5-2018

Classifying textual fast food restaurant reviews quantitatively using text mining and supervised machine learning algorithms

Lindsey Wright

UNDERGRADUATE HONORS THESIS AVAILABILITY AGREEMENT and NON-EXCLUSIVE DISTRIBUTION LICENSE

ETSU
East Tennessee State University

Please select one of the following honors thesis availability options:

_✓_ No restriction on availability

____ 1 year embargo

____ 2 year embargo

I agree to the terms of this Non-Exclusive Distribution License and Availability Agreement:

Lindsey Wright	Lindsey Wright	5/1/18
Honors Thesis Author Signature	Print Name	Date
Michele Joyner	Michele Joyner	5/2/18
Honors Thesis Director	Print Name	Date

rev. 2013-Dec-13

Classifying textual fast food restaurant reviews quantitatively using text mining and supervised machine learning algorithms

---

An Honors thesis presented to the faculty of the
Department of Mathematics and Statistics
East Tennessee State University

In partial fulfillment of the requirements for the University Honor Scholar Program for a Bachelor of Science in Mathematics

---

by

Lindsey Brooke Wright

April 2018

Lindsey Wright

Lindsey B. Wright

Michele Joyner

Dr. Michele Joyner, Mentor

Jeff Knisley

Dr. Jeff Knisley, Faculty Reader

Dr. Chris Wallace, Faculty Reader

Classifying textual fast food restaurant reviews quantitatively using text mining and supervised machine learning algorithms

An Honors thesis presented to the faculty of the

Department of Mathematics and Statistics

East Tennessee State University

In partial fulfillment of the requirements for the University Honor Scholar Program for a Bachelor of Science in Mathematics

by

Lindsey Brooke Wright

April 2018

Lindsey B. Wright

Dr. Michele Joyner, Mentor

Dr. Jeff Knisley, Faculty Reader

Dr. Chris Wallace, Faculty Reader

Abstract

Companies continually seek to improve their business model through feedback and customer satisfaction surveys. Social media provides additional opportunities for this advanced exploration into the mind of the customer. By extracting customer feedback from social media platforms, companies may increase the sample size of their feedback and remove bias often found in questionnaires, resulting in better informed decision making. However, simply using personnel to analyze the thousands of relative social media content is financially expensive and time consuming. Thus, our study aims to establish a method to extract business intelligence from social media content by structuralizing opinionated textual data using text mining and classifying these reviews by the degree of customer satisfaction. By quantifying textual reviews, companies may perform statistical analysis to extract insight from the data as well as effectively address concerns. Specifically, we analyzed a subset of 56,000 Yelp reviews on fast food restaurants and attempt to predict a quantitative value reflecting the overall opinion of each review. We compare the use of two different predictive modeling techniques, bagged Decision Trees and Random Forest Classifiers. In order to simplify the problem, we train our model to accurately classify strongly negative and strongly positive reviews (1 and 5 stars) reviews. In addition, we identify drivers behind strongly positive or negative reviews allowing businesses to understand their strengths and weaknesses. This method provides companies an efficient and cost-effective method to process and understand customer satisfaction as it is discussed on social media.

Contents

1 Introduction **3**

1.1 Motivation . . . 3

1.2 Text Mining . . . 5

1.3 Predictive Modeling . . . 7

2 Data Overview **8**

3 Preparing the Data **10**

3.1 Pre-Processing . . . 10

3.2 Frequency Term Matrix . . . 11

3.3 Manual Clustering . . . 13

3.4 Sentiment Analysis . . . 22

4 Methods **24**

4.1 Decision Tree Classifier . . . 24

4.2 Random Forest Classifier . . . 27

5 Results and Discussion **28**

6 Conclusion **33**

7 Future Work **34**

1 Introduction

1.1 Motivation

Social media content establishes a unique opportunity for companies. Currently, companies tend to place a strong emphasis on customer feedback in order to improve their business model. In fact, Semio Corporation maintained,"Increasingly, companies are looking for strategic solutions that will help them leverage that information in order to be more nimble to market demands" [21]. For this reason, consumers are often requested to fill out a survey advertised on the bottom of a receipt or within an app or email. This method allows companies to gather data on their customers' demographics and satisfaction as well as the efficiency of their business. However, this feedback can be problematic. First, the questions posed to the customer may result in bias. In addition, the number of responses is often too small of a sample size to be statistically significant. This method requires companies to be dependent upon the proactive effort of their customers to provide feedback. While gathering customer feedback through surveys is financially and statistically difficult, alternative methods of customer satisfaction data collection already exist in social media. Social media produces overwhelming amounts of text everyday with roughly 500 million Tweets (6,000 per second) [11], 95 million Instagram photos [18], and 4.75 billion pieces of content shared on Facebook per day [4]. Often these posts discuss satisfaction with a certain product or an experience at a business, mimicking reviews.

While review-like data is readily available and easily accessible on social media platforms, combing through thousands of online comments is expensive in both time and money. Claude Vogel, founder and CEO of Semio Corporation explains, "The problem today is that there is too much information overload. Increasingly, companies are looking for a solution that will help them leverage their legacy data" [21]. Large companies may be discussed in thousands of online posts per day, and dedicating personnel time to interpret, respond, or collect the information is logistically inefficient. Furthermore, technology is capable of storing and processing text but is unable to interpret the meaning or opinion. Thus, businesses need an efficient and cost-effective method to extract the readily available insight from social media.

This study aims to develop a method to extract business intellect from review-like posts on social media. We employ Yelp! reviews to quantify textual data on a discrete scale from one to five indicating the degree of customer satisfaction [24]. By assigning a numerical value to convey negative to positive sentiment, businesses may track trends in customer satisfaction over time and understand the general populace's opinion of the company. In addition, our method looks to extract key features within reviews that lead to highly satisfied/dissatisfied reviews, allowing businesses to easily visualize their strengths and weaknesses. Initially, we text mine the Yelp! reviews and construct a matrix representative of the data. This matrix is then utilized to build and test a predictive model that assigns each review a star rating. By implementing supervised machine learning algorithms, we hope to develop an accurate model that reflects the opinion of text quantitatively. We have two purposes of this study. First, we want to accurately quantify textual reviews. This model may then be used to predict star ratings of textual data from other sources such as Facebook comments, Tweets, or Instagram posts. If a less than satisfactory review is identified through the model, then the post may be flagged to company personnel that may handle the situation as needed. Second, we want to understand what causes the customer to be satisfied or dissatisfied. Thus, we will analyze and interpret the model and results through these two different lenses.

1.2 Text Mining

Since language is highly irregular in spelling, length, and meaning, textual data is unstructured thus unable to be interpreted by technology; yet the majority of available data is textual. Therefore, textual data must first be structured before it can be analyzed. Text mining is an important technique for extracting information and key concepts from collections of textual data. Text mining bridges the gap from large unstructured textual data to structured data that allows for understanding of relationships and themes within the data [7]. After data has obtained a structure, we may implement predictive modeling, classification algorithms, clustering analysis, and other techniques to extract insight.

Text mining is used in many different fields to accomplish a variety of purposes. For example, Zi Ning of East China Normal University and his team constructed an an intelligent interface named "OncoViz" that text mines health literature available online in real time and customizes the resulting information for cancer patients. The team constructed a data structure containing a collection of relevant terms and calculated the Term Frequency-Inverse Document Frequency to determine the importance of each term in the documents. In addition, Ning and team considered word associations to relate drugs to their side effects. Finally, the team developed a user-friendly visualization that allows users to easily navigate and understand the risks of cancer-related drugs and alternative medications. Text mining plays an important role to the "OncoViz" interface by allowing patients easy access to the most recent information on cancer-related drugs and their side effects [13].

