THE UNIVERSITY OF CHICAGO

What Makes Consumer Reviews Trustworthy?

By

Li Liu

May 2020

A thesis submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Philip Waggoner

Preceptor: Joshua Mausolf

What Makes Consumer Reviews Trustworthy?

Li Liu*

6/1/2020

Abstract

Consumer reviews are ubiquitous in online shopping. They help consumers evaluate product qualities and make purchase decisions. However, consumers are also complaining that many reviews are likely to be untrustworthy. While they read reviews, they can vote for trustworthy reviews by clicking the "Helpful" button. Using the number of votes as the proxy for review trustworthiness, this paper develops an ensemble variable selection method to evaluate what characteristics of the reviews form the trustworthiness. Based on the analysis using sample reviews on Amazon, the paper highlights that consumers' trust formation for reviews is likely conditional on the review ratings. The method and result could potentially improve Amazon and other e-commerce websites' algorithm for recommending most trustworthy reviews for consumers.

JEL classification: M31, C21, L81

Keywords: Consumer Reviews, Variable Selection, Text Mining

*M.A. Student in Computational Social Science, The University of Chicago
I would like to thank Philip Waggoner, Joshua Mausolf, and classmates in Computational Social Science M.A. program for helpful comments and suggestions. All remaining errors are mine.
Email: lliu95877@gmail.com

Replication files are available at Github.

1 Introduction

Consumer reviews are ubiquitous in online shopping websites, such as Amazon. They show the real user experience of previous consumers to the potential consumers. As a result, they serve as the wisdom of crowd that are supposed to be trustworthy and helpful. However, consumers are losing trust in reviews as they think many reviews are fake. Studies have shown that people are more skeptical about whether the positive reviews are legitimate as companies inflate ratings by rewarding consumers if they post a positive review (Dragon, 2016). Even for negative reviews, many studies have argued that consumers should not really trust negative online reviews (Beaton, 2018).

My research question is what makes some reviews more trustworthy than others? Also, whether consumers' criteria for trustworthy reviews differ when they selectively read the reviews by star ratings?

Review trustworthiness is estimated by the adjusted number of "Helpful" votes. While consumers read reviews, they can vote for a trustworthy review by clicking the "Helpful" button. Thus, the reviews for one product would have variations in the number of votes. After adjusting for the potential bias from different written time and displayed positions, I constructed a trustworthiness rank for the reviews of the sample review data (around 3 thousand reviews for one popular keyboard product).

By applying text mining methods, I extracted characteristics variables from the sample review data that essentially describe the most straightforward information that consumers would pay attention to when reading reviews. To discover the most predictive variables for the trustworthiness, I developed an ensemble variable selection method that aggregates results from three methods (random forest, multiple linear regression, and LASSO regression). I found that consumers' trust formation for reviews is likely conditional on the review ratings. In particular, when reading positive reviews, consumers trust more with reviews that are shorter and more readable. In comparison, when reading negative reviews, they look for reviews that are shorter but with attached images. After validating the results with a new sample review data (another keyboard product with 4 thousand reviews), I found initial evidence that consumers' criteria for trustworthy reviews differ

when they read the reviews with different star ratings.

The method and findings will be helpful for helping data scientists and software engineers at e-commerce websites to build consumers' trust in reviews. For example, they can use the method to nudge consumers into writing better reviews. Also, they can further improve the text generation algorithm for summarizing the consumer reviews by putting more weights to the most important characteristics variables.

2 Literature Review

Broadly speaking, this paper contributes to the literature on product reviews and UGC (user-generated content) by discovering what variables of reviews make them more trustworthy. Traditionally, in order to study what affects consumers' purchase decisions, researchers would conduct surveys, focus groups, or lab experiments. Using computational methods, many recent studies in business and economics have illustrated that reviews are useful for studying consumer behaviors in online shopping.

2.1 Text as Data for Social Scientists

The advances in natural language processing, combined with traditional data and models, have many applications in the social and business research. By extracting information from the text as new input variables, text are new sources of data for social scientists (Gentzkow, Kelly, & Taddy, 2019). Computational approaches to text analysis help us discover interesting patterns of the social world (Evans & Aceves, 2016).

Consumer reviews, as one common example of the user-generated content, is valuable for capturing marketing insights and creating values for both businesses and consumers (Balducci & Marinova, 2018). User-generated content (including reviews, images, etc) analyzed with machine learning methods is a better alternative to identify consumer needs, compared with interviews and focus groups (Timoshenko & Hauser, 2019).

2.2 Studies on Review Helpfulness

Many studies have explored what are the important variables for helpful reviews (Trenz & Berger, 2013). These variables are usually extracted from the review text as consumers read them rather than just looking at the summary statistics (Chevalier & Mayzlin, 2006).

One early study used the regression model and found review extremity, review depth, and product type affect the review helpfulness (Mudambi & Schuff, 2010). Textual features such as polarity, subjectivity, entropy, and reading ease are generated and proven to help predict the helpfulness by ensembles learning techniques (Singh et al., 2017).

The limitation in the existing literature is that the important variables are determined usually from a small number of variable candidates, which makes the results differ when using different sample data. Also, even some variables generated by natural language processing methods are predictive (such as entropy and vectors from word embedding), they are not necessarily used by consumers for their voting decisions. Thus, this paper will only consider the most straightforward information that consumers would consider when reading review.

3 Theory

I developed my theory of voting from utility-based choice modeling (Chintagunta & Nair, 2011). Consumers would only vote for a review when the utility from voting the review is greater than not voting. In other words, consumer i are faced with two decisions after reading the review: vote or not vote. Assuming the decisions are influenced only by the consumer reviews, the utility function is defined as the following:

$$u(X) = \alpha + \beta X + \epsilon \tag{1}$$

In Equation 1, X is a vector of review variables that would affect consumers' voting decision. Thus, the choice probabilities are

$$P(y = vote) = Pr(u_{vote} > u_{not}) = \frac{1}{1 + exp(-\alpha - \beta X)} \tag{2}$$

After transforming Equation 2, the logit form of the model is

$$trustworthiness \propto In(\frac{Q_{vote}}{Q_{not}}) = In(\frac{P(X)}{1-P(X)}) = \alpha + \beta X \tag{3}$$

The research question of this paper is what makes some reviews more trustworthy than others. Mathematically, the question is what are the most important elements of X in the utility function $u(X)$ defined in Equation 1. I will start with a variety of variables explored in the previous literature (Mudambi & Schuff, 2010; Eslami, Ghasemaghaei, & Hassanein, 2018; Wang, Wang, & Yao, 2019). However, since consumers would only consider some of these variables into their decision process, I applied three variable selection methods (linear regression by p-value, LASSO regression, random forest) to find the most important variables X for modeling the decision.

The second question is to test whether the X within $u(X)$ will be dependent on the star ratings. If consumers' preferences change when they filter reviews before reading the reviews, Equation 1 should be $u(X|rating)$. This implies that selecting different X when modeling consumers' behavior when reading positive, neutral, and negative reviews, respectively.

$$X|positive \neq X|neutral \neq X|negative \tag{4}$$

4 Data

4.1 Population

There are two major parts in the Amazon product review data: consumer reviews and product metadata. The data is scraped by several computer science researchers at the University of California, San Diego and made available for the public to use (Ni, Li, & McAuley, 2019).

There are two major advantages of using this large data. Firstly, it has great breadth, as it contains 233 million reviews for various products in almost 30 product categories.

Secondly, it has excellent depth, as it has reviews ranging from May 1996 to October 2018.

The short way to describe the data is that it captures all information that consumers would normally see on the product page, such as price, images, descriptions, reviews, and so on. For each product in the data, the data provides detailed information on its characteristics, such as color, price, package type, descriptions, technical details, similar products, image features, categories information, sales rank, to name a few. For each review associated with the product, it has the reviewer's name, ratings, text, summary, attached image URL, and helpfulness votes, the label for verified purchase, and so on.

Most of the existing studies with this data are in the fields of recommendation systems and natural language processing. The provider of the data Jianmo Ni and Julian McAuley have three published paper in top Computer Science conferences with this data (Ni & McAuley, 2018; Ni, Lipton, Vikram, & McAuley, 2017; Ni et al., 2019). As a result, this paper shows this data could be used for understanding consumer behavior in evaluating reviews.

4.2 Sample

For this paper, the relevant population is reviews of all online product products. However, running the analysis on the whole data is not appropriate for this paper for two reasons. The first is that the raw review data is too large in scale to process. The second is that there exist many potential con-founders, such as product types and prices, influencing how consumers vote for trustworthy reviews. Thus, the sample for this paper would be all reviews for one particular product. However, the method and analysis are applicable to all products with consumer reviews on Amazon or other online shopping websites.

When choosing the product for the sample, my priority criteria is that it should be popular and representative to make sure there are enough amount of review data. Another factor is the product doesn't have many variations in terms of the styles, size, color, versions, and so on. This ensures consumers who wrote the reviews received the same type of product.

As a result, I chose the electronics product: Microsoft Natural Ergonomic Keyboard 4000 as my sample, which is priced around $46 with a high average rating (around 4.3 out of 5) and 3 thousand reviews as shown in the product screenshot (Figure 1) and review screenshot (Figure 2) in March 2020.

Within my data for this product, there are 2567 reviews ranging from October 2005 to May 2015. The word cloud visualization of the sample reviews (Figure 3) shows consumers have diverse opinions and attitudes towards the product.

Figure 1: Product Page of Microsoft Natural Ergonomic Keyboard 4000 on Amazon

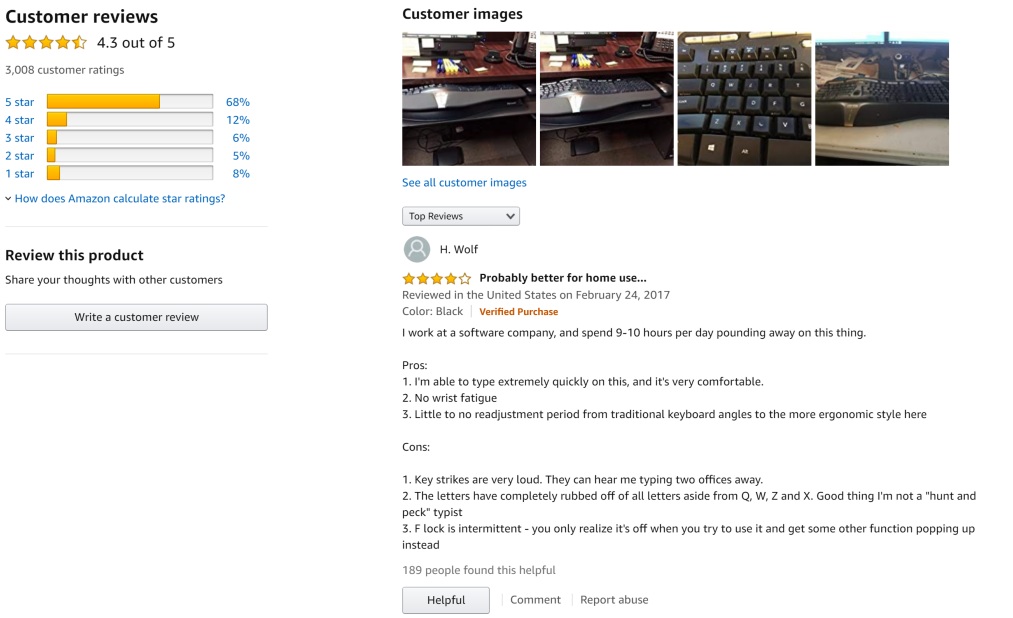

Figure 2: Sample Reviews of Microsoft Natural Ergonomic Keyboard 4000 on Amazon

Figure 3: Word Cloud Visualization of Sample Reviews

5 Variables

5.1 Trustworthiness

The outcome variable of interest is the trustworthiness of the reviews. Since the reviews are written at different time and displayed at different positions by the ranking algorithm, they don't have equal chances of being read and voted. In other words, vote counts alone is not enough for measuring the trustworthiness. As a result, I need to construct the rank of adjusted vote counts by accounting for different reviews' written dates and displayed locations.

