so that customers would be more likely to leave high star ratings and positive reviews. Hence, to inform the company on its online sales strategy, we need to identify the common features of competing products that have both higher frequency of star ratings and reviews, and better, more positive star ratings and reviews. This would then translates to features that the company may consider enhancing to improve the quality of their products.

1.4 Key definitions and approach

- We use regression methods that treat the sales as the explained variable to determine the measures and patterns that are important for the firm to track.
- Sentiment analysis of reviews is the measurement of the review polarity. The review polarity then becomes one of our indicators for prediction of success or failure of a product along with star ratings and helpfulness rating.
- Topic modelling is used to identify a combination of several product features within all reviews. In this case, our "product features" are "topics" to be modelled. This will be done using Latent Dirichlet Allocation model, which models the distribution of topics over reviews and distribution of words over topics.
- Product success is determined by a function of expected purchases and the satisfaction of customers based on customer feedback indicators. This will be done using logistic regression model that classifies whether a product is more or less likely to be purchased based on a combination of customer feedback factors.
- Reputation of a product is determined by its average star rating, how quantity and quality of review is influenced by star ratings and vice versa. This will be done using a system of discrete homogenous linear equations.

2 Data processing

2.1 Preprocessing

For further analysis in the logistic regression model, we filter our these following data:
- any reviews that has less than 50 characters.
- ratio of helpful to total votes must be at least 98%.
- number of helpful votes is at least 1.

2.2 Extract sentiments from text

An important measure of customer satisfaction lies in the polarity of reviews. A review can be assigned a numeric value that rates its polarity. A generic way to do this is through a simple dictionary lookup method to evaluate the frequency of positive ("happy", "satisfied", "awesome") and negative words("sad", "disappointed", "broken") appearing in a review. However, we also need to take into account of valence shifters, such as negators (do **not** like), amplifiers (**really** like), deamplifiers (**barely** works) and adversative conjunctions (I like it **but** it's expensive) [1].
According to Rinker (2019), a word in each review is denoted by $w_{i,j,k}$(i^{th} word in j^{th} sentence in k^{th} review). We compare each word to a dictionary of polarized words, with positive word and negative word valued at 1 and -1 respectively, then form a polarized cluster of words for each sentence ($c_{i,j}$). This cluster includes n_b number of words before the polarized word and n_a number of words after. Polarity scores for each type of valence shifter are calculated as follows:

- Negator ($w^n_{i,j,k}$): $(-1)^k + 2$, $k = \sum w^n_{i,j,k}$, assuming odd number of negators is negative and even number of negators is positive.
- Amplifier ($w^a_{i,j,k}$): $\sum w^n_{i,j,k}(z_1 \cdot w^a_{i,j,k})$, z_1 being the weight of amplifier on the word, which is defaulted to be 0.8.
- Deamplifier ($w^d_{i,j,k}$): $\sum z_1(-w^n_{i,j,k} \cdot w^a_{i,j,k} + w^d_{i,j,k})$, z.
- Adversary conjunction ($w^a_{i,j,k}c$): $w_b = 1 + z_2 \cdot \sum |w^a_{i,j,k}c|, w^p_{i,j,k}, ..., w^a_{i,j,k}c| - 1$. This weighs the sentiment of the part after adversary conjunction more than the one before.

3 Patterns, relationships between star ratings, helpfulness rating and review

3.1 Relationship among ratings, word count, helpful votes and review sentiments

The main takeaway from the investigation we conduct here is that for all products ratings have a negative correlation with word count, albeit this relationship is very small $-.16 < r < 0$. However, the regression highlights that this is still an important relationship since our variables are significant in all products even at the 1 significance level. The result is not surprising given that research shows how bad news is more likely to spread compared to good news[4]. But this has broad impact on a product branding as this highlights how negative ratings can lead to more reviews. Further, we also find the relationship between word count and helpful votes. Here we want to find whether our assumption of longer reviews containing more information also means that they are helpful. Our correlation albeit being small $0.37 > r > 0.22$, for all products shows a positive relationship that exist between helpful votes and word count. The result here has important implications. The result is showing us that high ratings are less likely to encourage future purchases because people are less likely to write helpful reviews, where as dissatisfied consumers more helpful reviews that influence the future consumer not to purchase the product.

Microwaves			
	Star ratings	Word Count	Helpful votes
Star ratings	1		
Word Count	-0.16994	1	
Helpful Votes	0.011731	0.340327	1

Hairdryers			
	Star ratings	Word Count	Helpful votes
Star ratings	1		
Word Count	-0.11047	1	
Helpful Votes	-0.04426	0.277535	1

Pacifiers			
	Star ratings	Word Count	Helpful votes
Star ratings	1		
Word Count	-0.10693	1	
Helpful Votes	-0.07008	0.218778	1

3.2 Multiple linear regression to determine useful measures

The objective for Sunshine firm when launching their products in the market is to maximize their sales. Therefore, we have to decide to observe the data measures that help the firm reach this objective. We are aware that consumers use ratings and reviews when making their purchasing decisions. However, we are not sure which of the two influences the consumer the most, when making a choice. Obviously, it is important for a product to have a good rating because algorithms used by firms such as Amazon mean that such a product will likely appear on the top in a search. Further, consumers then use reviews to distinguish which of the highly recommended goods they want to buy. However, we have to determine when a review is most likely to have higher influence on the consumer. We assume that reviews that have received helpful votes are considered to be trustworthy, therefore, have a bigger influence on purchasing decisions compared to a review that did not receive any helpful votes. Thus products which have higher reviews that are deemed to be helpful are more likely to be witness an increase in purchases or not depending on overall sentiments of the review. To establish any relationship or measure, we decided to group products by their brand so as to gather any meaningful measures or relationship.