Text mining has also been used to increase efficiency in business settings by simplifying email threads. University of British Columbia professor Giuseppe Carenini and associates utilized text mining and clue words to summarize email conversations. They first constructed a fragment quotation graph (a directed graph) to represent conversation threads. Then Carenini and associates developed Algorithm ClueWordSummarizer which first text mines the email thread then assigns a quantitative value to each term and sentence. The sentence assigned the highest score is returned as the email conversation summary. This allows users to save time by browsing summaries rather than reading the entire thread in search of vital information [2].

Furthermore, companies may implement text mining techniques to utilize the overwhelming majority of internet content to their advantage. In the past, business intelligence has been gathered from structured data only. However, the overwhelming majority of data is unstructured text data. Thus, Byung-Kwon Park of Dong-A University and Il-Yeol Song of Drexel University extract business intelligence from both unstructured and structured data allowing businesses to study a broader, more inclusive range of information. Parks emphasizes, "Through analyzing the reports on market trends, news articles, and web pages in Internet, business people can obtain important business information such as new competitors or competitive products coming out in market or consumer demand patterns changing" [15]. Combining both structured data and unstructured textual data allows companies to see a greater overall picture of their customers desires and satisfaction and equips the company to advance with current market demands.

Similarly, our study looks to improve business intelligence through text mining. By combining text mining with predictive modeling, we may understand the opinions within reviews of fast food restaurants. Companies continuously leverage structured data to make informed decisions, but multitudes of textual data exist on social media that often remain untouched. In addition, assigning a quantifier to represent textual reviews allows for greater statistical analysis. We have sound advanced methods for analyzing numerical data whereas processes to analyze text are still being developed. Thus, text mining is the first step in achieving our goal of gaining business intelligence from social media.

1.3 Predictive Modeling

Predictive modeling is a method to assign a probability of a given outcome using mathematical modeling and algorithms [9]. Predictive modeling allows for the consideration of a combination of factors and discovery of underlying trends and relationships. Applications of predictive modeling include real-time face recognition [23], the discovery of higher order and nonlinear genome-wide associations [5], predicting consumer financial behavior [12], and a variety of others. An analyzation of texual reviews suggests a need for modeling. We cannot classify reviews simply by the presence or absence of words. Language is very complex which uses combinations of words and phrases to communicate an idea. In addition, sarcasm makes it increasingly difficult to detect the tone of the author. Thus, we need a process that considers word/phrase combinations within reviews to extract the communicated satisfaction level.

Specifically, supervised machine learning algorithms are of special interest to this study. Supervised machine learning uses data to train the predictive model on how to make predictions accurately by already knowing the correct outcomes [10]. If we can use a dataset that has both a textual review and a numerical value that reflects the level of customer satisfaction, we can test our predictive model for accuracy and precision. Specifically, our study will test the ability of decision trees and random forest classifiers to accurately rate reviews by level of customer satisfaction.

2 Data Overview

The data used in this project is an open Yelp! dataset. Yelp! is a social media platform dedicated to reviewing and recommending restaurants, bars, hotels, shopping, nightlife, etc. Currently, Yelp has over 135 million business reviews [24]. Each review contains both a textual review and a discrete quantitative star rating from one to five. In addition, each entry has a series of attributes associated with the user, the business, and the review itself. These attributes, displayed in Table 1, allow the user to identify important information about the establishment, the services offered, or the trustworthiness of the reviewer. Specifically, our data set contains 56,414 reviews of fast food restaurants represented in Arizona, Illinois, North Carolina, Nevada, Pennsylvania, South Carolina, and Wisconsin as seen in Figure 1.

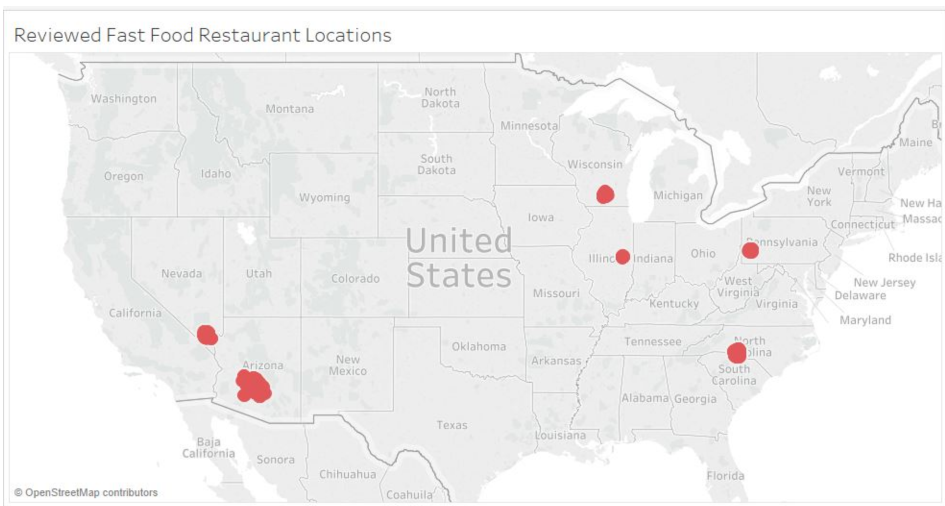

Figure 1: Locations of Fast Food Restaurants Reviewed

Figure 2 displays the distribution of the star ratings within our data set. We observe an emphasis on polarizing ratings (1, 4, and 5 stars) and fewer apathetic reviews (2 or 3 stars). Since predictive modeling is increasingly more difficult

Table 1: Data Set Attributes

Review Attributes	**Business Attributes**	**User Attributes**
Business ID	User ID	Business ID
Address	Name	Date
Category of Business	Date Joined Yelp	Review ID
Open	Number of Reviews	Stars
Category	Average Star Rating	Text
City	Type of Account	Type
Review Count	Fan Count	User ID
Name	Cool Votes	Cool Votes
Neighborhood	Funny Votes	Funny Votes
Longitude & Latitude	Useful Votes	Useful Votes
State	Cool Compliments	
Stars	Cute Compliments	
Wi-Fi	Funny Compliments	
Take-Out	Hot Compliments	
Drive-Thru	List Compliments	
Takes Reservations	Note Compliments	
Delivery	Photo Compliments	
Noise Level	More Compliments	
BYOB	Profile Compliments	
Corkage	Writer Compliments	
Dogs Allowed	Plain Compliments	
Caters	Elite Status	
Recommended Meal Time		
Ambience		
Type of Parking		
Good for Kids or Groups		
Dogs Allowed		
Coat Check		
Smoking		
Wheelchair Accessible		
Has TV		
Outdoor Seating		
Attire		
Alcohol		
Accepts Credit Cards		
Price Range		
Open 24 Hours		

with more than two classification classes, we study only strongly positive and strongly negative reviews (5 and 1 stars, respectively). Due to computational limitations, we randomly selected, without replacement, 20,000 reviews from the total subset of strictly 1 or 5 star reviews. Figure 3 depicts the distribution of stars from the selected subset. We observe a similar trend in review frequency as reflected in the entire data set. Since this problem is computationally expensive and increasingly more difficult as classification classes are added, simplifying to two highly polarized star ratings allows us to initially build an effective method that may later be expanded.