5.1.1 Helpful Votes

When reading a review, consumers would evaluate the trustworthiness by reading the text of the reviews and vote by clicking the "Helpful" button if they find a certain review is helpful for them. For example, in the sample review page (Figure 2), the sentence on the bottom of the first review text: "189 people found this useful" suggests this review is quite trustworthy as 189 consumers have clicked the"Helpful" button.

As the helpful votes are integers within a fixed period, event count model is a good option for predicting the values (King, 1989). In particular, hurdle model or zero-inflated

model could account for excess zeros as only 6.7% of reviews have more than one vote, as shown in Figure 4.

However, the issue of estimating trustworthiness by helpful votes is that the number is likely to be biased. Firstly, the reviews with more votes tend to be older as shown in Figure 6. Secondly, consumers could only read several reviews instead of all of them. Due to the review ranking algorithm, the ones displayed on the first few pages are likely to get more votes than the ones that are ranked in the bottom few pages. As a result, the helpful votes do not precisely measure the actual trustworthiness of the reviews. For instance, a newly-written review has 0 vote currently, but it might be as trustworthy as the one which was written two years ago and got 20 accumulative votes.

In the following paragraphs, I will further analyze the effect of position and time on the helpful votes and define a formula to approximate the trustworthiness by adjusted helpful votes.

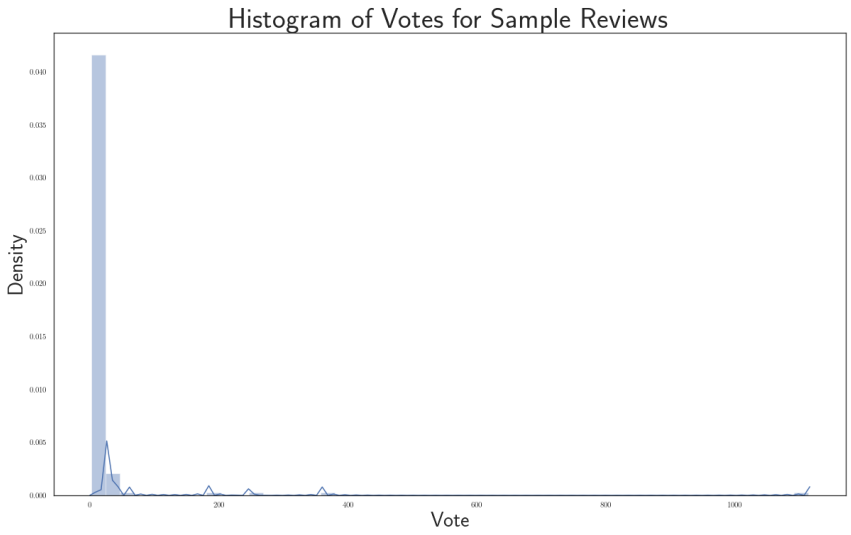

Figure 4: Histogram of Votes for Sample Reviews

5.1.2 Position Effect

To account for reviews' different displayed positions, I created a position index indicating the page number that the review is located, assuming there are 10 reviews in each page.

Figure 5 illustrates that the reviews showed up in the top and bottom pages received more votes, while most of the reviews ranked in the middle received almost no vote at all.

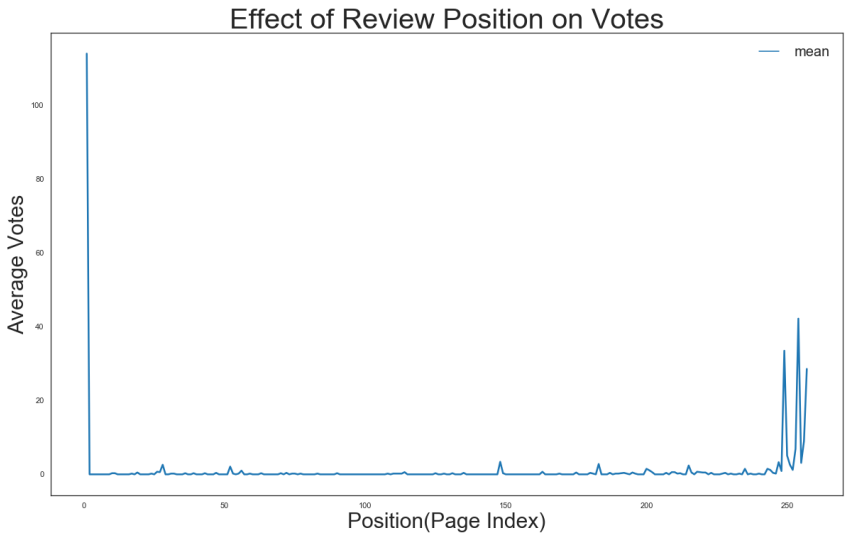

Figure 5: Effect of Review Position on Votes

5.1.3 Time Effect

The review written dates range from October 2005 to May 2015. To account for reviews' different written dates, I created a time index indicating the number of quarters since the date that first review was written. In Figure 6, the blue solid line is the average of votes at each of the 53 quarters. There is a sharp spike in the beginning quarters, but the average votes decreases to almost zero after 10th quarter. In comparison, the orange dotted line is the counts of reviews written at each of the 53 quarters. Interestingly, few consumers wrote reviews at the beginning quarters. However, during the 35th to 45th quarters, more than 100 new reviews were written at each quarter. After that period, the counts start to decrease. One possible reason is consumers didn't feel motivated to write new review when there are already thousands of reviews written by others.

5.1.4 Trustworthiness Rank

My intuition is that when quarter increases, the votes should be inflated to compensate for the shorter time of being read by consumers. When page index increases, the votes

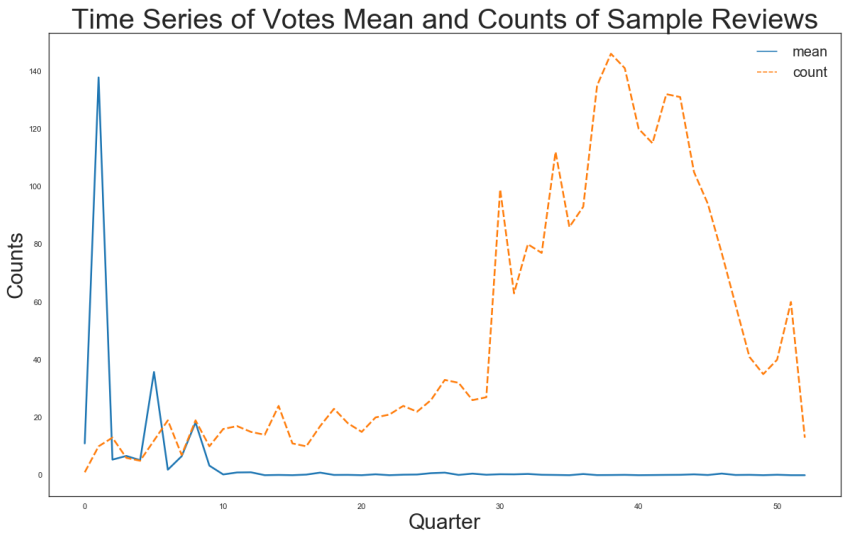

Figure 6: Time Series of Votes Mean and Counts of Sample Reviews

should also be inflated to compensate for the fewer chances of being read by consumers.

As a result, the trustworthiness score function is defined as

$$trustworthiness \propto Votes * (1 + Quarters + Position) \tag{5}$$

In Equation 5, Votes, Quarters, Positions are normalized to the scale of 0 to 1. For example, for the first written reviews on the first page, the trustworthiness score will just be the raw votes as no adjustment is needed (Votes *(1+0+0)). In comparison, for the newest written reviews shown on the bottom page, the trustworthiness score will be three times larger than the raw votes by adjusting for the time and position (Votes*(1+1+1)).

Figure 7 is a Gaussian kernel density estimate of the trustworthiness scores, compared with the vote counts. The orange curve (representing trustworthiness scores) captures the trend of the blue solid curve (representing vote counts) as well as magnifies some spikes.

Still, the trustworthiness scores are highly right-skewed after the adjustment. Thus, I used the rank of trustworthiness scores instead of the numeric scores. This helps to overcome the issue of large gap among scores. The average rank of the group was used when they have some vote counts.

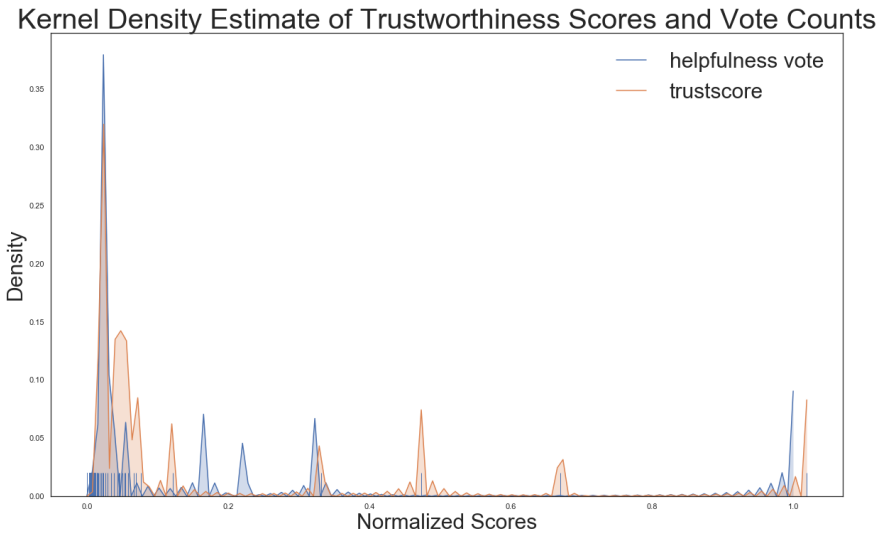

Figure 7: Kernel Density Estimate of Trustworthiness Scores and Vote Counts

5.2 Variables of Reviews

In this section, I used text mining method to extract variables to describe information that consumers would see when reading reviews. The variables are informed by previous research to further study what characteristics variables make reviews helpful.

5.2.1 Latent Topics

Different reviews talk about different topics. As a popular tool for extracting latent topics, the topic model is a type of unsupervised machine learning method that groups the words in the reviews into several latent topic groups (Alghamdi & Alfalqi, 2015). To find these latent topics, the topic model calculates the statistical correlations and groups the words that used together more frequently (Gentzkow et al., 2019). In particular, as each review often contains several topics, the model will assign different distributions to each review for a given number of topics.

