Yelp's Review Filtering Algorithm

Yao Yao[1], Ivelin Angelov[1], Jack Rasmus-Vorrath[1], Mooyoung Lee[1], Daniel W Engels[1]

[1] Southern Methodist University, Dallas, TX, US
{yaoyao, iangelov, jrasmusvorrath, mooyoungl, dwe}@smu.edu

Abstract. In this paper, we investigated which features have the most weight in influencing Yelp's review filtering algorithm. Misclassification of recommended reviews affect average rating, consumer decisions, and business revenue. We systematically sampled and scraped for Yelp's restaurant reviews. Features are created from the reviews' metadata, text sentiment analysis, and text classifier scores. To compare recommended and not recommended reviews, features are adjusted, scaled, and balanced. At 78% prediction accuracy, the multivariate logistic regression model was used to find the coefficient of weights. The weights for filtering reviews are logical towards Yelp's efforts to promote quality and reliable information to help consumers gain insight and make decisions [1].

1 Introduction

Algorithms can inherently have unethical procedures when filtering out deceptive reviews. Yelp is a third-party online platform where users find and review local businesses. People seeking advice or businesses seeking feedback will find crowd-sourced 1-to-5 star ratings paired with written context. Contributed reviews vary in detail and opinion, where some are deceptive or disruptive. Yelp filters for recommended reviews to promote quality and reliable information to help consumers gain insight and make decisions [1]. Reviews are filtered out based on user activity to reduce fraudulent accounts that submit deceptive, disruptive, and paid reviews [2]. Not recommended reviews are still accessible but are not calculated towards average rating [1]. Each star increase in average rating increases the corresponding business revenue by 5 to 9% [3].

Yelp's filtering algorithm can misclassify credible reviews as not recommended and deceptive reviews as recommended. Adjustments to the filtering algorithm will change which reviews are recommended thus affecting the average rating [1]. Yelp has the power to influence consumer decisions and impact business revenue based on their corresponding recommended reviews and average ratings.

For a guideline on how to submit a recommended review, we investigated which features have the most weight in influencing Yelp's review filtering algorithm. We systematically sampled and scraped for Yelp's restaurant reviews. Features are created from the reviews' metadata, text sentiment analysis, and text classifier scores. To compare recommended and not recommended reviews, features are adjusted for data distribution asymmetry and scaled from 0 to 1. The reviews are then balanced for equal observations and modeled with multivariate logistic regression to find the coefficients, which reflect feature importance.

The reduced model has an prediction accuracy score of 77.61% and a F1-Score of 76.79%. To submit a review that is more likely to be recommended, compose an overall positive message in multiple sentences that express variations in sentiment. Rating a business higher than the average rating or having too many sentences would result in a not recommended review. Users that have a larger number of friends, reviews, and photos submitted also increases their likelihood to be recommended. Recommended reviews also are based on recent submission, higher text readability, and less stop words. Reviews are less recommended if many reviews already exist per business and if the review is edited.

The features important for filtering reviews are logical towards Yelp's efforts to promote quality and reliable information to help consumers gain insight and make decisions [1]. Quality of text is promoted by reviews with higher readability and less stop words. Reliability of content is promoted by recent reviews from users with more activity and submitted data. Insight is gained by the variation of sentimental context of the collective experience. Ultimately, consumers' decisions are based on personal discretion and Yelp's filtering algorithm only help to create more informed decisions. Although business revenue is affected by information on Yelp, the filtered reviews and average rating only serves as a justified reflection of the collective experience.

The remainder of this paper is organized as follows: Section 2 outlines Yelp's business model, third-party representation of businesses, and reasons to filter reviews. Section 3 covers the sampling procedure to maintain consistency with the total dataset and pre-balancing the observations prior to predictive modeling. Section 4 explores attributes of the data file and generates enumerated features from the metadata. Section 5 explains the workflow of feature creation, model selection, and coefficient analysis. Section 6 generates features from the review text with the bag of words model, naive Bayes text classifiers, and sentiment natural language processing (NLP). Section 7 evaluates how mean and correlation determines the difference between recommended from not recommended reviews. Section 8 determines the features that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients. Section 9 describes the guideline to write a recommended review and explain the insignificant features of the reduced model. Section 10 describes Yelp's role in helping society make better informed decisions while filtering reviews. Section 11 concludes that Yelp has justified reasons to filter reviews that promote quality and reliable information.