Figure 2: Distribution of Star Ratings in Entire Data Set

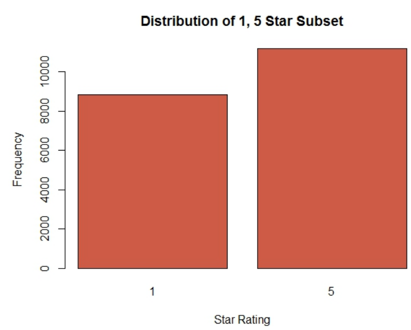

Figure 3: Distribution of Star Ratings in 1, 5 Subset

3 Preparing the Data

3.1 Pre-Processing

To begin implementing text mining techniques, we first pre-process the reviews. Pre-processing is a series of steps that remove unnecessary language such as punctuation, repeating letters, and plural forms. As our first text mining step, we create a corpus data structure containing all the reviews. A corpus is a data structure that contains each individual text document. Next, we extract each term from the corpus through tokenization. Tokenization is simply the process of recognizing each term as a "token" and extracting each term from the overall document. This way each term is seen as an individual rather than a part of the whole [6]. Next, we pre-process the corpus with the following method:

1. Remove white space, punctuation, and special characters.
2. Covert all letters to lower case.
3. Remove unnecessary repeating letters.
4. Remove stop words.
5. Stem all words to the root word by removing suffixes, demonstrated in Figure 4.

This process removes unnecessary items from textual data [6]. Punctuation and many words in the English language are not informative to the overall value of a sentence. Thus, by removing these items we reduce the size of our data and unnecessary repetition.

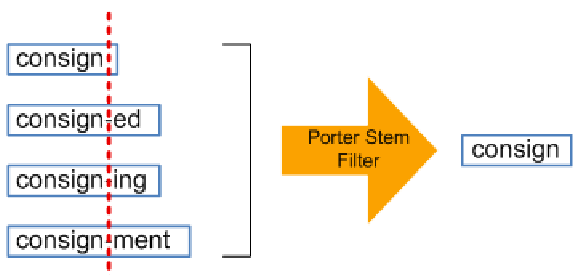

Figure 4: Example of Stemming [25]

3.2 Frequency Term Matrix

After pre-processing the Yelp! reviews, we construct a frequency term matrix. This allows the data to be structured for utilization in predictive modeling and further analysis. Within the frequency term matrix, rows represent each review while columns represent each tokenized pre-processed term. We summed the occurrence of term j within review i and place the frequency within cell i, j. For example, suppose we have the following reviews:

1. "The food was bad."
2. "The food was good."
3. "The food was very, very good."

First, all punctuation and white space are removed. Next, the stop words "was" and "the" are removed. Finally, each term is represented as a column and the frequency of each term is placed in the proper cell.

Table 2: Sample Frequency Term Matrix

	bad	food	good	very
Review 1	1	1	0	0
Review 2	0	1	1	0
Review 3	0	1	1	2

By assigning a column to every individual term present within thousands of reviews, the frequency term matrix becomes abundantly large. Thus, we reduce the matrix without losing valuable information. If the sum of column i is 1, we delete column i. Essentially, we throw out all words that only occur once across the entire collection of reviews. This allows for greater speed and less required storage space for predictive modeling. However, we must be aware that decreasing the size of the dataset, decreases the variance and increases bias. In the future, we would like to use a sparse representation of the data to minimize computational time and retain all the data.

3.3 Manual Clustering

In order to improve predictive modeling capabilities and extract relevant business intelligence, we manually cluster terms within the frequency term matrix. Our hope is to essentially provide a thesaurus to group similar words of importance when discussing fast food restaurants. Specifically, we desire to extract topical subjects informative of the customer experience. This allows us to target specific aspects of running an efficient fast food restaurant by combining like ideas that would be separated within the frequency term matrix due to differing terminology. For this reason, we manually clustered the data using common customer experience themes when visiting a fast food restaurant. We considered cleanliness of restaurant, speed of service, employee appearance, attitude and work ethic of employees, food taste, food temperature, order accuracy, occurrence of problem, problem resolution, overall satisfaction, and likelihood to return. In order to cluster these themes, we determined both a negative and positive term that embodies each theme then assembled a list of synonyms for each term. Table 3 displays the chosen terms for each customer experience theme. In addition to the terms in Table 3, we considered other words that do not fit in any one of these categories such as "forever," "staff," and "crowded" for a total of thirty-six clustered terms. To cluster the customer experience themes, we summed the occurrence of synonyms for term i in each review j and placed the total value at position j, i. The column heading for each negative and positive representative term is denoted as i. We refer to this new matrix as the Filtered Matrix - containing both the frequency term matrix and the customer experience filters.

These filters provide greater insight as to what the reviews are trying to communicate. For example, if there is an abundance of words from the negative "speed of service" filter, we can conclude that the customer was most likely

Table 3: Customer Experience Themes

Category	**Positive Term**	**Negative Term**
Cleanliness of Restaurant	clean	dirty trash unkempt
Fast Service	fast	slow
Attentive & Courteous Employees	friendly attentive courteous sincere listen	unfriendly inattentive disrespectful neglect careless
Taste of Food	tasty	distasteful
Order Accuracy	accurate understand	incorrect misunderstanding wrong
Experience a Problem		problem
Problem Resolution	resolve respond	unresolved
Overall Satisfaction	satisfied	unsatisfied
Likelihood to Return	return	leave

displeased with their visit due to the slow service time. By combining synonyms of customer satisfaction, we hope to provide insight into drivers of customer satisfaction.

Upon completing the Filtered Matrix, we construct word clouds to visualize the overall trends of the data. As the frequency of a term increases within the subset of reviews, the font size also increases. Figure 5 represents the most commonly used words within the basic frequency term matrix. The most common terms include “food,” “order,” “place,” “get,” “service,” and “time.” Yet, these terms are not beneficial in analyzing sentiment towards a particular fast-food restaurant; we are unable to determine if the service was slow or perhaps the customer disliked their food. Moreover, as the font size decreases, we observe specific food terms such as “chicken,” “fries,” and “burger” begin to emerge. Although the over-arching principle of a restaurant is food, a specific food without the sentiment attached to it is useless for understanding customer satisfaction. Thus, the word cloud for the frequency term matrix is not beneficial in analyzing the opinion of a review since we are unable to observe any sentiment. However, when we add in the customer experience filters seen in Figure 6, themes begin to emerge that reflect emotion or customer experience. We observe many reviews discuss friendliness, cleanliness, order accuracy, and clear communication. Based on these results, businesses may make informed decisions to place emphasis on maintaining a clean facility or hiring friendly employees.