There are three major steps for running topic model on the review text. Firstly, I preprocessed the text of the sample reviews, including tokenization, lemmatization, stemming, removal of short words, punctuation, and stopwords. This procedure ensures the latent topics can capture the most essential information in the reviews. Secondly, I used

TF-IDF (frequency-inverse document frequency) (Li, 2018) to re-calculate each word's frequency. The TF-IDF scores measure how important a word is to a review in the corpus. Thirdly, by using the Latent Dirichlet Allocation model (LDA) (Blei, Ng, & Jordan, 2003), I extracted the proportions of latent topics for each review. LDA is a specific type of probabilistic topic model where each topic is a combination of keywords and each keyword contributes a certain weight to one topic (*Topic Modeling in Python with Gensim*, 2018).

Further, I chose the optimal number of latent topics k by calculating both log perplexity values and coherence scores for models with different number of topics (*gensim: topic modelling for humans*, 2020). The steps are summarized as the following:

1. Split prep-processed text randomly into training (80%) and testing sets (20%)
2. For k in a range of values ([2,8] in this case), calculate the log perplexity score and coherence scores of the associated topic model (*gensim: topic modelling for humans*, 2020).
    - Coherence score measures the semantic similarity among top words of the topics (TechnovativeThinker, 2019). A higher score indicates better model performance.
    - Perplexity score measures the uncertainty of the model on predicting unseen data. A lower score indicates better model performance (*Topic Modeling in Python with Gensim*, 2018).
3. Plot the coherence score (Figure 8) and log perplexity score (Figure 9).
4. Pick optimal k based on the following considerations
    - Find k that produces high coherence score and low perplexity score
    - Check whether topic allocations are robust in the neighborhood of k, such as $k-1$ and $k+1$
    - Link with social theory or knowledge about the generation process of the text

5. Creat k new variables for each review with the topic ratios (the sum of ratios is 1)

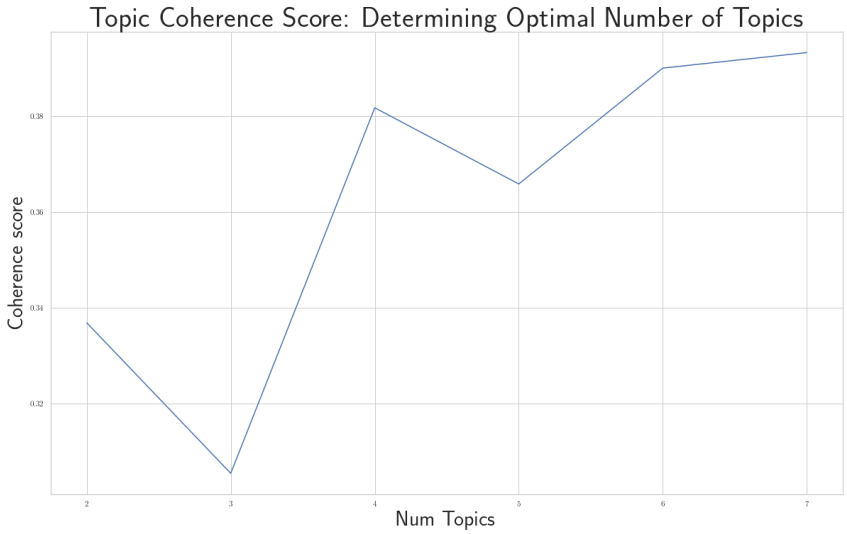

Figure 8: Coherence Scores for Determining Option Number of Topics

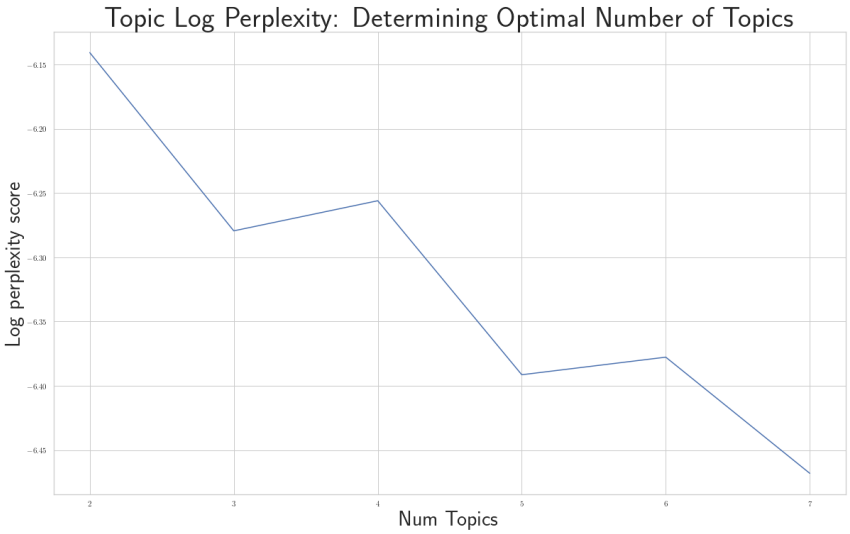

Figure 9: Log Perplexity Scores for Determining Option Number of Topics

In the topic visualizations of LDA model by pyLDAvis (Python library for interactive topic model visualization) (Mabey, 2020), each bubble on the left-hand side plot represents a latent topic and the right-hand bar chart lists the weights of the associated words for each latent topic (See Appendix A).

The most relevant terms of the first latent topic, as presented in Figure 25, are "space", "key", "type", "microsoft", "great", "button". This topic is likely about the functionality of the keyboard. Since keyboard is a material product, consumers talk about how this keyboard actually performs. The most relevant terms of the second latent topic, as presented in Figure 26, are "good', "year', "wrist', "ergo', "spacebar', "home". This topic is likely about the positive experience of this keyboard. The most relevant terms of the third latent topic, as presented in Figure 27, are "mouse', "excel', "actual', "previous', "lock". This topic is likely about the performance of the keyboard when working with the mouse. The most relevant terms of the fourth latent topic, as presented in Figure 28, are "love', "stiff', "quick', "favourite". This topic is likely about strong subjective feeling about the keyboard.

For example, the following is an extract of one review selected randomly from the sample reviews:

"*...Absolutely love this keyboard. I had an older one and the letters were worn off. Didn't bother me, but my wife was very pleased when she saw that it had the letters still on the keys. This keyboard feels great and is just designed so well....*"

For this sample review, the five topic ratios are 11%, 42%, 11%, 36%. The ratios suggest this review mainly talks about user experience and subjective feeling about the keyboard components. Also, the histograms of the topic ratios in Figure 10 suggest that the most prevalent topic is Topic 2, followed by Topic 4, Topic 3, and Topic 1, respectively.

5.2.2 Number of Associated Images

Consumers could upload images as the attachments to the reviews. As shown in Figure 2, there are 3 images associated for the first review. However, although images usually make a review more trustworthy, attaching images requires extra effort from the consumers. As a result, 99% of the reviews in the sample don't have associated images and the average number of associated images is only 0.6.

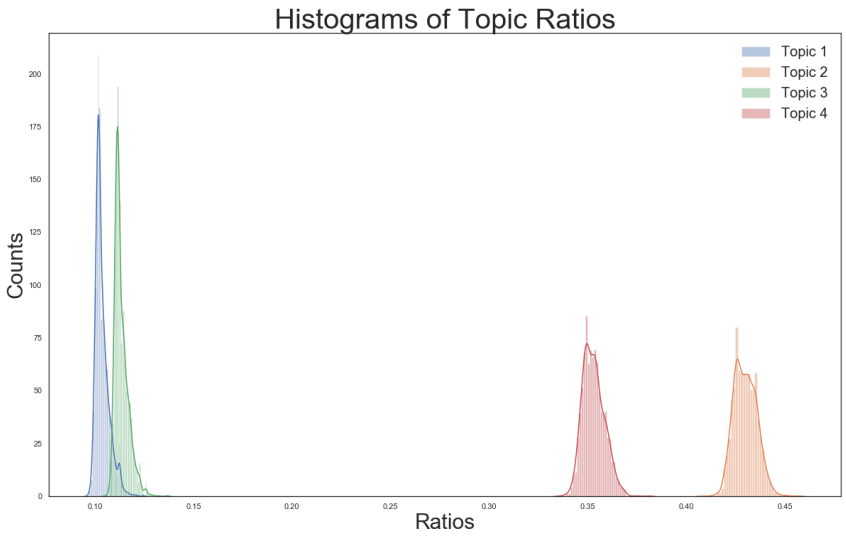

Figure 10: Histograms of Topic Ratios

5.2.3 Verified Purchase

Amazon introduced the the verified purchase labels option around November 2016. From the official definition on Amazon (Service, 2020), consumers could voluntarily add this label to their review only if they purchased the products at Amazon and didn't get the price at a deep discount. Amazon implemented this option to make reviews more trustworthy. In particular, fake reviews are hopefully excluded and consumers can rely on this label to trust reviewers' qualification of evaluating the product. 86% of the reviews in the sample data have this label.

5.2.4 Reviewer Identity

Reviewers' identity information is influential for changing other consumers' judgment of products and reviews (Forman, Ghose, & Wiesenfeld, 2008). When submitting a Amazon review, consumers can choose to either keep their name or remain anonymous when their reviews are shown to others. However, it is likely that consumers would perceive reviews written by people who share their names are more trustworthy. In order to determine whether a name is "real" or not, I define a "real name" by two conditions:

1. The gender of the first part of the name could be identified by gender-guesser

algorithm (PyPI, 2016);

2. The name has at least two parts (ex. first name, middle name, and last name).

For example, a username like "Ben Wills" is predicted to be a real name as "Ben" could be guessed as a male name and the name has two parts. However, a username like "Amazon customer" will be classified as not real name as the gender of "Amazon" is ambiguous. Finally, I classified 7.6% of the reviewers' names as "real".

5.2.5 Sentiment

The consumers' sentiment matter for trustworthiness as consumers might believe the extreme sentiments are more authentic. I applied the VADER algorithm (Valence Aware Dictionary and sEntiment Reasone) to infer the sentiment level in the review. VADER is a lexicon and rule-based sentiment analysis tool trained with social media data (Hutto & Gilbert, 2014) and returns the magnitude of the sentiment intensity normalized between -1 and 1. Both extremely positive (0.99) and negative (-0.99) reviews exist in the sample. The distribution is left-skewed and most of the reviews are positive (average is 0.45), as shown in Figure 11. Moreover, Figure 12 shows the density of the sentiment scores across different star ratings. Densities for reviews with low ratings are spread-out more widely than the ones with high ratings, suggesting consumers have various levels of sentiments expressed in the negative reviews.

5.2.6 Length

On the one hand, consumers would think lengthy reviews to be more trustworthy as reviewers put more sincere effort in writing. On the other hand, consumers would skip reading them at all as they don't want to put too much effort in reading reviews. I used the number of unique words as the measure of the review length. Figure 13 (Gaussian density plot) shows that the average length is 81 and the distribution is right-skewed.