Therefore, we use the multiple linear regression to determine which measure has more influence on product sales. We naturally choose average sentiment of reviews and review rating as two of our independent variables. We also add quality of review as another independent based on the assumption we established in the prior paragraph. Thus we have

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

in which y, x_1, x_2, x_3 are sales, average star ratings, average quality of reviews, average sentiment respectively. However we deem that it might be more important for a firms to know semi-elasticity of their product. Therefore,

$$\ln y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

3.3 Results and mathematical explanations

Overally, all our regression highlighted that the functional model was statistically significant albeit explaining a small variation in our explained variable. Another noteworthy observation from our analysis is that quality of reviews is an important measure as its weight coefficient is always the highest. For both models and all the products we realize the quality of reviews has economic significance in the model and at the same time, it is statistically significant at the 5 percent level.

Hair dryer (Sales) regression			
Variables	Coefficients	T-stat	P value
Intercept	-5.2926	-0.5466	0.5849
Average rating	0.0736	0.0304	0.9748
Quality of review	50.1797	8.1579	$3.15 \cdot 10^{-15}$
Average sentiment	33.1113	1.94122	0.05283

Pacifier (Sales) regression			
Variables	Coefficients	T-stat	P value
Intercept	-0.3295	-0.3619	0.7175
Average rating	0.4415	1.9090	0.0563
Quality of review	5.1721	8.5196	$2.05 \cdot 10^{-17}$
Average sentiment	0.5828	0.5313	0.5952

Microwave (Sales) regression			
Variables	Coefficients	T-stat	P value
Intercept	-16.9020	-.8046	.4248
Average rating	4.4052	0.5753	0.5676
Quality of review	38.3207	2.4708	0.0169
Average sentiment	49.8272	0.7225	0.4733

3.4 The impact of vine members

We measure the impact of review determined by whether the reviewer is in the vine group or not. We calculate the impact of reviews as being

$$= \frac{\sum \text{helpfulvotes}}{\sum \text{totalvotes}}$$

This is based on our hypothesis that people are likely to give helpful votes only when they deem the review to have helped them in making a decision.

Microwaves	
Category	Impact of review
Not vine member	0.8288
Vine member	0.9424

Hairdryers	
Category	Impact of reviews
Not vine member	0.852035
Vine member	0.768437

Pacifiers	
Category	Impact of reviews
Not vine member	0.73064
Vine member	0.77259

Psychology studies highlight that we are more persuaded by people with authority and the vine community is regarded as having this kind of influence on consumers. Thus our expectation would be that vine member quality of reviews or rather influence is greater than that of non-vine members. Interestingly, we do not see that much of a difference in the impact of reviews between non-vine and vine members. In fact for the case of hairdryers non-vine members have more influence that vine members.

3.5 Implications

To summarize, we notice from the tables that we notice that the most important variable is the quality of reviews. However, we have to mention here that our results should be reflected upon carefully. Our results though showing a positive relationship between sales and quality of reviews, it has to be mentioned that if the with quality reviews where average reviews sentiment for your products are negative, then with a more complex model this should imply a decrease in sales. Further, the firm does not need to really pay special attention to the vine members as their influence is almost comparable to non-vine members

4 Prediction Models

4.1 Identify product features using term frequency and topic modelling

4.1.1 Introduction

To identify important product features, a traditional and fundamental method would be looking at the frequency of the most common n-grams in the reviews. One of the models that makes use of this method is the tf-idf model (term frequency count-inverse document frequency). Such model, however, does not take into account the structural complexity within and between reviews, such as the probability of two or more topics to appear at the same time across all reviews. For the company to make a decision on improving quality of their products and writing an attractive product description, they may want to look at a set of features that correlate with each other instead of looking at individual, discrete features. A topic generative model such as Latent Dirichlet Allocation (LDA), which aims to find the best set of latent features that can fit the best with the observed data, would provide a more comprehensive approach to infer important product features. [2]

4.1.2 Term frequency model

For the term frequency model, we separated all words into n-grams (tokens) across all reviews in each data set, before removing all numbers and stop words which provide little insight (e.g. "the", "a", "she"). Further common generic descriptors or emotions were filtered, for example, "product", "love", "nice", "perfect" are filtered out for all data sets. Setting n to be 1, we then counted the frequency of all single words and picked out the top 10 words.

Pacifier		Microwave		Hair dryer	
word	frequency	word	frequency	word	frequency
son	2506	unit	444	blow	1968
easy	2498	door	392	time	1954
daughter	2356	time	391	cord	1697
cute	1990	oven	356	heat	1566
time	1964	service	272	hot	1433
seat	1320	kitchen	250	dryers	1352
hold	1150	cooking	244	price	1302
bag	1135	space	236	air	1207
clean	1026	easy	235	setting	1031
size	1009	samsung	221	light	979

Table 1. Top 10 words based on all reviews on competing pacifier, microwave and hair dryer products over the years.

Now set $n = 2$, we have the following results:

Pacifier	
word	frequency
car seat	512
diaper bag	391
stuffed animal	295
super cute	209
shower gift	191
washing machine	118
nipple confusion	94
life saver	91
bpa free	88
hard time	87

Microwave	
word	frequency
stainless steel	75
customer service	72
counter space	59
convection oven	41
control panel	36
counter top	33
dinner plate	29
service call	29
start button	22
circuit board/cook time/exhaust fan	20

Hair dryer	
word	frequency
retractable cord	441
heat setting(s)	409
light weight	287
drying time	193
night light	181
cool shot	177
air flow	162
wall mount	142
shot button	138
flat iron	134

Table 2. Top 10 bigrams based on all reviews on competing pacifier, microwave and hair dryer products over the years.

4.1.3 Latent Dirichlet Allocation model

According to Biel et. al. (2003), LDA is a generative probabilistic model that works on the assumption that every review is a mixture of latent topics, and each topic is represented as a distribution over words.
The following notation scheme and design are adapted from Biel et. al. (2003) and Steyvers and Griffiths (2007) [2] [3].

(a) Model design
First, we define the following:

- word: an element in the k x V matrix of k latent features and vocabulary size V of all reviews, denoted by $w_{i,j}$, the j^{th} word in the i^{th} feature.
- review: a vector of N_d words in the d^{th} review, denoted by $\mathbf{w^{(d)}} = (w_1, w_2, ..., w_{N_d})$, where $N_d \leq V$.
- corpus: a collection of reviews, denoted by $D = \{\mathbf{w^{(1)}}, \mathbf{w^{(2)}}, ..., \mathbf{w^{(n)}}\}$.
- topic: the i^{th} feature of j^{th} word, denoted by $z_{i,j}$.