Figure 5: Frequency Word Cloud

Figure 6: Filtered Word Cloud

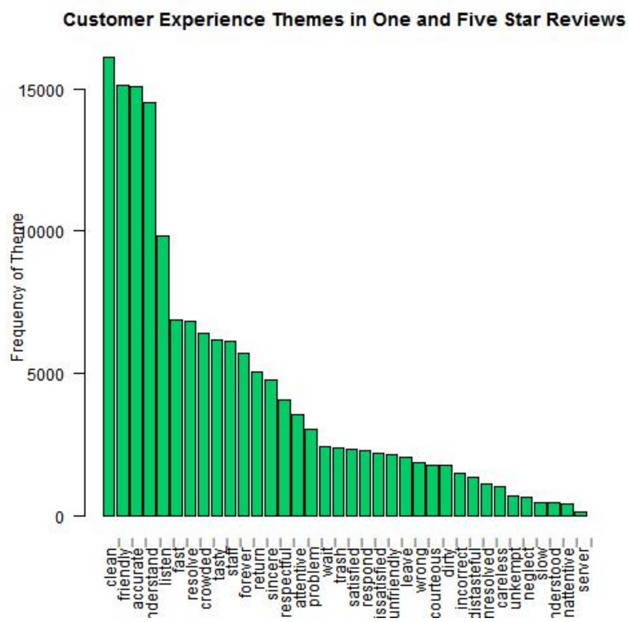

Figure 7: Most Frequent Customer Experience Themes in 1 Star and 5 Star Reviews

Figure 7 demonstrates the distribution of the frequency of each customer experience theme. We can see the most discussed customer experience themes are "clean," "friendly," "accurate," and "understand," and "listen". This suggests the cleanliness (or lack thereof) of a fast-food restaurant often prompts customers to write a review on Yelp!. In addition, many of the common themes are related to the customer's interaction with the employees such as friendliness, understanding, listening, and even order accuracy. Each of these themes are dependent upon solid communication between employee and customer. Therefore, hiring competent, well-mannered employees with strong communication skills is most likely favorable for high customer satisfaction at fast food restaurants. Interestingly, the taste of food is rarely discussed. One might assume better food results in better reviews, but it appears as if the treatment of the customer prompts a written review rather than the quality of the food.

Upon examination of Figure 7, we observe the most frequently discussed customer experience themes: clean and friendly. The following figures explore the frequency of these themes in one and five star reviews. We hope to see how frequent these themes occur within reviews and whether these ideas are more prevalent in one or five star reviews. Figure 8 indicates the amount of times cleanliness is discussed within a review, the more likely the review is to be a one star review. However, Figure 9 warns that there are few cases in which cleanliness is discussed more than three to four times. Figure 10 suggests that as the customer discusses friendliness, the more pleased they are with their visit, as expected. However, we are unable to determine the star rating of a review based solely on the presence/absence or frequency of these themes within a review. We may simply conclude that friendliness and cleanliness are highly valued by the customer.

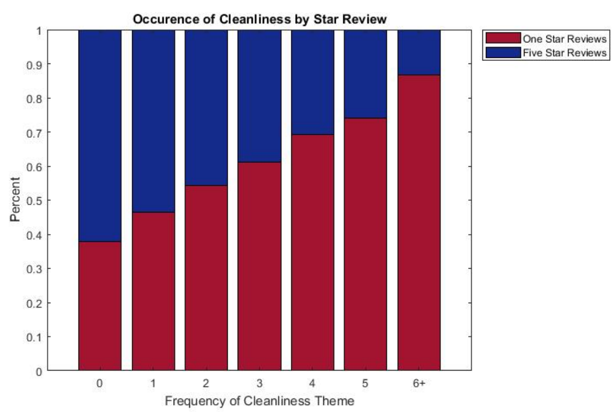

Figure 8: Percent of one and five star reviews that contain n instances of cleanliness

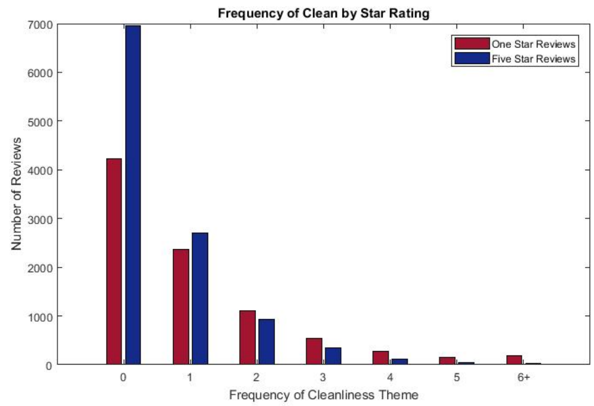

Figure 9: Count of one and five star reviews that contain n instances of cleanliness

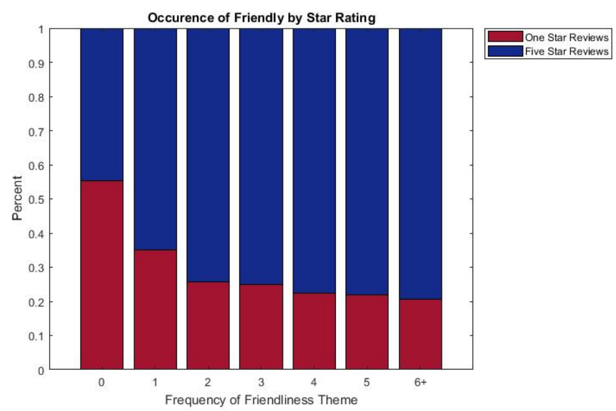

Figure 10: Percent of one and five star reviews that contain n instances of friendliness

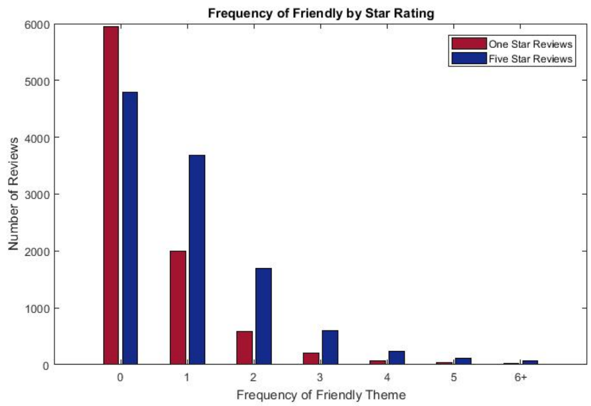

Figure 11: Count of one and five star reviews that contain n instances of friendliness

After observing the entire data set, we constructed word clouds to study the different terms and themes present in one star and five star reviews as seen in Figure 12 and Figure 13. We observe in both one and five stars order accuracy and cleanliness is highly mentioned. In one star reviews, we specifically observe discussions of crowded, time, worst, unfriendly, dissatisfied, forever, service, and attentiveness. On the other hand, five star reviews comment on quickness, understanding, friend, good, service, nice, sincerity, hot, and love. Yet, there is still overlap in terms. Both discuss crowded, service, time. Once again this indicates the inability to predict the star rating based on the presence of terms alone. We need a more intuitive method to distinguish between one and five stars.