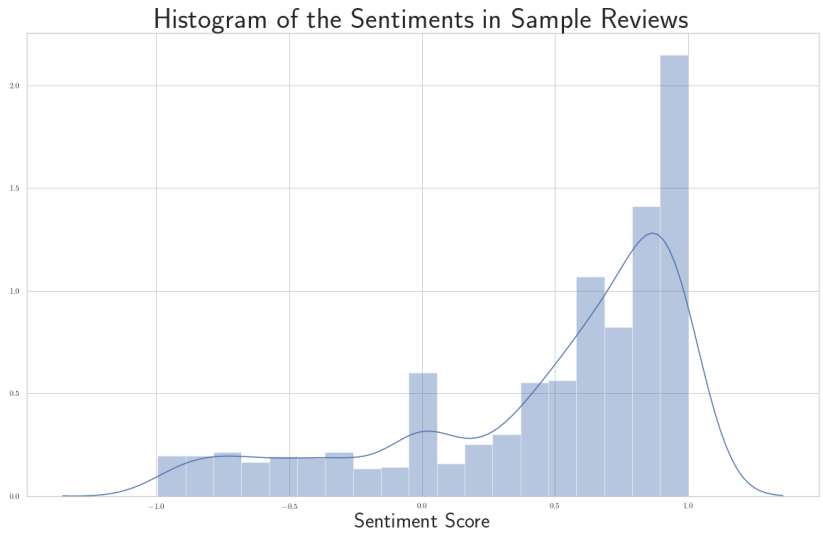

Figure 11: Histogram of the Sentiments in Sample Reviews

Figure 12: Densities of Sentiment Scores by Star Ratings

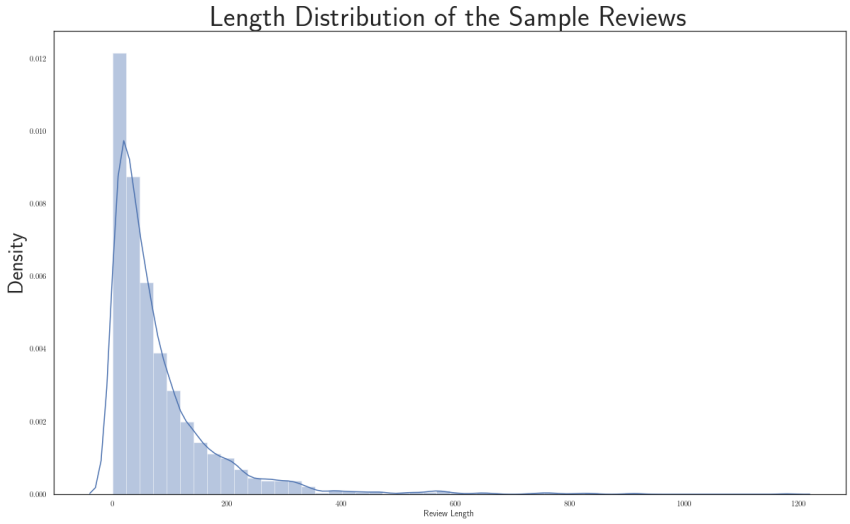

Figure 13: Length Distribution of the Sample Reviews

5.2.7 Counts of Digits

Counts of digits are quantified evidences or measures that consumers use when writing reviews. For example, one of the sample review writes:

"*...spend anywhere between 5-8 hours typing....*"

In particular, "5-8" are two digits that make the reviews more trustworthy as the reviewers are more likely to have used the product. Figure 14 shows 27% of the reviews have digits within the text.

5.2.8 Readability

The Flesch score (Aggarwal, 2020) is a common statistic of the text's reading easiness. The intuition is that consumers trust more readable reviews. As shown in Figure 15, the average score is 54 ('Fairly Difficult'). Although most of the reviews are well-written (with positive scores), the long left tail in in Figure 15 suggests few reviews are very bad written.

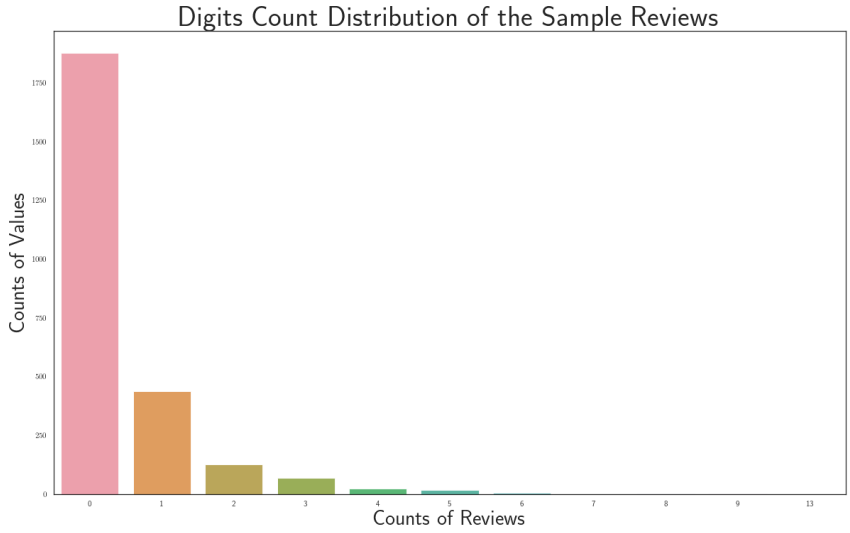

Figure 14: Digits Counts Distribution of the Sample Reviews

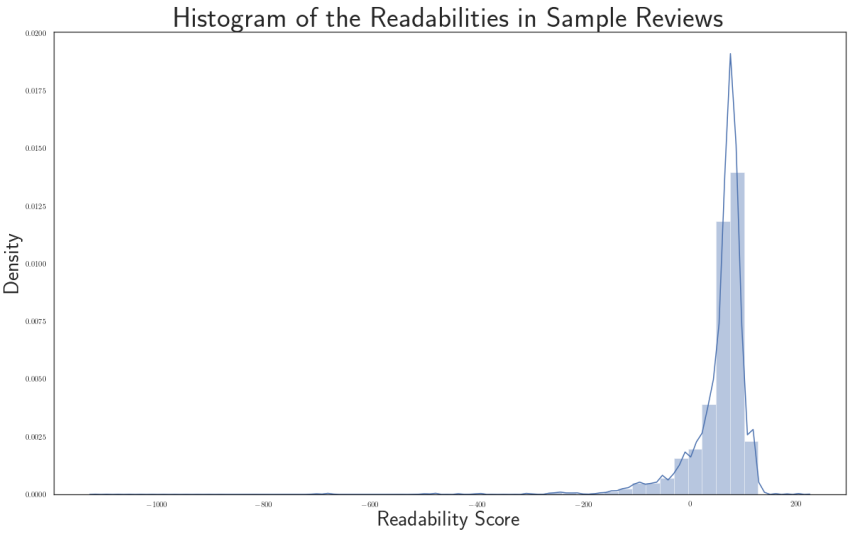

Figure 15: Histogram of the Readabilities in Sample Reviews

5.2.9 Length and Sentiment of Summary

Each review has a summary associated with it. Reviewers usually summarize their opinions and recommendations in one or two sentences in this part. For instance, the summary shown in Figure 2 is:

"*Probably better for home use...*"

Similarly to reviews text, I created two variables for the summary part: length (counts

of tokens) and sentiment. The histogram (Figure 16) shows that the length of summary is right-skewed and some summaries are quite lengthy. Figure 17 shows that 75% of the summary sentiments are positive.

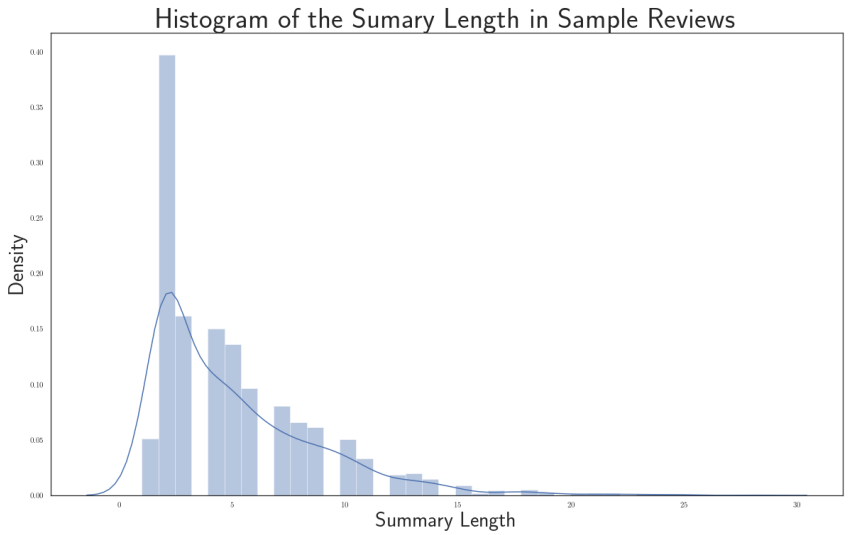

Figure 16: Histogram of the Summary Length in Sample Reviews

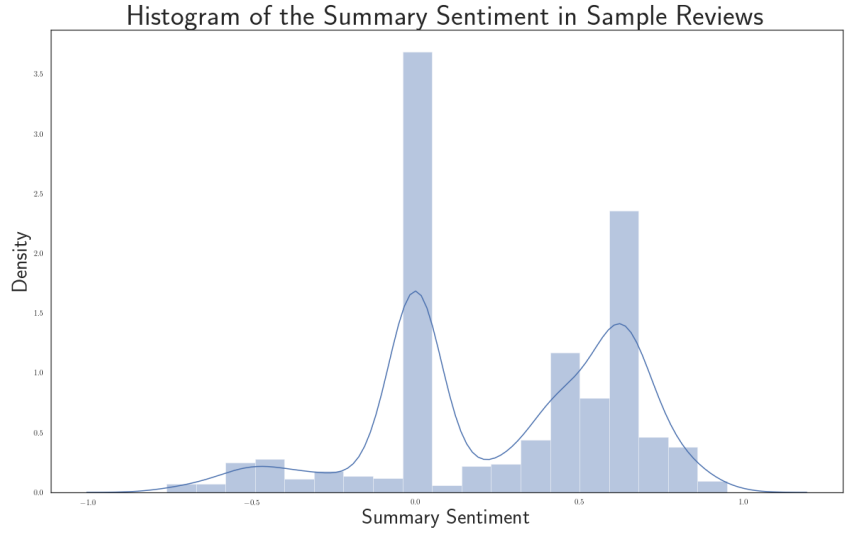

Figure 17: Histogram of the Summary Sentiment in Sample Reviews

5.2.10 Star Rating

Besides from writing the reviews, reviewers would also give a overall rating of the product and their experience. The number of 5-star reviews is 10 times as many as the number of 1-star reviews, as presented in (Figure 18). From my own online shopping experience, I would filter the reviews by ratings and then read the top reviews with different expectations in each categories (such as positive, neutral, and negative reviews). Thus, in the later analysis, I will divide the sample data by star ratings to further understand whether consumers' trust formation for reviews is conditional on the review ratings.