- k-dimensional topic distribution: distribution of features in review d based on observed data, denoted by a multinomial variable $\theta^{(d)}$.
- α: parameter matrix for prior estimated probability distribution of words in topics.
- β: parameter matrix for prior estimated probability distribution of topics in reviews.

The algorithm has the following assumptions:
1. For i^{th} feature in k latent topics, choose $\phi^{(i)} \sim \text{Dir}(\beta)$
2. For review d in corpus D: Choose $N_d \sim \text{Poisson}(\xi)$, choose $\theta^{(d)} \sim \text{Dir}(\alpha)$
3. For j^{th} word in the i^{th} feature in review d of N_d words, choose $z_{i,j} \sim \text{Multinomial}(\theta^{(d)})$, choose $w_{i,j}$ Multinomial($\phi^{(i)}$).
4. We use Bayes' Theorem to describe any conditional probability of an event based on the prior knowledge of the event.

To find the probability distribution of a feature in reviews and a word in features, we sample the feature i from the topic distribution, then sample word j from the topic distribution [3]. The probability distribution of words in a review, given topic distribution in a review is:

$$P(w_{i,j}) = \sum_{i=1}^{k} P(w_{i,j}|z_{i,j})P(z_{i,j}) \tag{1}$$

in which $P(z_{i,j})$ is the probability of sampling word j in feature i, and $P(w_{i,j})$ is the probability of j_{th} word under i^{th} feature.
Let $\theta^d \sim \text{Dir}(\alpha_i)$, since Dirichlet distribution is conjugate for the multinomial distribution. The following is the probability of features in reviews given prior probability of features in reviews:

$$P(\theta|\alpha) = \text{Dir}(\alpha_1, \alpha_2, ..., \alpha_k) = \frac{\Gamma(\sum_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \Pi_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{2}$$

in which $\Gamma(\alpha_i)$ is prior probability of i^{th} feature to appear in a review, which is assumed to be a Gamma function; $\Gamma(\sum_i \alpha_i)(\Pi_{i=1}^{k} \theta_j^{\alpha_i - 1})$ is the joint probability distribution of posterior and prior probability of all features in reviews.
The probability of words in reviews given prior estimated distribution is given by $P(w_n|z_n|\beta)$

$$P(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = P(\theta|\alpha)\Pi_{n=1}^{N} P(z_n|\theta)P(w_n|z_n, \beta) \tag{3}$$

We get the marginal word distribution over a review, by the integration over θ and summing over z:

$$P(\mathbf{w}|\alpha, \beta) = \int P(\theta|\alpha)(\Pi_{n=1}^{N} \sum_{z} {}_n P(z_n|\theta)P(w_n|z_n, \beta))d\theta \tag{4}$$

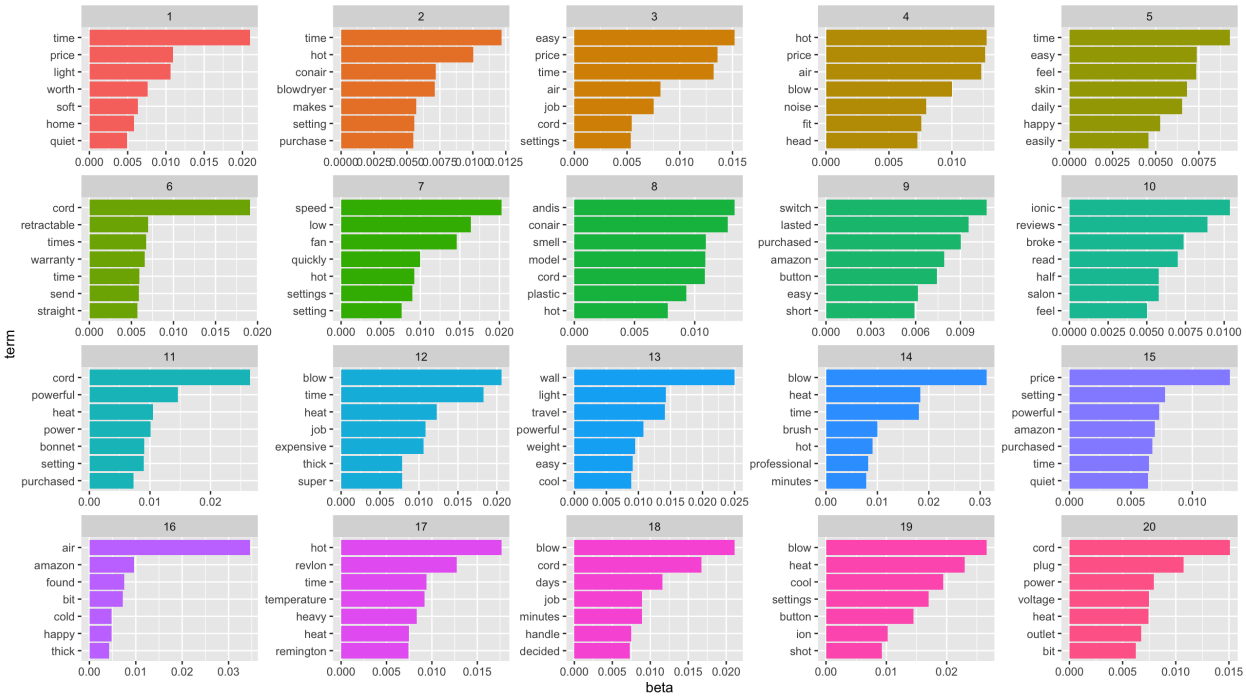
Taking the product of all the marginal distribution, we have the probability distribution of a corpus:

$$P(D|\alpha, \beta) = \Pi_{d=1}^{M} \int P(\theta^d|\alpha)(\Pi_{n=1}^{N} \sum_{z} {}_{dn} P(z_{dn}|\theta^d)P(w_{dn}|z_{dn}, \beta))d\theta^d \tag{5}$$