Figure 12: Filtered Word Cloud 1 Star

Figure 13: Filtered Word Cloud 5 Stars

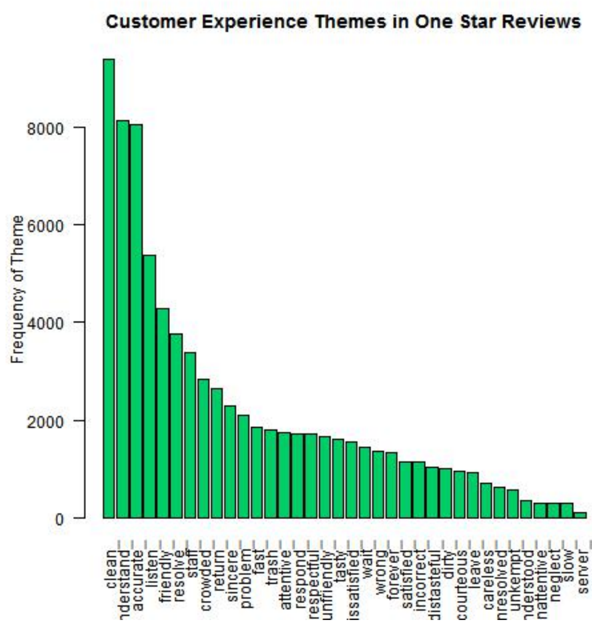

Figure 14: Most Frequent Customer Experience Themes in 1 Star Reviews

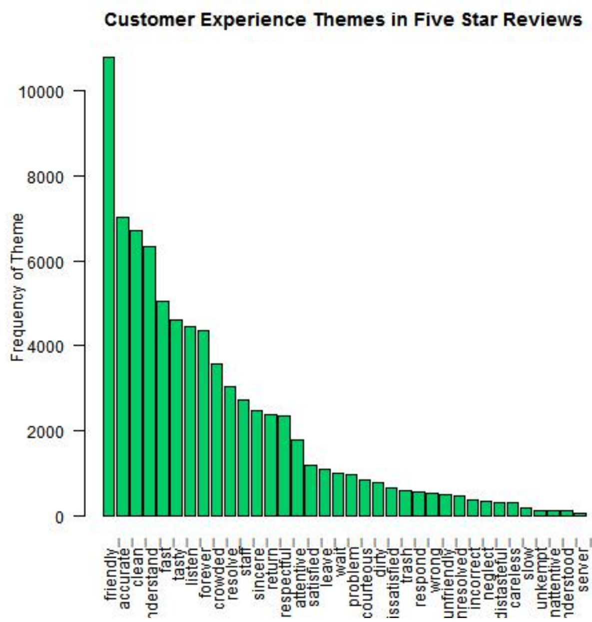

Figure 15: Most Frequent Customer Experience Themes in 5 Star Reviews

Figure 14 displays the frequency of the customer experience filters for one star reviews. We observe cleanliness, accuracy, understand, and order to be the most common themes in these reviews. The most common filters are for five star reviews are seen in Figure 15. We see an emphasis on friendliness, accuracy, cleanliness, and understanding. Both 1 Star and 5 Star reviews discuss cleanliness and order accuracy. Under our first purpose of prediction capability, this indicates the presence of a word is not sufficient enough to predict a star review, reiterating our need for predictive modeling. However, in terms of obtaining business intelligence, we note the cleanliness of the facility is very important to customers as well as order accuracy. Once again the taste of the food is not highly discussed, so businesses may place more of an emphasis in training employees to keep a clean restaurant rather than pouring money into altering/perfecting an already satisfactory recipe.

3.4 Sentiment Analysis

Since our data consists of opinionated reviews, we need to understand the emotion attached to words. For this reason, we utilize sentiment analysis to capture the positive or negative emotion conveyed in each review. Anuj Sharma of Indian Institute of Management claims, "Sentiment analysis is performed to extract opinion and subjectivity knowledge from user generated text content." Sentiment analysis differs from text mining because it seeks to extract and classify opinion rather than topical information [22].

Upon observation of specific reviews, it becomes clear as to why technology is unable to interpret language: sarcasm. For example, one Yelp! user review states, "This McDonald's is so bad it's amazing." This review contains what a fluent English speaker understands to be a very positive term "amazing" as well as a negative term "bad." While it is clear to the English-speaking mind that a sentence constructed in this manner communicates great disappointment and sarcasm, a computer is unable to understand such sarcasm and sentence structure. Thus, we need to design algorithms that can understand patterns and learn from language. To extract the overall meaning of a sentence, we cannot look for the presence or absence of a word, but rather we must understand the relationships and combinations of terms. For this reason, we employ sentiment analysis to assign numerical values to terms/phrases to reflect the emotion carried within reviews.

To each term we assign an affinity score, a numerical value that reflects the level of positive or negative emotion of the term. Our study specifically obtained affinity scores from the "sentimentr" package in R that uses Jockers 2017 dictionary of polarized words [8]. The algorithm assigns a sentiment score of $\{-1, 1\}$ to each term to reflect negative or positive emotion, respectively [20]. The algorithm considers a cluster of words surrounding each polarized term including four words preceding and two words following the polarized term. Words surrounding the polarized term are marked as either neutral terms, negators, amplifiers, or deamplifiers. Amplifiers increase the polarity by conveying stronger emotion while deamplifiers reflect a weaker expression of emotion thus decrease the polarity. Negators flip the sign of the polarity if the number of negators is odd [19]. By creating a cluster around each polarized word, the algorithm is able to account for sentiment conveyed in phrases. We then sum the sentiment of each polarized cluster within each review and add the score as a column on our matrix. Our final Complete Matrix consists of the frequency of each individual term, the customer experience themes (filters), and the net sentiment.

Figure 16 and Table 4 display net sentiment by star rating. We observe on average a higher net sentiment within the five stars than the one stars, as expected. However, there is still overlap within the net sentiment of one and five stars allowing for no clear separation based on net sentiment alone.

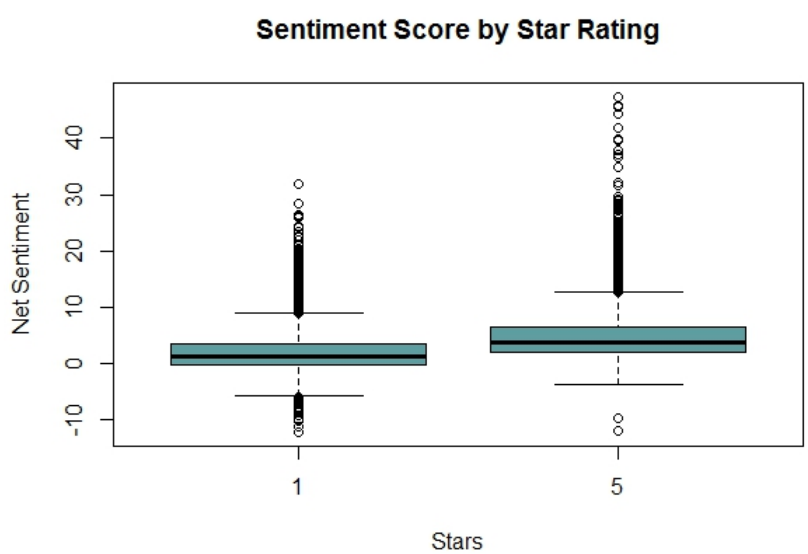

Figure 16: Net Sentiment of Reviews by Star Rating

Table 4: Net Sentiment Five Number Summary

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
One Star	-12.250	-0.350	1.250	1.804	3.350	31.950
Five Star	-11.950	2.050	3.800	4.717	6.350	47.400

4 Methods

Now that we have constructed our Complete Matrix, we convert to Python to build our predictive models [16]. We consider two different types of supervised machine learning algorithms: Decision trees and random forest classifiers. Decision trees are a weaker predictive model, but they are easily interpreted. Thus, we utilize decision trees for insight on the dataset. On the other hand, we use random forest classifiers as a more robust algorithm to build a predictive model with higher accuracy, but we sacrifice interpretability.