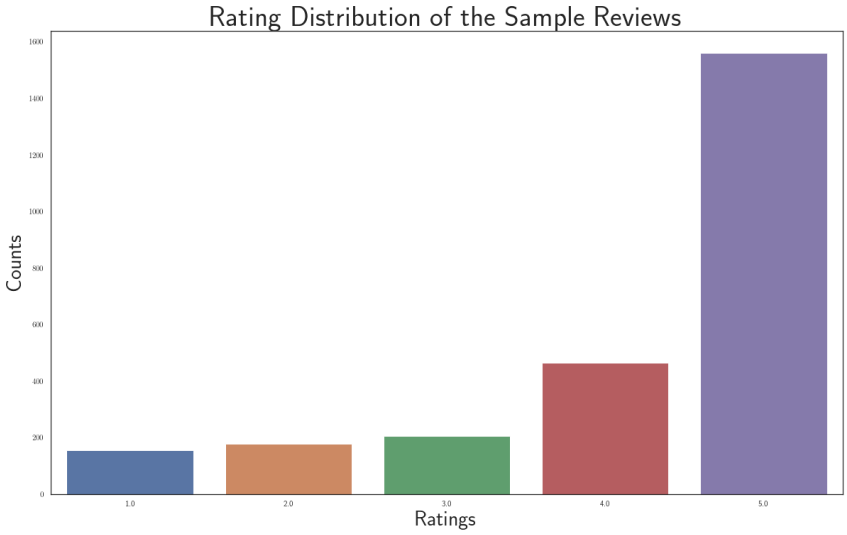

Figure 18: Rating Distribution of the Sample Reviews

6 Variable Selection

Consumers have limited time to process information presented to them. As a result, this paper focuses on finding what characteristics variables matter the most when they are evaluating reviews' trustworthiness.

6.1 Exploratory Data Analysis

The 14 variables described above are the candidates to model the review trustworthiness. As shown in the Table 6 and Figure 29, all variables are numerical and do not highly correlate with each other (cross-correlations are all less than 0.5). However, since the variables are in different scales, I normalized each variable to the 0 to 1 range.

6.2 Models for Selecting Variables

Selecting variables means selecting the ones with strongest relationships with the outcome variable. The major benefit of reducing the variable number is to obtain a more interpretable model for modeling consumers' review reading behavior. (James, Witten, Hastie, & Tibshirani, 2013).

6.2.1 Linear Regression

Multiple linear regression fits the n variables in the following form:

$$Y_i = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ji} + \epsilon_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_n X_{ni} + \epsilon_i \tag{6}$$

The estimated results of Equation 6 include the p-values of each variable's coefficient mean estimate. P-value is defined as the smallest level α_0 so that we could reject the null hypothesis at level α_0 (DeGroot & Schervish, 2011). I chose α_0 to be 0.05 (equivalent to 95% confidence interval) and the variables that are statistically significant at 0.05 level (from Table 1) are "Rating", "Review Length", "Summary Length", and "Summary Sentiment".

6.2.2 LASSO Regression

Multiple linear regression takes all features as input, which leads to potential over-fitting issue. Also, statistically significant variables are not guaranteed to be the optimal predictors for either modeling or prediction.

As a result, I applied the LASSO regression (Tibshirani, 1996) to filter out some

	Coefficients	Std	p_value	Significant at 95% level
Intercept	-18671290.0	28201322.0	1.0	0
Rating	-287.0	63.0	0.0	1
Image Number	-224.0	393.0	1.0	0
Verified Name	-7.0	35.0	1.0	0
Verified Purchase	-63.0	49.0	0.0	0
Sentiment	13.0	67.0	1.0	0
Review Length	2298.0	249.0	0.0	1
Digits Count	189.0	230.0	0.0	0
Readability Score	-322.0	384.0	0.0	0
Summary Length	-546.0	124.0	0.0	1
Summary Sentiment	258.0	85.0	0.0	1
Topic1	8813402.0	13310968.0	1.0	0
Topic2	17008831.0	25687993.0	1.0	0
Topic3	11219900.0	16945216.0	1.0	0
Topic4	15725887.0	23750435.0	1.0	0

Table 1: Linear Regression Results for All Reviews in the Sample

variables by imposing an L_1 norm penalty on the regression coefficients. In the ordinary least square, the linear regression tries to minimize the residual sum of square in Equation 7.

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{7}$$

In comparison, the new objection function (Equation 8) to minimize in LASSO regression is the residual sum of square and the penalty term (penalty weight λ times the L_1 norm) (James et al., 2013).

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{8}$$

Figure 19 shows that coefficients of some variables converge to zero as λ increases. With larger λ (more penalization), fewer variables will be kept. However, with smaller λ, the regularization term doesn't affect the results much, so it won't help with selecting the important variables (James et al., 2013).

To find the optimal λ, I tuned the penalty weight λ by grid search method. By iterating through the range of values from 0 to 10 with 50 interval values, this tuning method eval-

uates the performance by the average mean square error of 5-fold cross validation (Geron, 2017). Finally, the λ that yields best prediction result is 2.45 and variables that are "important" (with non-zero coefficients) at this λ are "Review Length', "Rating', "Summary Length', and "Summary Sentiment".

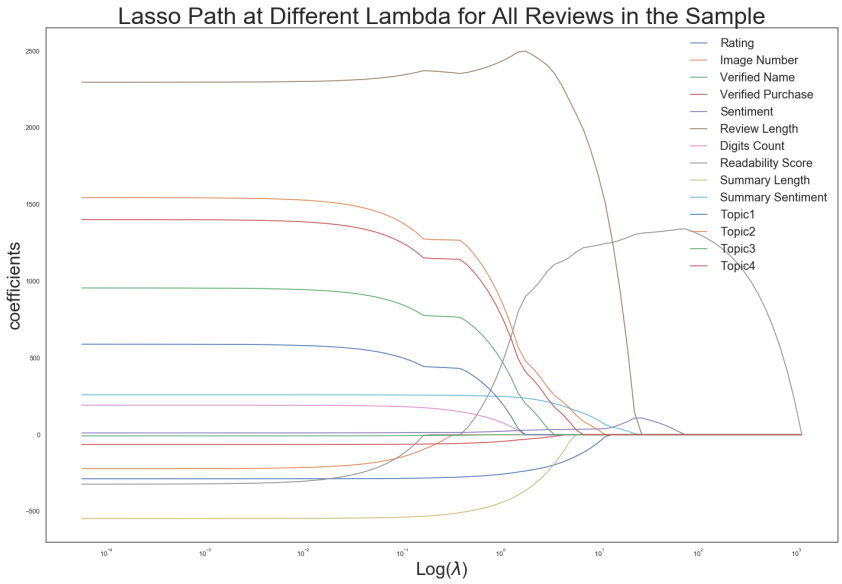

Figure 19: LASSO Path at Different Lambda for All Reviews in the Sample

6.2.3 Random Forest

Random Forest algorithm (Breiman, 2001) is a tree-aggregation method that averages the results from many de-correlated tree trained with bootstrapped subsets (Hastie, Tibshirani, & Friedman, 2009). In particular, random forests reduces variance to improve accuracy by randomly selecting only a subset of variables to consider at each node split (Breiman, 2001).

In particular, the algorithm could measure the importance of each variable by accumulating each splitting variable's improvement in split-criterion (Hastie et al., 2009). After the iterations, I found the top 5 important predictors are "Review Length", "Topic 4", and "Topic 1", "Topic 2".

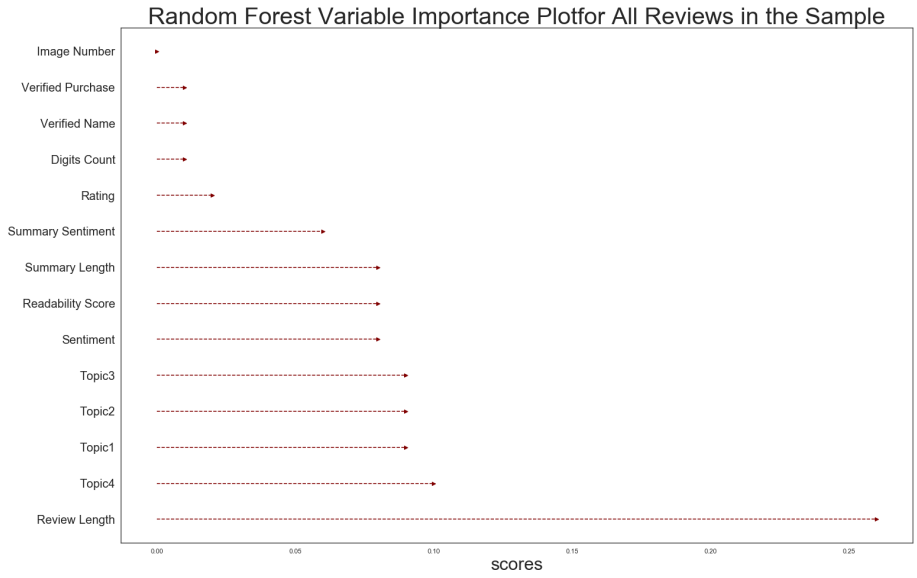

Figure 20: Random Forests Variable Importance Plot for All Reviews in the Sample

7 Ensemble Results

As shown in Table 2, the results from three methods (Linear Regression, LASSO Regression, Random Forest) are different. Motivated by the idea of ensemble learning in machine learning (Witten, Frank, Hall, & Pal, 2016), combining the results of multiple models would achieve better result than using just one single model.

	random_forest_imp	linear_imp	lasso_imp	aggregated_results
Review Length	1	1	1	3
Rating	0	1	1	2
Summary Length	0	1	1	2
Summary Sentiment	0	1	1	2
Topic1	1	0	0	1
Topic2	1	0	0	1
Topic4	1	0	0	1
Image Number	0	0	0	0
Verified Name	0	0	0	0
Verified Purchase	0	0	0	0
Sentiment	0	0	0	0
Digits Count	0	0	0	0
Readability Score	0	0	0	0
Topic3	0	0	0	0

Table 2: Ensemble Variable Importance Results for All Reviews in the Sample

For each method, I indicated the "important" variables as 1 and other ones as 0, which could be understood as a "vote" for this variable. Then I aggregated the votes from the three methods to construct the relative importance score. With the scores, the variables could be ranked in terms of importance for decision modeling.

More generally, for each variable, the relative importance score is

$$\sum_{i=1}^{j} w_j * \mathbb{1}_{Important} \tag{9}$$

In Equation 9, j is the the number of all models and w_i is the associated weight for one model. As j increases, the results become more reliable by considering more votes (Witten et al., 2016). Also, the researchers could assign different weights w_i to the models if they have prior belief or preference towards certain methods. As $\sum_{i=1}^{j} w_i = 1$ and I don't have any preference towards the three methods, the aggregation will be unweighted and $w_{linear} = w_{LASSO} = w_{randomforest} = \frac{1}{3}$.

Only "Review Length" variable received all votes from three methods, indicating it to be the most important variable for modeling trustworthiness. Since the coefficients of "Review Length" in linear regression and LASSO regression are positive and the outcome is the rank, longer reviews are less trustworthy for consumers.

"Rating", "Summary Length", and "Summary Sentiment" received 2 out of 3 votes, indicating they are fairly important as well. Since these variables are shorter information that review writers use to summarize the review paragraph, this result suggests that first impression of the review plays an important role.

Among the three methods, LASSO regression performs the best as its indicator variable correlates best with the aggregated result (0.79).