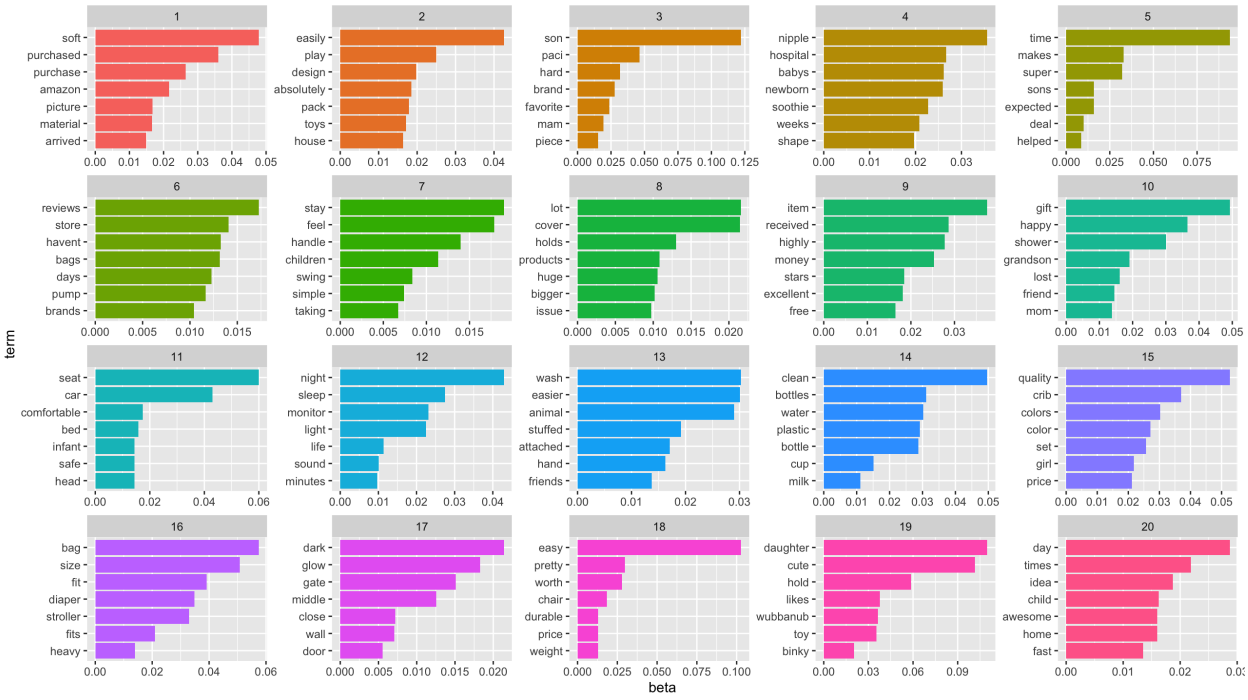
(b) Results

For each dataset, we make a corpus of all reviews of all products over the years, with all reviews as strings to undergo text processing, including removal of punctuations, contractions, stop words, numbers, white spaces and lowercase transformation. Then, we ran LDA over a document-term-matrix of the corpus (excluding empty reviews) with k being any number from 2 to 20 and computed a perplexity score for each k value. The lower a perplexity score is, the better the probability model is at predicting the sample. From Figure 3 (Appendix), as k increases, perplexity increases for all products.

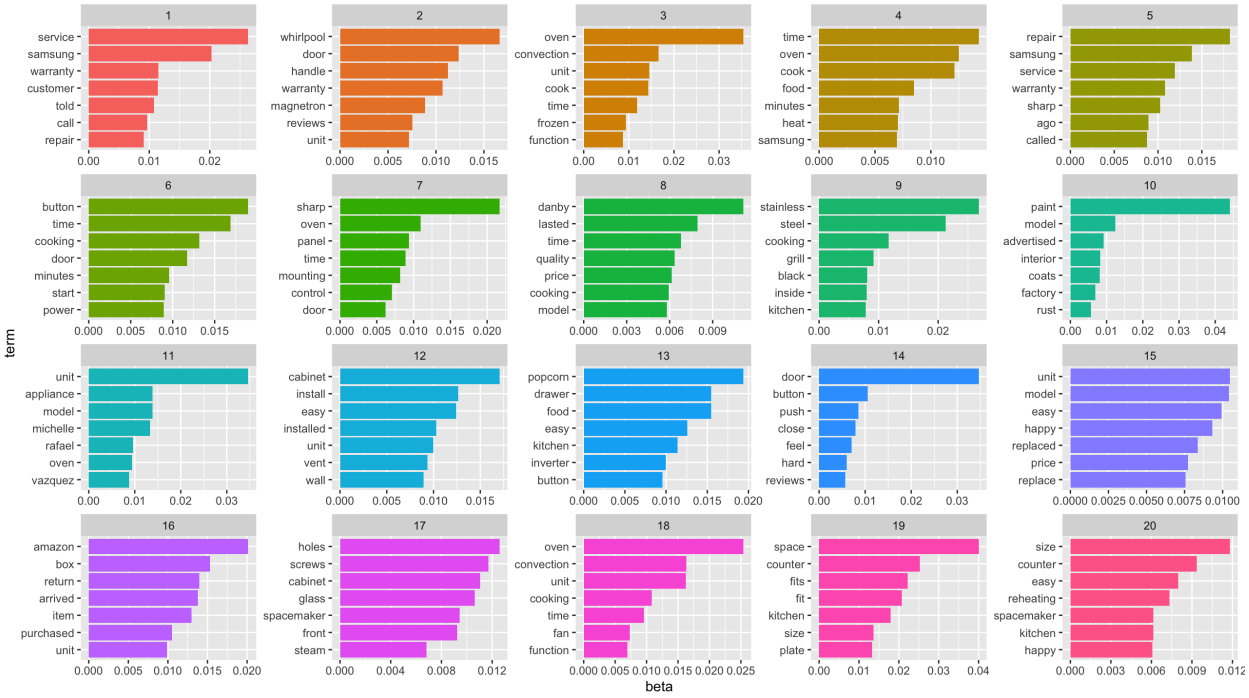
Figure 1: Beta distribution of words over topics, k = 20