4.1 Decision Tree Classifier

For our purpose of understanding the common themes and relationships within reviews, we employ Decision Tree Classifiers. While decision trees are a weaker machine learning algorithm, they allow for clear visualization of the classification process. Thus, the interpretability of the decision trees is beneficial to businesses by providing insight into customers' desired experience at fast food restaurants. A decision tree is a predictive model that constructs a series of true/false statements with the purpose of classifying an object. To build an accurate model, we first randomly partition the data into a training set and testing set. We arbitrarily chose two-thirds of the data to be used as the training set. Next, the algorithm selects one factor that best divides the entire data set into an ideal 50/50 split. This factor is termed the root node or best predictor. Since this factor results in a roughly 50/50 split, there is maximum uncertainty within the root node. The goal is to split the data in such a way that the following intermediate nodes have the least amount of impurity possible [17]. We choose to utilize entropy as our measure of impurity in this study. Entropy is defined as a measure for randomness, impurity, or uncertainty, calculated with the following formula

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where p_i is the probability of class i, X is a discrete random variable, and n is the number of classes [9]. As entropy approaches one, we obtain maximum uncertainty in a binary split, yet we approach absolute purity as entropy goes to zero [17]. We desire to minimize entropy, so we may maximize confidence in our model's ability to classify accurately. After the initial division within the root node, the algorithm searches for the next best factor that results in the highest information gain. In a binary split, information gain is defined as

$$IG\Big(D_p, x_i\Big) = H\Big(D_p\Big) - \frac{N_{left}}{N_p} H\Big(D_{left}\Big) - \frac{N_{right}}{N_p} H\Big(D_{right}\Big)$$

where x is the factor on which the algorithm performs the split, N_p is the number of samples in the parent node, H is the entropy function, D_p is the subset of training samples in the parent node, and D_{left} and D_{right} are the subset of training samples in the left and right child nodes after the split, respectively [17]. Essentially, the algorithm searches for the factor that maximizes information gain by minimizing entropy in the children nodes. This algorithm continues splitting into intermediate nodes until a predicted category, also known as a leaf, is obtained [9].

While decision trees are helpful in visualizing important factors for prediction, decision trees have high variance. In order to combat this problem, we bootstrap the decision tree. The process of bootstrapping involves building multiple decision trees and allowing each tree to classify the object [14]. This allows us to account for the variability in the tree by using multiple different training and testing sets. Since decision trees are dependent upon the training set used, we randomly select a training set from the original data for each individual tree constructed. Although there exists variability among the training and testing sets used for each bootstrapped decision tree, this process still results in highly correlated trees since each tree is built from a small subset of strong predictors [14]. For the purpose of our study, we bootstrap 100 decision trees to study relationships between the selected factors.

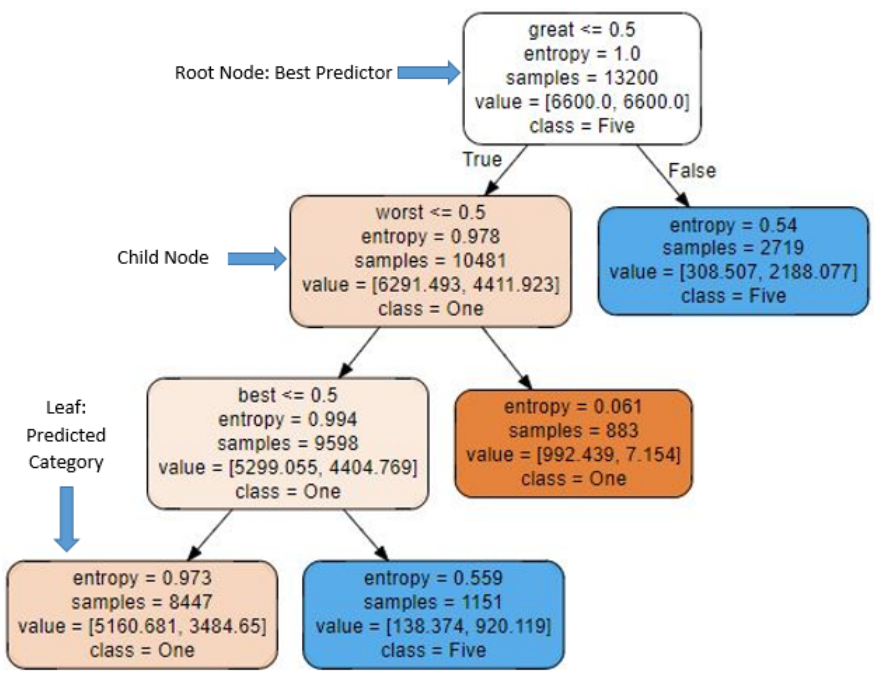

Figure 17: Decision Tree Example

Figure 17 is an example of a Decision Tree with four leaves. We begin with the best predictor, in this case "great." If the root term is present within the review, we proceed down the tree to the right. If the root term is not present, we proceed to the left and continue likewise through the children (intermediate) nodes until a leaf is reached.

It is important to note that it is possible to achieve a perfect prediction score through decision trees as a result of over-fitting. Essentially, the model has learned both the relevant relationships and the noise within the training data. This is problematic, because the model is faulty when other data is inserted into the decision tree. Thus, the best predictive model is obtained when the testing and training score are closest together [9]. In order to avoid over-fitting, we must prune the trees. Essentially, we limit the number of leaves on the trees so the algorithm does not grow to perfectly fit the data. We desire to use a decision tree in which we prune the tree to minimize the ratio defined as follows

$$\text{ratio} = 1 - \frac{\text{training score}}{\text{testing score}}.$$

Therefore, we study the performance and ratio of the decision tree by varying the number of leaves. We define the best predictive model by selecting the optimal number of leaves where the prediction accuracy is maximized and the ratio is minimized [9].

4.2 Random Forest Classifier

In order to overcome the high variance and correlation found in bootstrapped decision trees, we utilize Random Forest Classifiers to achieve our second purpose of accurately classifying reviews. Similar to decision trees, we first randomly select a training and testing set. A Random Forest is essentially multiple decision trees with one initial difference. Rather than selecting the best divisor from the entire list of factors, the algorithm selects a random subset of the factors. The best predictor is then chosen from the random subset of factors [14, 9]. The algorithm chooses a random subset of factors for each split [1]. Thus, a random forest consists of a collection of less correlated decision trees. In order to classify an object, each tree within the forest casts a vote. The object is classified by the majority vote. The idea is that many weak predictors perform better than one strong predictive tree [17, 9]. Based on previous studies, we constructed a random forest with an optimal number of 128 trees [14]. In order to avoid over-fitting within the random forest, we analyzed the trees performance by varying the number of leaves from two to sixteen leaves.

5 Results and Discussion

To fulfill our first purpose of interpreting relationships within the reviews for business intelligence, we observe the decision tree of the complete matrix with twelve leaves displayed in Figure 18. This tree has a training score of 77.82%, testing score of 77.00%, and a ratio of 0.34. Our goal for the decision tree is interpretablity, so we desire a tree with a small number of leaves. Thus, we analyze the twelve leaf tree which is later deemed optimal by the random forest results. We see the greatest divider of the data is the net sentiment. If the net sentiment is less than or equal to 0.475, then we proceed down the left side of the tree and classify the review as One Star, regardless of the presence of the term "worst." We note the only factors considered in the decision tree are net sentiment, worst, great, order, delicious, best, friend, and love. Looking at these terms, we are able to extract minimal business intelligence from the reviews. For this reason, Figure 19 examines a decision tree built only with the customer experience themes.