8 Results by Ratings

Consumers might have different perceptions of trustworthiness for reviews with different ratings. In addition, consumers usually tend to read several reviews under different rating categories. From the rating bar chart on the left panel in Figure 2, consumers can click on

each bar to display reviews under this rating.

Broadly, I categorized the reviews with ratings of 4 and 5 as positive reviews (79%), reviews with ratings of 3 as neutral reviews (8%), and reviews with ratings of 1 and 2 as negative reviews (13%).

8.1 Positive Reviews

Table 3 shows the results of selecting the important variables for positive reviews. "Review Length" variables received all votes from three methods, indicating it to be the most important variable for modeling trustworthy in this subset. Interestingly, the length of the reviews has a negative impact on trustworthiness while the impact of the length of the summaries is positive. "Readability" and "Summary Length" received 2 out of 3 votes, indicating they are fairly important as well.

	random_forest_imp	linear_imp	lasso_imp	aggregated_results
Review Length	1	1	1	3
Readability Score	0	1	1	2
Summary Length	0	1	1	2
Rating	0	1	0	1
Digits Count	0	0	1	1
Topic1	1	0	0	1
Topic3	1	0	0	1
Topic4	1	0	0	1
Image Number	0	0	0	0
Verified Name	0	0	0	0
Verified Purchase	0	0	0	0
Sentiment	0	0	0	0
Summary Sentiment	0	0	0	0
Topic2	0	0	0	0

Table 3: Ensemble Variable Importance Results for Positive Reviews in the Sample

8.2 Neutral Reviews

Table 4 shows the results of selecting the important variables for neutral reviews. Only "Review Length" variable received all votes from three methods, indicating it to be the

most important variable for modeling trustworthy in this subset. "Sentiment', and "Summary Length" also received 2 out of 3 votes, indicating they are fairly important as well.

	random_forest_imp	linear_imp	lasso_imp	aggregated_results
Review Length	1	1	1	3
Sentiment	1	1	0	2
Summary Length	0	1	1	2
Verified Purchase	0	0	1	1
Summary Sentiment	0	0	1	1
Topic2	1	0	0	1
Topic4	1	0	0	1
Rating	0	0	0	0
Image Number	0	0	0	0
Verified Name	0	0	0	0
Digits Count	0	0	0	0
Readability Score	0	0	0	0
Topic1	0	0	0	0
Topic3	0	0	0	0

Table 4: Ensemble Variable Importance Results for Neutral Reviews in the Sample

8.3 Negative Reviews

Table 5 shows the results of selecting the important variables for negative reviews. Only "Review Length" variable received all votes from three methods, which suggests it to be the most important variable for modeling trustworthy in this subset. "Image Number" and "Summary Sentiment" received 2 out of 3 votes, indicating they are fairly important as well.

8.4 Aggregated Results

Figure 21 illustrates that the important variables for trustworthiness are different for reviews with different ratings. In a nutshell, length and information of summary are definitely influential for consumers to gain trust. When they are reading positive reviews, more readable reviews are more trustful, as some fake reviews are generated by computer bots and not readable. When they are reading neutral reviews, sentiment is crucial as consumers need to tell exactly about the reviewers' attitudes. When they are reading negative

	random_forest_imp	linear_imp	lasso_imp	aggregated_results
Review Length	1	1	1	3
Image Number	0	1	1	2
Summary Sentiment	0	1	1	2
Verified Purchase	0	1	0	1
Readability Score	0	0	1	1
Topic1	1	0	0	1
Topic2	1	0	0	1
Topic3	1	0	0	1
Rating	0	0	0	0
Verified Name	0	0	0	0
Sentiment	0	0	0	0
Digits Count	0	0	0	0
Summary Length	0	0	0	0
Topic4	0	0	0	0

Table 5: Ensemble Variable Importance Results for Negative Reviews in the Sample

reviews, reviews with images are more trustworthy as reviewers provide evidences of why the product or experience is not satisfactory.

Using dummy variables $\mathbb{1}_{neg}$ as indicators, Equation 10 extends Equation 3 by incorporating the findings in this sample data to better model of how review trustworthiness is formed.

$$\begin{aligned} trustworthiness = \alpha + \beta_1 * length + \beta_2 * ImageNumber * \mathbb{1}_{neg} \\ +\beta_3 * Readability * \mathbb{1}_{pos} + \beta_4 * SummarySentiment * \mathbb{1}_{neg} \\ +\beta_5 * SummaryLength * \mathbb{1}_{pos} * \mathbb{1}_{neu} + \beta_6 * Sentiment * \mathbb{1}_{neu} \end{aligned} \tag{10}$$

8.5 Validation

To validate whether my results on this sample data could be generalized to other samples, I applied the analysis on another keyboard product: "Logitech MK270 Wireless Keyboard and Mouse Combo". This Logitech keyboard is similar with the Microsoft keyboard in terms of functions, prices, and appearances (Figures 22 and 23). The new product has slightly more number (4624) of reviews in the data.

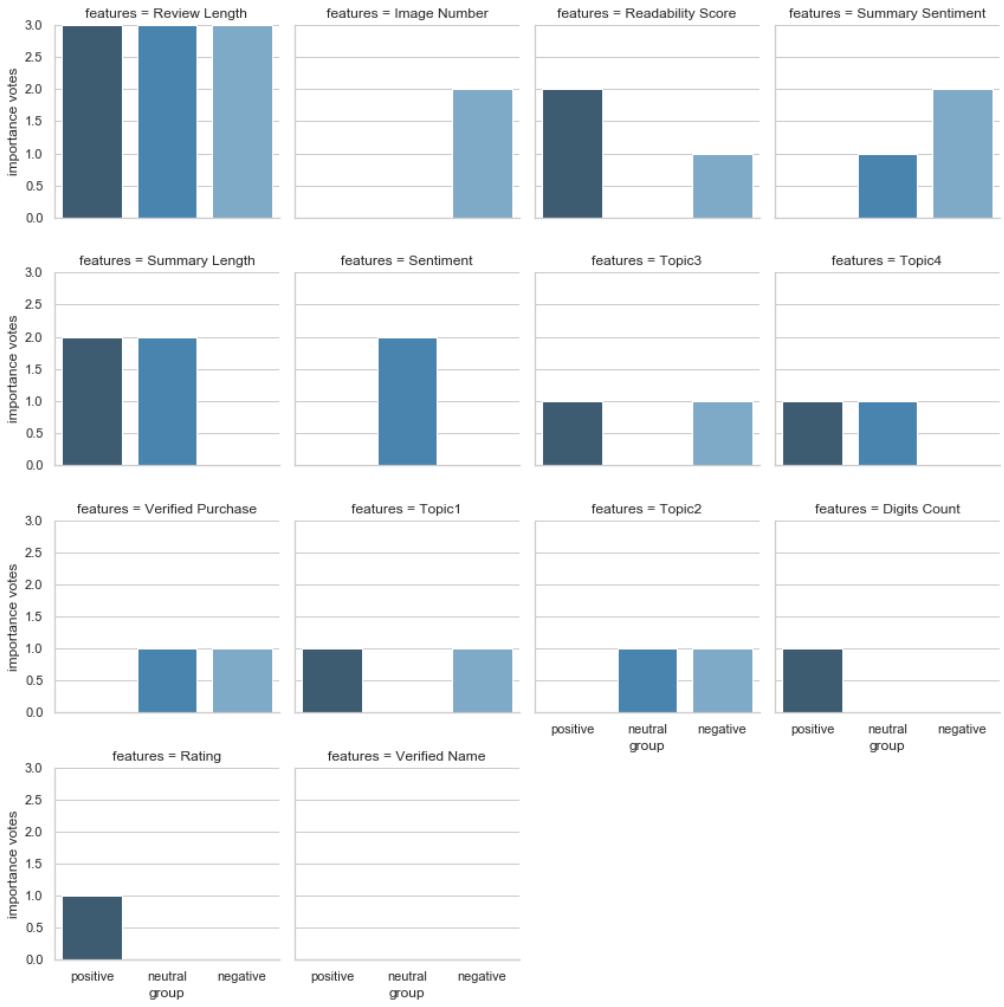

Figure 21: Importance Votes of Variables at Positive, Neutral, and Negative Reviews

Figure 22: Sample Product Page of Logitech MK270 Wireless Keyboard and Mouse Combo on Amazon

Figure 24 shows the empirical results of applying the method on the Logitech keyboard reviews. Although the selected important variables are different from the results with Microsoft keyboard sample data, the results validated that consumers pay attention

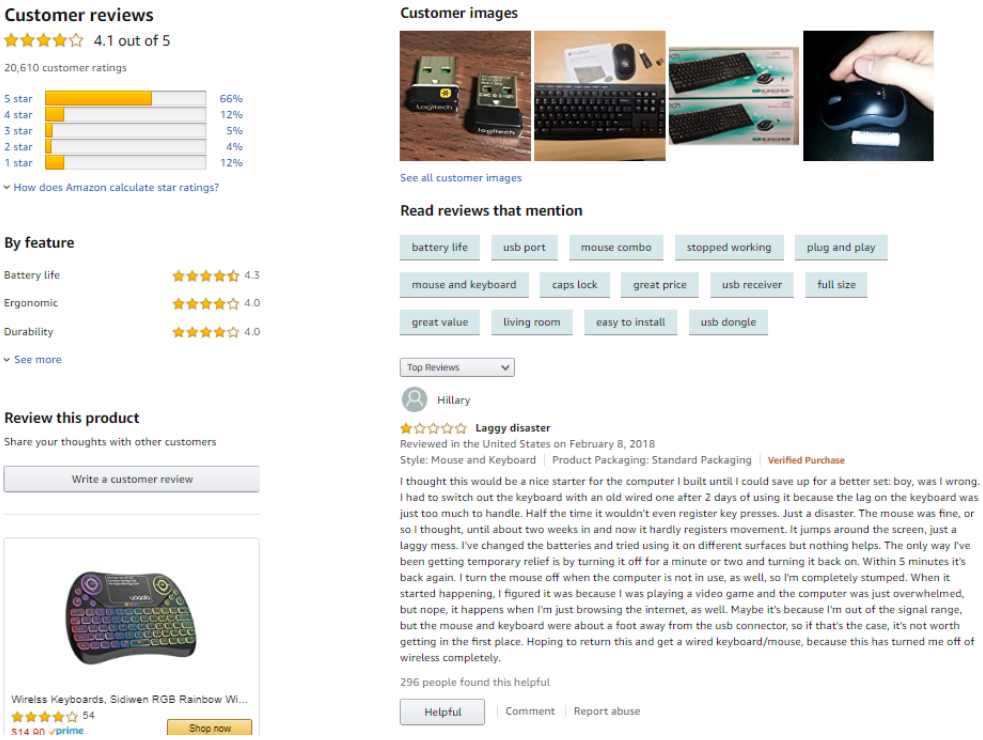

Figure 23: Sample Reviews Page of Logitech MK270 Wireless Keyboard and Mouse Combo on Amazon

to different review characteristics when reading positive, neutral, and negative reviews, respectively.

9 Discussion

9.1 Small Sample

Currently, the results are just based on reviews of two popular keyboard products. The results might be different for products in other categories. For example, number of attached images are only important for this keyboard's negative reviews. However, for clothes, the images within all reviews could be informative as consumers want to know how it fits on actual consumers.