(a) Hair dryer



(b) Pacifier


1
service
samsung
warranty
customer
told
call
repair
2
whirlpool
door
handle
warranty
magnetron
reviews
unit
3
oven
convection
unit
cook
time
frozen
function
4
time
oven
cook
food
minutes
heat
samsung
5
repair
samsung
service
warranty
sharp
ago
called
6
button
time
cooking
door
minutes
start
power
7
sharp
oven
panel
time
mounting
control
door
8
danby
lasted
time
quality
price
cooking
model
9
stainless
steel
cooking
grill
black
inside
kitchen
10
paint
model
advertised
interior
coats
factory
rust
11
unit
appliance
model
michelle
rafael
oven
vazquez
12
cabinet
install
easy
installed
unit
vent
wall
13
popcorn
drawer
food
easy
kitchen
inverter
button
14
door
button
push
close
feel
hard
reviews
15
unit
model
easy
happy
replaced
price
replace
16
amazon
box
return
arrived
item
purchased
unit
17
holes
screws
cabinet
glass
spacemaker
front
steam
18
oven
convection
unit
cooking
time
fan
function
19
space
counter
fits
fit
kitchen
size
plate
20
size
counter
easy
reheating
spacemaker
kitchen
happy
term
beta

c) Microwave

4.2 Product success prediction with binary logistic regression algorithm

There are two aspects of product success: expected sales, which is based on whether verified purchase is made, and customer satisfaction based on the polarity of reviews.
Logistic regression is a supervised learning classification algorithm, which we use to find conditional probability of verified purchase, which ranges from 0 to 1, given one or a combination of feedback indicators. Let this probability be $P(y = 1|x; w) = p(X)$, where $X =$ a verified purchase is made and $X \sim B(1, p)$.

4.2.1 Model design

This model is based on a standard logistic Sigmoid function, defined by $y = \frac{1}{1+e^{-x}}$. Let $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ be our feature vector of n independent variables. Similarly to linear regression model, our response variable, which is verified purchase, can be expressed as a a linear combination of a weight vector $\mathbf{w} = \{w_1, w_2, ..., w_n\}$ and the feature vector, plus a bias term w_0:

$$w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n = w^T \cdot \mathbf{x}$$

Substituting into the logistic function, we have:

$$p(X) = \frac{1}{1 + e^{-\mathbf{w}^T \cdot x}}$$

From the Sigmoid function, $p(X)(1 + e^{-\mathbf{w}^T \cdot \mathbf{x}}) = 1 \implies \frac{1}{e^{-\mathbf{w}^T \cdot x}} = \frac{p(X)}{1-p(X)}$. Thus, the log-odds ratio, which is the log ratio of whether a purchase is likely to occur to whether a purchase is not likely to, is defined by:

$$\ln \frac{p(X)}{1 - p(X)} = \mathbf{w}^T \cdot \mathbf{x}$$

We must estimate the parameter vector such that the likelihood function $L(\mathbf{w}|y) = \Pi_{i=1}^{n} P(y = 1|x; w)_i^{y_i}(1 - P(y = 1|x; w))_i^{1-y_i}$ is maximized. Taking the natural logarithm of the function, we have the log-likelihood function defined as:

$$\ell(\mathbf{w}|y) = \ln(\sum_{i=1}^{n} P(y = 1|x; w)_i^{y_i} + \sum_{i=1}^{n}(1 - P(y = 1|x; w))_i^{1-y_i})$$

The cost function, which is the negative log-likelihood, is to be minimized by a method called gradient descent. We take the partial derivative with respect to each w_n term of the weight factor to eventually get:

$$\frac{\partial \ell(w)}{\partial w} = \sum_{i=1}^{n}(y - P(y = 1|x, w))\mathbf{x}^T$$

The gradient descent is then defined as:

$$w_{i,j} := w_{i,j} \frac{\partial \ell(w)}{\partial w}$$

4.2.2 Implementation and results

We fixed our dependent variable to be verified purchase, which is either 0 or 1, and varied our set of independent variables among star ratings, helpful votes, year, average review sentiment, assuming these variables do not have very strong correlation with each other.
We denote the following:

- y: probability of verified purchase, ranging from 0 to 1

- x_1: log scale of helpful votes
- x_2: star rating
- x_3: year
- x_4: exponential scale of average sentiment of review

We look at the magnitude of z-value and estimated coefficient to infer the significance of each feature in predicting verified purchase. Hence, for pacifier and hair dryer, the most important feature is year, while it is star rating for microwave. The second most important feature is x_1, x_4 and x_2 for pacifier, microwave and hair dryer respectively.

Pacifier				
Variables	Estimated coefficients	Standard Error	z-value	p-value
Intercept	-0.18270	0.07854	-2.326	0.0200
x_1	-0.18270	0.07854	-2.326	0.0200
x_2	0.06205	0.04731	1.312	0.1896
x_3	0.34024	0.02363	14.401	$< 2e^{-16}$
x_4	0.55304	0.23712	2.332	0.0197
Microwave				
Variables	Estimated coefficients	Standard Error	z-value	p-value
Intercept	-4.76166	1.09822	-4.336	1.45^e-5
x_1	-0.22997	0.16528	-1.391	0.1641
x_2	0.79494	0.09383	8.472	$< 2e^{-16}$
x_3	0.13326	0.07394	1.802	0.0715
x_4	1.30509	0.69129	1.888	0.0590
Hair dryer				
Variables	Estimated coefficients	Standard Error	z-value	p-value
Intercept	-2.66928	0.36250	-7.364	$1.79e-13$
x_1	-0.08111	0.06827	-1.188	0.235
x_2	0.20460	0.04594	4.454	$8.43e-6$
x_3	0.33513	0.02369	14.146	¡ 2e-16
x_4	0.02487	0.24038	0.103	0.918