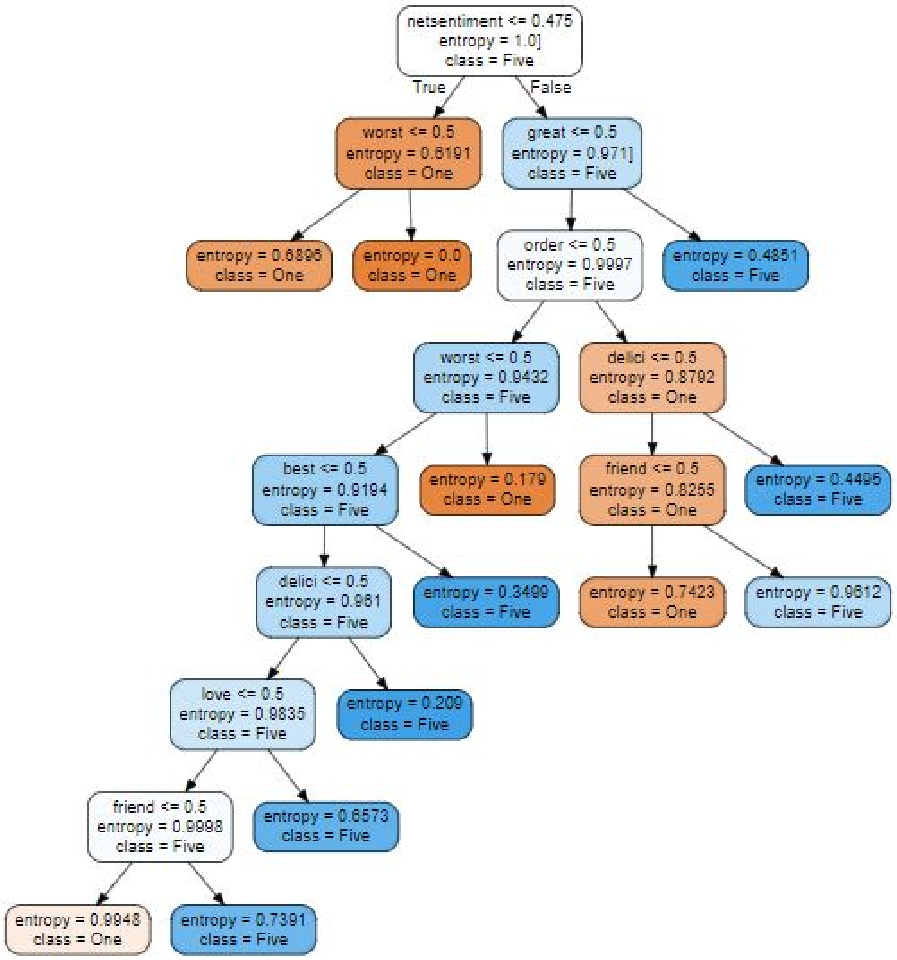

Figure 18: Decision Tree with 12 Leaves

The customer experience theme decision tree has a training score of 67.77%, a testing score of 67.53%, and a ratio of 0.37. Although the accuracy of the themes decision tree is lower than the complete matrix decision tree, we are able to determine underlying themes within the data set that is helpful for the business. In this figure we note two prevalent effective business models. If there is no mention of the root node “friendly,” then five star reviews mention “tasty” and “fast.” This implies that customers leave satisfied if they receive tasty food quickly. This business model suggests being treated kindly by employees is not usually factored into the satisfaction level of the customer. On the other side of the decision tree, we note the customer discusses friendliness, responsiveness, and order accuracy. This implies customers are also highly satisfied when the restaurant’s staff is competent and kind.

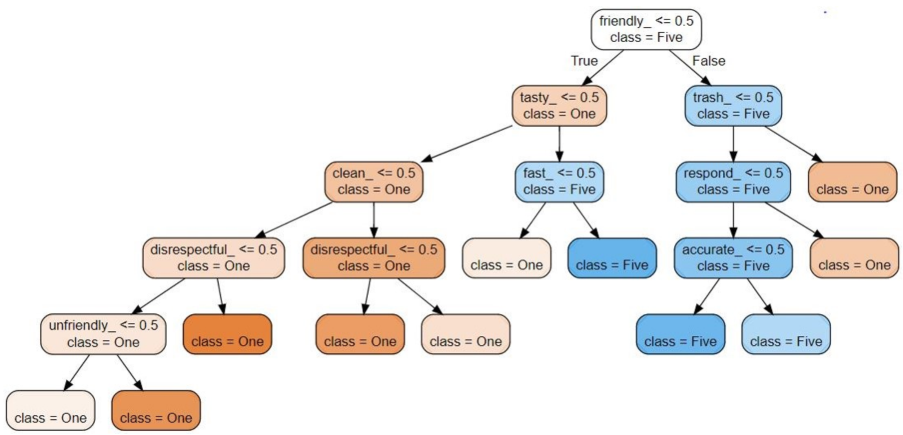

Figure 19: Customer Experience Themes Decision Tree with 11 Leaves

Figure 20 displays the performance of 100 bagged averaged decision trees against a random forest of 128 trees while varying the number of leaves. We see that random forests perform significantly better than decision trees. At only two leaves, the random forest already performs 15% better than the decision trees. For this reason, we conclude random forest should be used over decision trees to accurately classify between one and five star reviews.

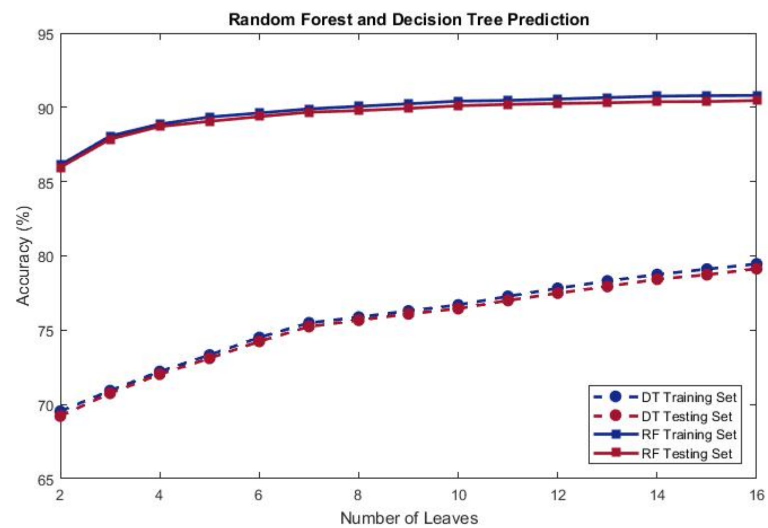

Figure 20: Comparison of Performance of Bootstrapped Decision Trees and Random Forests

Figure 21 shows a closer examination into the performance of the random tree algorithm with our data set. We seek to determine the optimal number of leaves that does not overfit the data but has a high prediction rate. We first note that the algorithm predicts with about 88% accuracy at four nodes with a very small ratio. The prediction accuracy begins to stabilize at twelve leaves with about 90% accuracy. After twelve leaves, the testing set score only increases by 0.01 yet the ratio increases at a much higher rate, so we achieve our best prediction rate at twelve leaves without risking overfitting. It also benefits us to prune the number of leaves to save on computational power. Thus, we conclude with two recommendations. If the company has limited computing power, four leaves are suggested. However, if the company has the ability to compute a random forest of 128 trees, then twelve leaves leads to a higher prediction accuracy when studying textual reviews.