Although the analysis results might be different for other products, the methods and procedures could be easily applied for products with consumer reviews. As a result, my future research agenda is to aggregate the results for a wide range of products, such as all keyboards, electronic products, clothing, books, and so on. This links to the literature

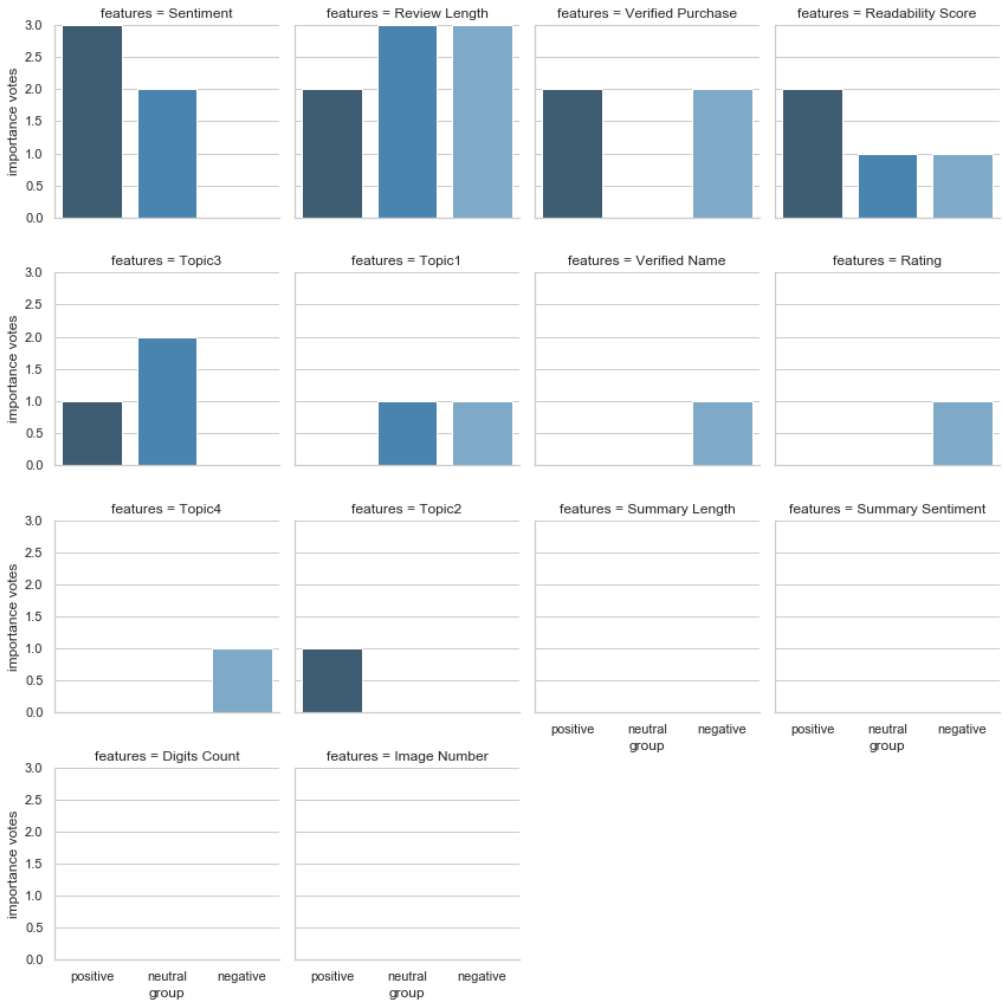

Figure 24: Logitech Keyboard: Importance Votes of Variables at Positive, Neutral, and Negative Reviews

on how product types affect people's probability in trusting reviews. For example, one study ran controlled experiments on Amazon Mechanical Turk to show people rely less on consumer reviews for experiential than material purchases (Dai, Chan, & Mogilner, n.d.). As a result, it would be interesting to further understand how consumers' trust in reviews are dependent on the product metadata.

9.2 Evolving Product Features and Prices

The reviews content are highly conditional on product features and prices, which are continuously changing. This data does not provide the associated features and prices at the time that reviews were written. As a result, evolving product features and prices are

potential unobservable confounders for this analysis.

9.3 Sampling Bias

consumers who wrote reviews might not be a representative sample of the online shoppers population. Study has shown only about 1.5% of the consumers left a review in the past 5 years on a large private label retailer's website (Anderson & Simester, 2014). They are also usually regarded as people who have extreme experience of the experience. Also, consumers might be less motivated to write reviews if there are already thousands of reviews for the product.

9.4 Different Versions of Data

Consumers might see different information under Amazon's thousands of online experiments at any given time (Clarke & Clarke, 2016). Since the data was scraped at one point by the researchers around November 2018, each consumer might see different prices and different orderings of the reviews. It is likely that the data only presents one version of many A/B tests of page information.

10 Application

10.1 Reviews Ranking

The potential application of this paper is to improve Amazon and other e-commerce websites' algorithm for ranking the reviews to display the most trustworthy ones for consumers. By using the ensemble variable selection method developed in this paper, the algorithm could efficiently find the reviews that are most trustworthy for consumers when shopping certain product.

10.2 Nudges for Reviewers

Reviewers write reviews to share their opinions with others. As a result, this result could be used to improve customer review systems by helping consumers write better reviews, which is an important but unanswered topic in related literature (Trenz & Berger, 2013). For instance, if the system finds that among the current reviews, number of images and short reviews are more trustworthy, the system can show some nudges when reviewers are writing, such as "Make your reviews stand out by uploading a image', "150-word is the right length for catching attentions".

10.3 Review Summarizing

The insights from this paper could also help to build better text summarizing applications for product reviews. The current approach in the natural language processing research community usually focus on the textual features and treat all input data as homogeneous. However, this paper shows that the reviews are heterogeneous from the consumers' perspectives and thus there should be different weights for reviews when using natural language processing techniques to summarize the reviews.

11 Conclusion

So what makes Amazon reviews trustworthy? It depends. Using the novel variable selection method developed in this paper, researchers could easily evaluate what variables of the reviews form the trustworthiness in their own review data. In particular, the method ranks the importance of the variables by aggregating results from linear regression, LASSO regression, and random forest. Based on the results from two keyboard product's reviews, the paper highlights that consumers' trust formation for reviews is conditional on the review ratings.