Table A. Logistic regression model on prediction of sales of 3 products

4.3 Predict future customer behavior in response to star rating

We first define the following, for a certain product:

- n: time step
- S_n: star rating at time step n
- R_n: review count at time step n
- Q_n: review quality at time step n, determined by total helpful votes of all reviews
- T_n: review sentiment at time step n

If we hypothesize that future customer behavior, including future review quantity, quality and sentiment, would be immediately influenced by current star rating, we can formalize a multi-variable, deterministic discrete linear model as follows:

$$\begin{aligned} S_{n+1} &= a_1 \cdot S_n + b_1 \cdot R_n + c_1 \cdot Q_n + d_1 \cdot T_n + e_1 \\ R_{n+1} &= a_2 \cdot S_n + b_2 \cdot R_n + c_2 \cdot Q_n + d_2 \cdot T_n + e_2 \\ Q_{n+1} &= a_3 \cdot S_n + b_3 \cdot R_n + c_3 \cdot Q_n + d_3 \cdot T_n + e_3 \\ T_{n+1} &= a_4 \cdot S_n + b_4 \cdot R_n + c_4 \cdot Q_n + d_4 \cdot T_n + e_4 \end{aligned}$$

We treat our system of linear equations as a system that evolves over time, thus we can use the diagonalization to describe such a system. We use this method to identify how our variables change over time.
Assuming the constant terms e_1, e_2, e_3, e_4 are all 0, our system is first-order homogeneous linear equation. Let us treat our system of linear equations as the matrix A, where A=

$$\begin{bmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{bmatrix}$$

then by diagonalization, $A = P \cdot D \cdot P^{-1}$. Therefore we can write $A^k = P \cdot D^k \cdot p^{-1}$, where k represents time. By theorem we know that A is diagonalizable if and only if has eigenvectors $x_1, x_2, ...x_n$such that the matrix $P = [x_1 x_2 ... x_n]$ is invertible and since P is invertible then we have P^{-1}. When this happens $P^{-1} \cdot A \cdot P = diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$, where for each i, λ_i is the eigenvalue of A corresponding to x_i the eigenvectors. Our matrix D will inform us on how our system will evolve. It is more likely that if a variable is associated with a $\lambda_i > 1$ then that variable will grow.

5 Letter to Sunshine company

Dear Sunshine Company,
We write this letter to inform you on important data measures and important design features you need to know to enhance your products. We use linear and logistic regression models to identify important data measures. To predict future sales, you may want to look at review sentiments, number of helpful votes for reviews and may not consider looking at star rating as star rating does not seem to impact the prediction models much. We identify the important features based on frequency of words over all reviews. The most common features across all products that customers focus on is ease of use. Specifically, pacifier's features include durability, material and appearance. Microwave's features include material and occupied space. Hair dryer's important features include heating element and weight.
We hope this helps in informing you to make sound marketing decisions. Thank you for your time.
Yours sincerely,
Team 2022849

6 Appendix

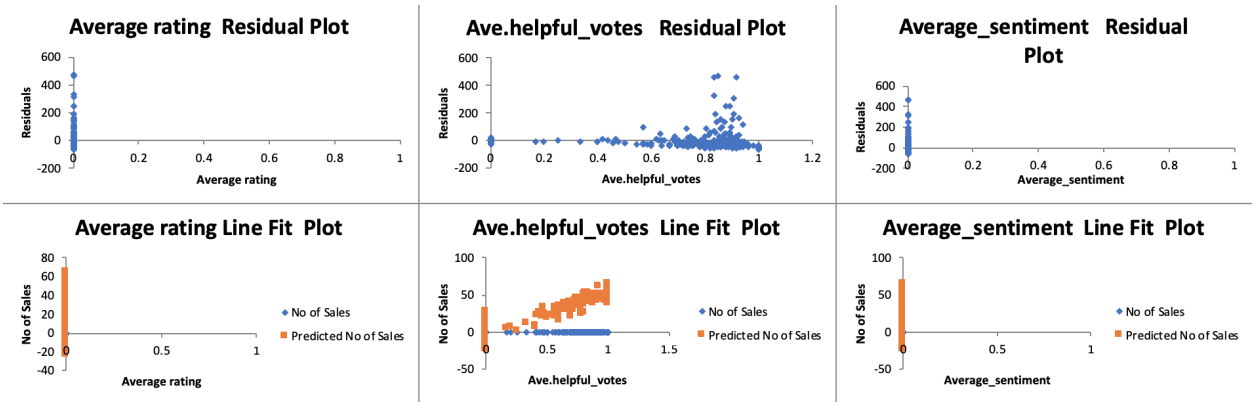
6.1 Softwares and packages

We processed our data and built our prediction models with the help of Excel and the following Python and R packages:

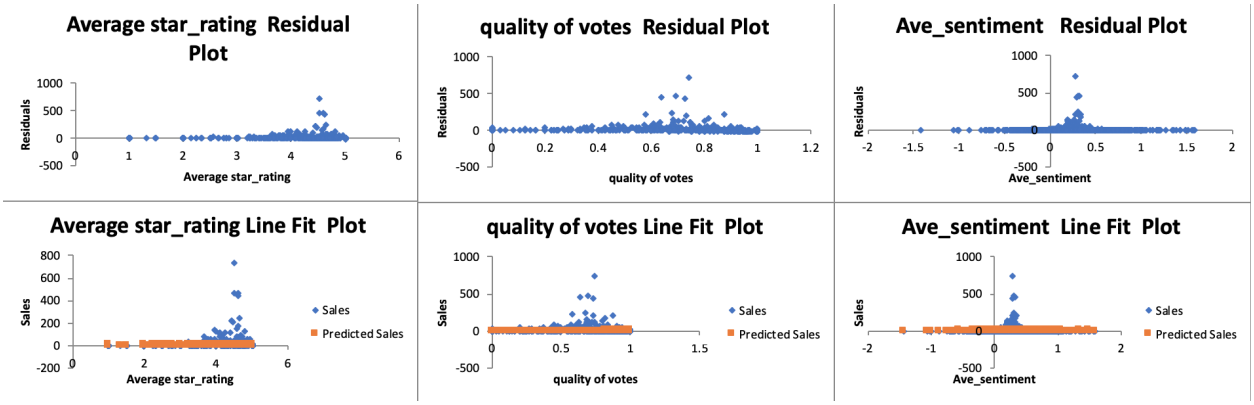
Packages	Functions	Documentation
pandas	data analysis and manipulation	http://github.com/pandas-dev/pandas
numpy	complex mathematical functions	https://numpy.org/
tm	text processing, create corpus over all reviews	https://cran.r-project.org/web/packages/tm/tm.pdf
dplyr	manipulation of data sets	https://cran.r-project.org/web/packages/dplyr/dplyr.pdf
tidytext	text mining for word processing and sentiment analysis	https://cran.r-project.org/web/packages/tidytext/tidytext.pdf
qdap	text processing	https://cran.r-project.org/web/packages/qdap/qdap.pdf
sentimentr	sentiment analysis of reviews	http://github.com/trinker/ sentimentr
ggplot2	visualization of prediction models	https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
tidyverse	a set of packages that work in harmony	https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf
topicmodels	create document-terms-matrix, build LDA models	https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf
broom	summarizing key information in the form of tidy tibbles for faster reporting	https://cran.r-project.org/web/packages/broom/broom.pdf

6.2 Figures

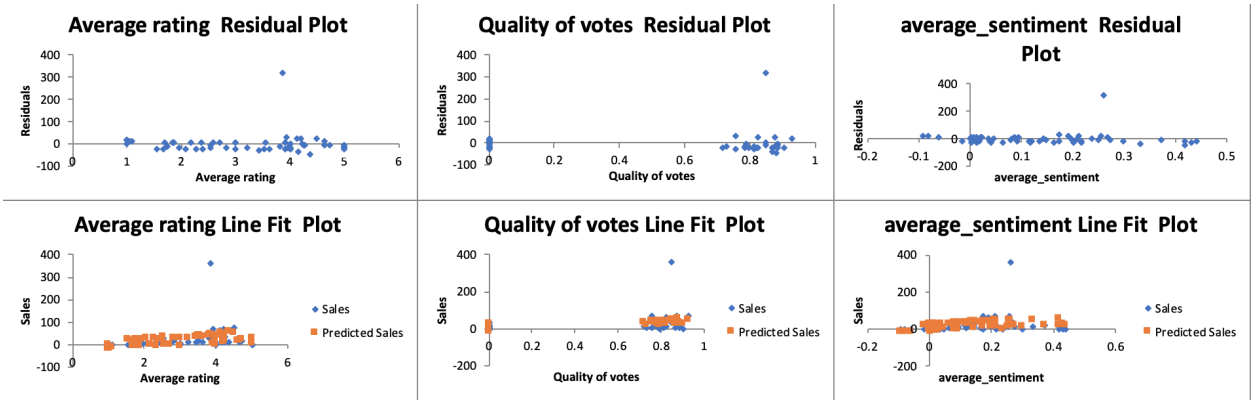
Figure 2: Lines and residuals plots for 3 products:



(a) Hair dryer

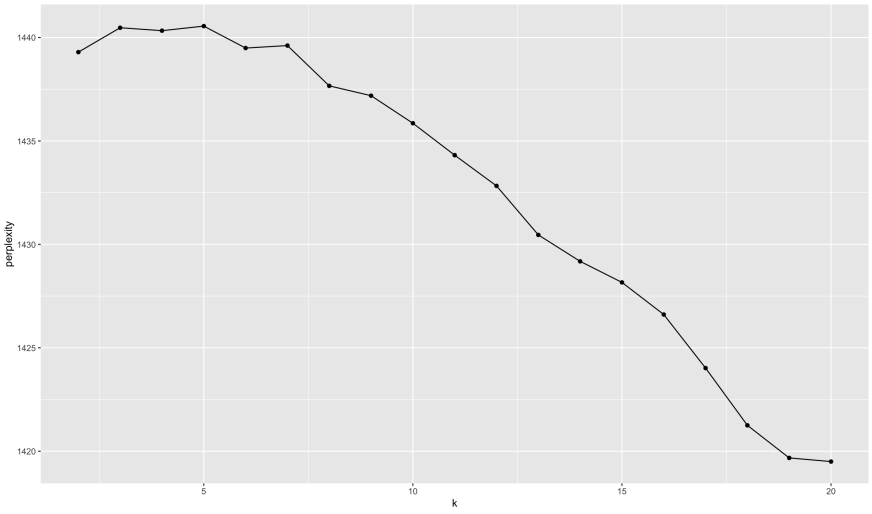


(b) Pacifier

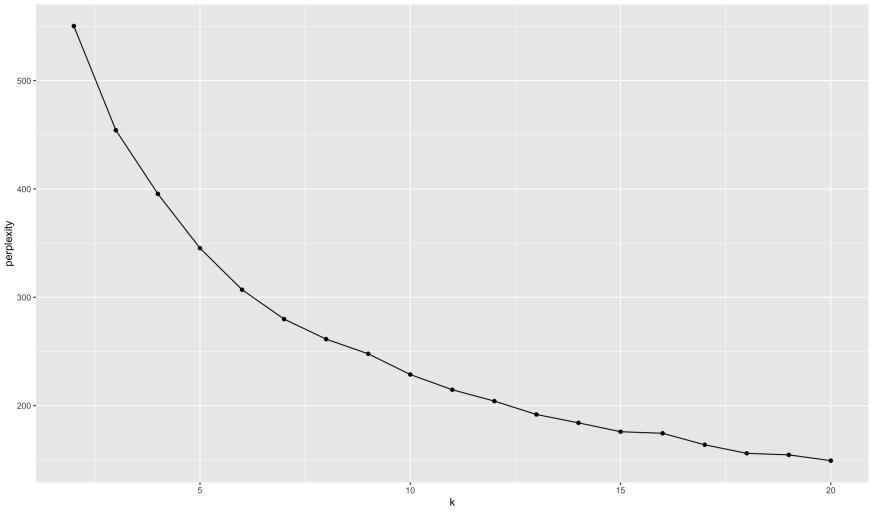


(c) Microwave

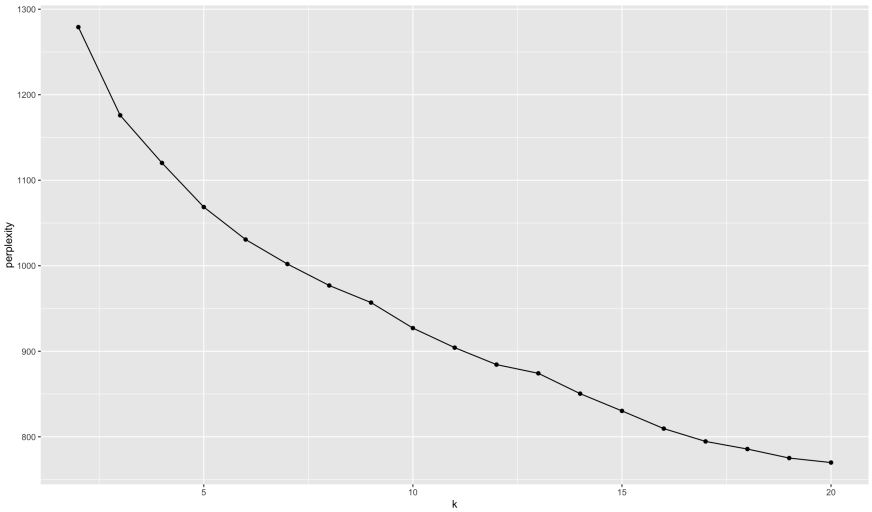
Figure 3: Perplexity score against k for 3 products:



(a) Hair dryer

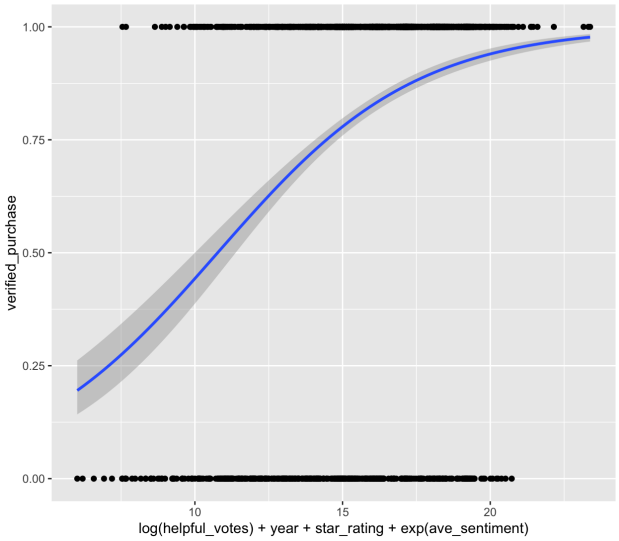


(b) Pacifier

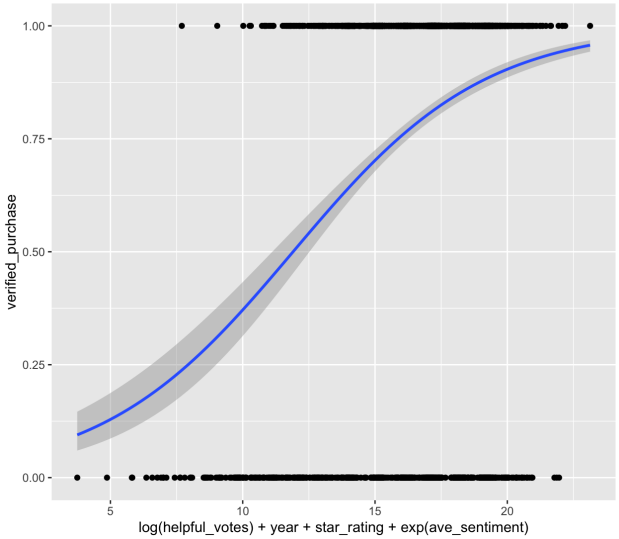


(c) Microwave

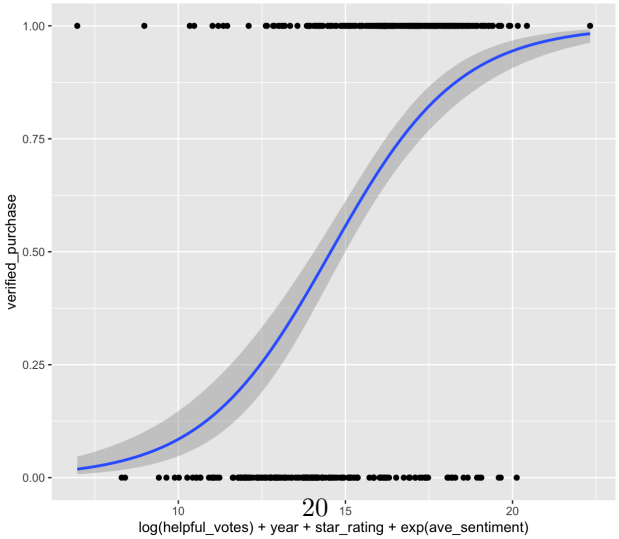
Figure 4: Logistic regression model for 3 products:



(a) Hair dryer



(b) Pacifier



(c) Microwave

7 References

[1] Rinker, T. W. (2019) *sentimentr: Calculate Text Polarity, Sentiment version 2.7.1.* `http://github.com/trinker/sentimentr`.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. *"Latent dirichlet allocation." Journal of machine Learning research*, 3 Jan 2003: 993-1022.

[3] Steyvers, Mark, and Tom Griffiths. *"Probabilistic topic models." Handbook of latent semantic analysis* 427.7 (2007): 424-440.

[4] Fang, Anna, and Zina Ben-Miled. "Does bad news spread faster?." *2017 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2017.