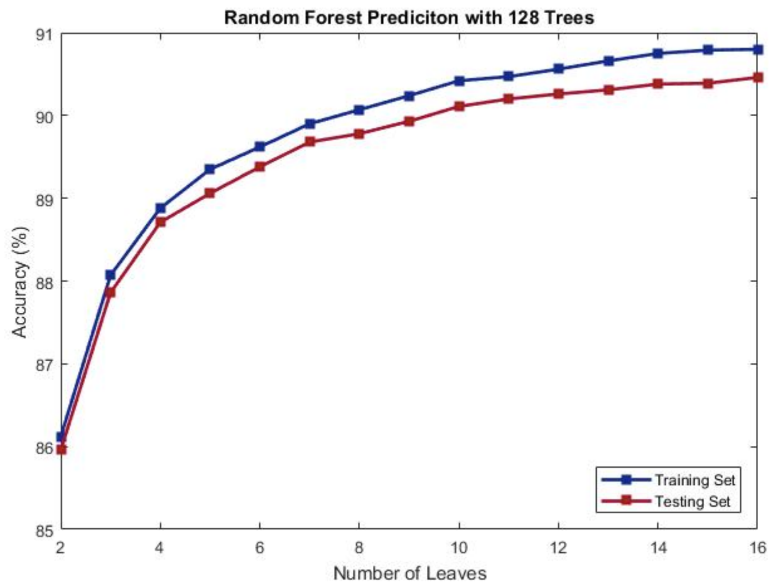

Figure 21: Random Forest Prediction with 128 Trees

Table 5 displays the prediction of the random forest algorithm with 128 trees with 12 leaves each with a training score of 90.53%, a testing score of 90.12%, and a ratio of 0.38. When analyzing the performance of the algorithm through the confusion table, we want to consider the precision and true positive rate of each star rating. Precision is defined as the number of true positive divided by the number of true and false positives [3]. In our definition, we will define a "positive" as a star rating. For example, the precision of five stars is found by dividing the number of accurately predicted five star reviews by the total number of predicted five star reviews. When the algorithm predicts a review to be a one star rating, it predicts correctly with a precision of 0.8820. Yet when a review is predicted to be a five star rating, the algorithm has a precision of 0.9200. Thus, the random forest classifier has a greater chance of classifying a five star review correctly. Next, the true positive rate communicates the number of correctly classified reviews for each star rating. The true positive rate is defined as the

Table 5: Confusion Table of Random Forest Prediction with 12 Leaves

		True Star Rating		
		One Star	Five Star	Total
Predicted Star Rating	One Star	$7,955$	$1,064$	$9,019$
	Five Star	878	$10,103$	$1,0981$
	Total	8833	$11,167$	$20,000$

number of true positives over the sum of true positives and false negatives [3]. Therefore, we have a true one star rate of 0.9006 and a true five star rate of 0.9047. We observe the random forest classifiers ability to distinguish between highly negative and highly positive reviews with relatively high accuracy and precision thus fulfilling our second purpose.

6 Conclusion

Through text mining and predictive modeling, we are able to uncover the fast food customer's satisfaction level is highly determined by the cleanliness of the restaurant as well as the kindness of the employees. Interestingly enough, the taste of the food is not highly discussed within fast food restaurant reviews. This may be attributed to the expectations a customer has when approaching a fast food restaurant. Usually, the quality and taste of a fast food restaurant is understood before entry, yet the environment the customer enters may vary. Therefore, fast food restaurants should emphasize a welcoming, clean environment when seeking to improve customer satisfaction levels. In addition, we recommend the use of random forest classifiers with 128 trees and 12 leaves with at least 20,000 initial one or five star reviews to build an algorithm to flag highly negative reviews on social media. To save computational power, we would also like to explore the use of a sparse representation of the data.

7 Future Work

Now that we have developed an effective model using only one and five star reviews, we would like to expand the study in an attempt to distinguish 5 star reviews from all others. We desire to study this difference because if a review is not a strong five, then the customer has walked away with some sort of dissatisfaction. We assume all dissatisfaction to be worth noting for the benefit of the company. Also, we would like to be able to predict each individual star rating from one to five, although more than two classifiers results in greater difficulty and usually lower accuracy. In addition, we may expand our study and test other predictive models to determine the model with the best predictive capability.

References

[1] Leo Brieman. Random forests. *Machine Learning*, 45:5–21, 2001.

[2] Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. Summarizing email conversations with clue words. *Association for Computing Machinery*, pages 91–100, 2007.

[3] Jesse Davis and Mark Goardrich. The relationship between precision-recall and roc curves. *ICML Proceedings of the 23rd international conference on machine learning*, pages 233–240, 2006.

[4] Facebook, Menlo Park, California. *Facebook Reports Fourth Quarter and Full Year 2017 Results.*

[5] Benjamin A. Goldstein, Alan E. Hubbard, Adele Cutler, and Lisa F. Barcellos. An application of random forests to a genome-wide association dataset: Methodological considerations new findings. *BMC Genetics*, 11, 2010.

[6] Vairaprakash Gurusamy and Subbu Kannan. Preprocesing techniques for text mining. 2004.

[7] IBM Knowledge Center. *About text mining.* Accessed February 2, 2018.

[8] M. L. Jockers. *Syuzhet: Extract sentiment and plot arcs from text*, 2017.

[9] Michele L. Joyner and Nicholas Joyner. Using predictive modeling to examine the potential implications of using ranked choice voting in the 2016 presidential election. 2017.

[10] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[11] Raffi Krikorian. *New tweets per second record, and how!* Twitter, 2013. Accessed January 11, 2018.

[12] Michael A. Lazarus, A. U. Mattias Blume, Kenneth B. Brown, William R. Caid, Ted E. Dunning, Larry S. Peranich, Gerald R. Russell, and Kevin L. Sitze. *Predictive modeling of consumer financial behavior*, 1999.

[13] Zi Ning, Ji Soo Yi, Mathew J. Palakal, and Anna McDaniel. Oncoviz: A user-centric mining and visualization tool for cancer-related literature. *Symposium of Applied Computing*, pages 1827–1828, 2010.

[14] Thais Mayumi Oshiro, Pedro Santoro Perez, and Jose Ausgusto Baranauskas. How many trees in a random forest? *MLDM*, 2012.

[15] Byung-Kwon Park and Il-Yeol Song. Toward total business intelligence incorporating structured and unstructured data. *Business Intelligence and the Web*, pages 12–19, 2011.

[16] Python Language Reference. *Python Software Foundation.*

[17] Sebastian Raschka. *Python Machine Learning.* Packt Publishing, Birmingham, UK, 2016.

[18] Reuters. *Instagram's user base grows to more than 500 million.* Accessed April 10, 2018.

[19] Tyler W. Rinker. *Package 'sentimentr'*, 2018.

[20] David Robinson. Does sentiment analysis work? a tidy analysis of yelp reviews, 2017.

[21] Semio Corporation. *Semio Introduces Intranet-Based 'Text Mining' Solution For Increasing the Value of Corporate Information*, 1998.

[22] Anuh Sharma and Shubhamoy Dey. A comparative study of feature selection and machine learning techniques for sentiment analysis. *Reserach in Applied Computation Symposium*, 2012.

[23] Ying Tang. *Real-Time Automatic Face Tracking Using Adaptive Random Forests*. Department of Electrical and Computer Engineering, 2010.

[24] Yelp! *Yelp! Dataset Challenge*.

[25] Chen Yong. Stemming, 2015. Accessed March 8, 2017.