Appendix A Latent Topics

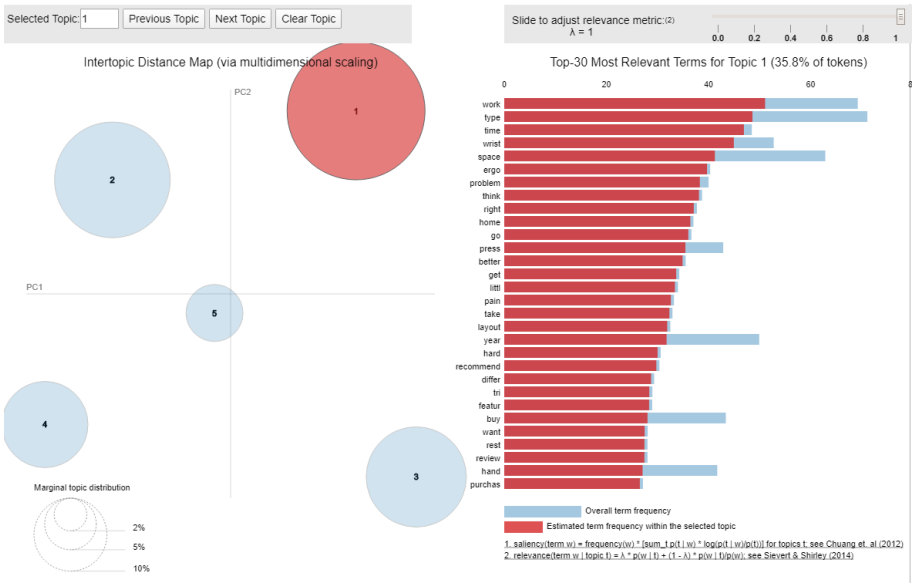

Figure 25: Visualization of Topic 1

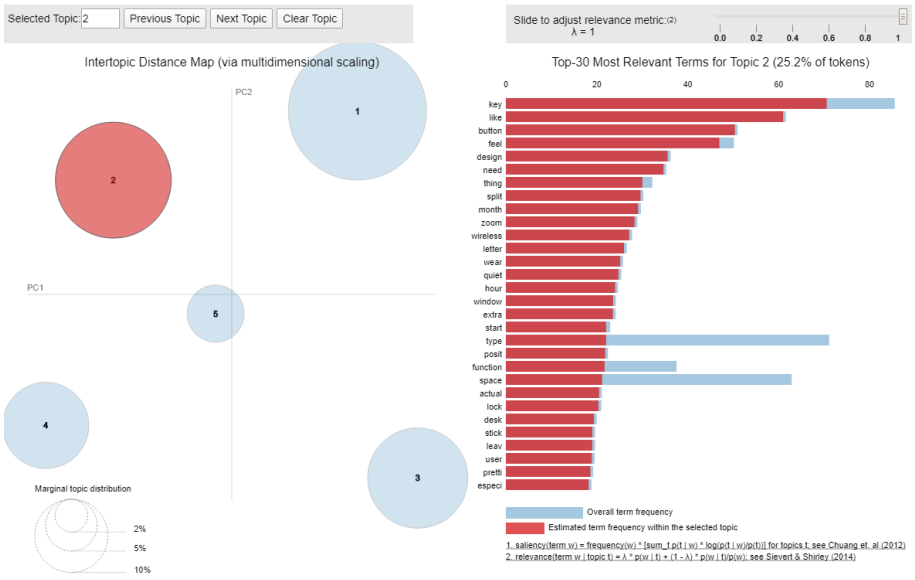

Figure 26: Visualization of Topic 2

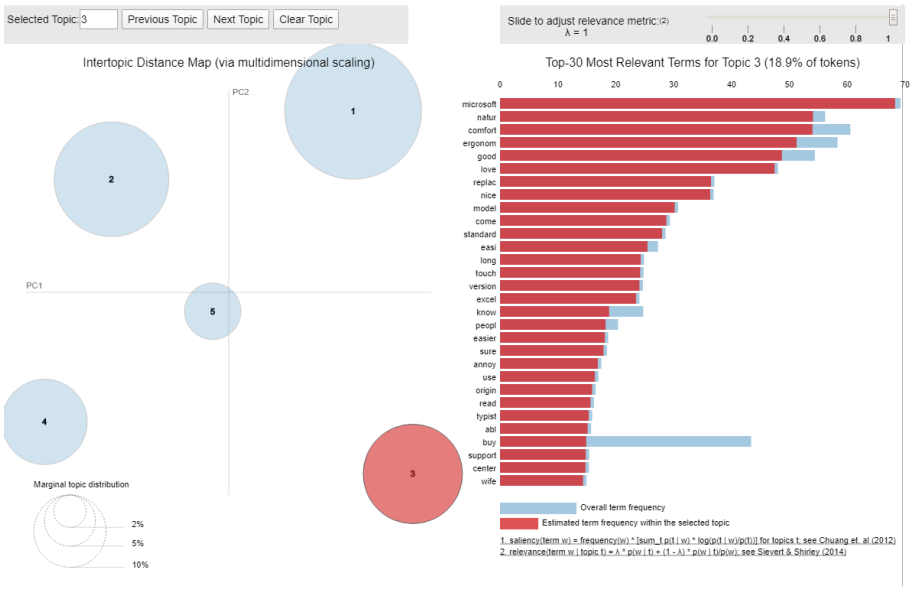

Figure 27: Visualization of Topic 3

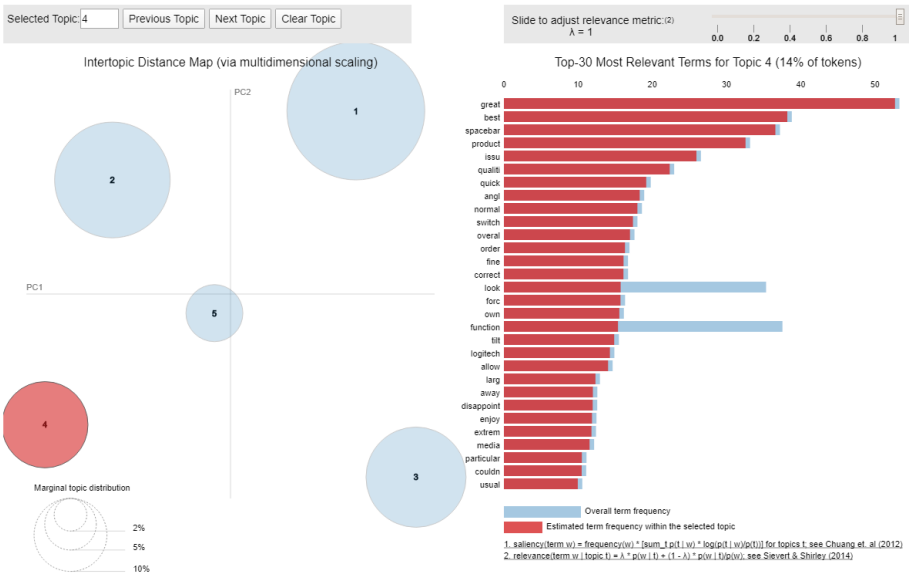

Figure 28: Visualization of Topic 4

Appendix B Summary Statistics

	mean	std	min	25%	50%	75%	max
Rating	4.21	1.21	1.00	4.00	5.00	5.00	5.00
Image Number	0.55	9.56	0.00	0.00	0.00	0.00	370.00
Verified Name	0.32	0.47	0.00	0.00	0.00	1.00	1.00
Verified Purchase	0.86	0.34	0.00	1.00	1.00	1.00	1.00
Sentiment	0.45	0.54	-0.99	0.14	0.64	0.88	1.00
Review Length	81.37	103.03	0.00	20.00	49.00	102.00	1179.00
Digits Count	0.46	1.01	0.00	0.00	0.00	1.00	13.00
Readability Score	54.32	65.76	-1108.30	46.44	71.48	83.66	206.84
Summary Length	5.08	3.70	1.00	2.00	4.00	7.00	28.00
Summary Sentiment	0.27	0.37	-0.76	0.00	0.36	0.62	0.95
Topic1	0.10	0.00	0.10	0.10	0.10	0.11	0.12
Topic2	0.43	0.01	0.41	0.43	0.43	0.43	0.46
Topic3	0.11	0.00	0.11	0.11	0.11	0.12	0.14
Topic4	0.35	0.01	0.34	0.35	0.35	0.36	0.38

Table 6: Summary Statistics of Variables Candidates

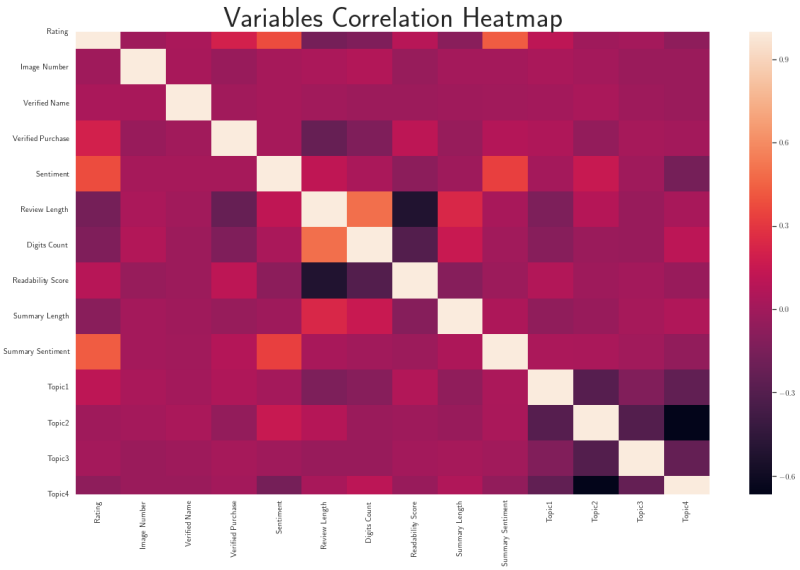

Figure 29: Variables Correlation Heatmap

References

Aggarwal, S. B., Chaitanya. (2020). *textstat: Calculate statistical features from text.* Retrieved 2020-03-29, from https://github.com/shivam5992/textstat

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1). Retrieved 2020-05-01, from http://thesai.org/Publications/ViewPaper?Volume=6&Issue=1&Code=ijacsa&SerialNo=21 doi: 10.14569/IJACSA.2015.060121

Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research*, *51*(3), 249–269.

Balducci, B., & Marinova, D. (2018). Unstructured Data in Marketing. *Journal of the Academy of Marketing Science*, *46*, 557–590. doi: (2018)46:557âĂŞ590https://doi.org/10.1007/s11747-018-0581-x

Beaton, C. (2018, June). Why You Can't Really Trust Negative Online Reviews. *The New York Times*. Retrieved 2019-11-18, from https://www.nytimes.com/2018/06/13/smarter-living/trust-negative-product-reviews.html

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. (Publisher: Springer)

Chevalier, J. A., & Mayzlin, D. (2006, August). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354. Retrieved 2019-10-30, from https://doi.org/10.1509/jmkr.43.3.345 doi: 10.1509/jmkr.43.3.345

Chintagunta, P. K., & Nair, h. S. (2011). Structural Workshop Paper—Discrete-Choice Models of Consumer Demand in Marketing. *Marketing Science*, *30*(6), 977–996.

Clarke, B., & Clarke, B. (2016, September). *Why These Tech Compa-*

nies Keep Running Thousands Of Failed Experiments. Retrieved 2020-05-01, from https://www.fastcompany.com/3063846/why-these-tech-companies-keep-running-thousands-of-failed (Library Catalog: www.fastcompany.com)

Dai, H., Chan, C., & Mogilner, C. (n.d.). People Rely Less on Consumer Reviews for Experiential than Material Purchases. *Journal of Consumer Research*. Retrieved 2020-04-06, from https://academic.oup.com/jcr/advance-article/doi/10.1093/jcr/ucz042/5567101 doi: 10.1093/jcr/ucz042

DeGroot, M. H., & Schervish, M. J. (2011). *Probability and Statistics* (4edition ed.). Pearson.

Dragon, L. (2016). *Let's Talk About Amazon Reviews: How We Spot the Fakes.* Retrieved 2019-11-18, from https://thewirecutter.com/blog/lets-talk-about-amazon-reviews/

Eslami, S. P., Ghasemaghaei, M., & Hassanein, K. (2018). Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems*, *113*, 32–42.

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*, *42*(1), 21–50. Retrieved 2020-05-02, from https://doi.org/10.1146/annurev-soc-081715-074206 (_eprint: https://doi.org/10.1146/annurev-soc-081715-074206) doi: 10.1146/annurev-soc-081715-074206

Forman, C., Ghose, A., & Wiesenfeld, B. (2008, September). Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research*, *19*(3), 291–313. Retrieved 2019-11-02, from http://pubsonline.informs.org/doi/10.1287/isre.1080.0193 doi: 10.1287/isre.1080.0193

gensim: topic modelling for humans. (2020). Retrieved 2020-05-01, from https://radimrehurek.com/gensim/models/coherencemodel.html (Li-

brary Catalog: radimrehurek.com)

Gentzkow, M., Kelly, B. T., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *forthcoming*.

Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Hutto, C., & Gilbert, E. (2014, June). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Ann Arbor, MI.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (1st ed.). Springer.

King, G. (1989). Event Count Models for International Relations: Generalizations and Applications. *International Studies Quarterly*, *33*, 123–147.

Li, S. (2018, June). *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python.* Retrieved 2020-03-18, from https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 (Library Catalog: towardsdatascience.com)

Mabey, B. (2020, March). *bmabey/pyLDAvis.* Retrieved 2020-03-28, from https://github.com/bmabey/pyLDAvis (original-date: 2015-04-09T22:48:03Z)

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com. *MIS quarterly*, *34*(1), 185–200.

Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 188–197).

Ni, J., Lipton, Z. C., Vikram, S., & McAuley, J. (2017, November). Estimating Reactions and Recommending Products with Generative Models of Reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 783–791). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved 2019-10-30, from https://

www.aclweb.org/anthology/I17-1079

Ni, J., & McAuley, J. (2018, July). Personalized Review Generation By Expanding Phrases and Attending on Aspect-Aware Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 706–711). Melbourne, Australia: Association for Computational Linguistics. Retrieved 2019-10-30, from https://www.aclweb.org/anthology/P18-2112 doi: 10.18653/v1/P18-2112

PyPI. (2016). *gender-guesser 0.4.0.* Retrieved 2019-05-21, from https://pypi.org/project/gender-guesser/

Service, A. H. . C. (2020). *Amazon.com Help: About Amazon Verified Purchase Reviews.* Retrieved 2020-03-18, from https://www.amazon.com/gp/help/customer/display.html?nodeId=202076110

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, *70*, 346–355.

TechnovativeThinker. (2019, December). *Topic Modeling: Art of Storytelling in NLP.* Retrieved 2020-05-01, from https://medium.com/@MageshDominator/topic-modeling-art-of-storytelling-in-nlp-4dc83e96a987 (Library Catalog: medium.com)

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved 2020-04-23, from https://www.jstor.org/stable/2346178 (Publisher: [Royal Statistical Society, Wiley])

Timoshenko, A., & Hauser, J. R. (2019). Identifying Customer Needs from user-Generated Content. *Marketing Science*, *38*(1), 1–20. doi: http://doi.org/10.1287/mksc.2018.1123

Topic Modeling in Python with Gensim. (2018, March). Retrieved 2020-03-18, from https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/ (Library Catalog: www.machinelearningplus.com Section:

NLP)

Trenz, M., & Berger, B. (2013). Analyzing Online Customer Reviews-An Interdisciplinary Literature Review And Research Agenda. In *ECIS* (p. 83).

Wang, Y., Wang, J., & Yao, T. (2019, June). What makes a helpful online review? A meta-analysis of review characteristics. *Electronic Commerce Research*, *19*(2), 257–284. Retrieved 2019-10-30, from https://doi.org/10.1007/s10660-018-9310-2 doi: 10.1007/s10660-018-9310-2

Witten, I., Frank, E., Hall, M., & Pal, C. (2016). "Ensemble Learning" Chapter 12. In *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed., pp. 351–371). Morgan Kaufmann. Retrieved 2020-04-23, from https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5