2 Yelp

Yelp is a multinational online platform where consumers voluntarily rate businesses on a scale of 1 to 5 stars, post pictures, and compose feedback in the form of short summary titles and long detailed reviews [4]. Headquartered in San Francisco, Yelp was founded in October 2004 by former PayPal employees Russel Simmons and Jeremy Stoppelman [5]. Yelp was created as an online "yellow pages" directory where people can solicit "help" and advice on finding the best local businesses [6]. Users are encouraged to self identify with their real name and profile picture for nominations to Elite Squad, where the frequency of writing quality reviews and visiting new establishments are met with benefits [7]. Online Yelp interactions include networking to friend local reviewers, complimenting reviews, and reporting reviews [4].

2.1 Business Model

According to Yelp's 2017 financial report, net revenue grew 19% since 2016 to $846.8 million, where advertising constitutes $771.6 million [8]. The other $75.2 million includes net revenue from other acquired services such as food delivery, a waitlist app, and sponsored WiFi [8]. Since 2016, paid advertising accounts grew 21% to 163,000 [8], where the average paid advertising account spends $4,730 a year.

Since inception, Yelp has accumulated 155 million reviews, where 72% are recommended, 21% are not recommended, and 7% are removed if they breach Yelp's term of service [9]. Yelp's metrics as of March 2018 indicate that per monthly basis, the Yelp app averages 30 million unique visitors, the Yelp mobile website averages 70 million unique visitors, and the Yelp desktop website averages 74 million unique visitors [9]. 79% of searches and 65% of reviews are on mobile devices [9]. The rating distribution of all reviews indicate that 48% are 5 stars, 20% are 4 stars, 9% are 3 stars, 7% are 2 stars, and 16% are 1 star [9]. The top 3 reviewed businesses by category are shopping at 21%, restaurants at 17%, and home and local services at 14% [9]. The top represented US demographics of Yelp reviewers are 37% for 35-54 year olds, where 59% finished college, and 49.6% have an income greater than $100k [9].

2.2 Representing Businesses

Businesses can claim their pages on Yelp, which allows them to add menu items, offer discounts, directly respond to reviews publically or privately, and see detailed traffic reports via Yelp's mobile app for businesses. [10]. Once verified, business owners are no longer allowed to submit reviews to Yelp [11]. To sign up for advertising, businesses are required to have at least an average rating of 3 stars for their sponsored ad listing to show at the top of the search results or on their competitors' Yelp profiles [11] [12].

In order to deter misconceptions that advertisers are able to marginalize negative reviews for pay [13][14], Yelp has a delicate balancing act to recommend and filter out reviews based on legitimacy while protecting the data of the reviewers [15]. Each star increase in Yelp rating leads to a 5 to 9% increase in revenue, where this effect is driven by independent restaurants and not affected by those with chain affiliation [3]. Only recommended reviews are calculated in the average star rating of businesses, where the non-recommended reviews are still accessible and could be recommended if the Yelp algorithm changes [1].

Yelp strives to be a platform for small and large businesses alike to be ranked and evaluated by the public on an even playing field. Many businesses say that Yelp has a conflict of interest because its main source of income is through ad sales, where businesses could pay their way into showing up on more search results and on the pages of rival businesses [11][16]. Yelp has denied any wrongdoing because the algorithm to filter out reviews are the same for everyone and ads are a way for the website to make revenue while providing a free service accessible by everyone [17].

Yelp receives an average of 6 subpoenas per month from businesses inquiring about their posted reviews that could lead into defamation cases from deceptive reviews [18][19]. Businesses have improperly used disparagement clauses to sue or fine customers who give negative reviews online [20]. Yelp helped invoke a 2014 California State law that protects the user from business scrutiny to promote site integrity that users are able to share reviews without the influence of bribery [21].

Harvard Business School found no significant correlation between advertising and having better ratings and reviews on Yelp [22]. Since non-recommended reviews are still accessible, Yelp does not censor free speech while prescribing recommended reviews [23]. Multiple court rulings and dismissals suggest that there were not enough substantial evidence that Yelp was manipulating the recommendation of reviews [24]. Businesses paying for Yelp advertising does not fall within the legal definition of extortion [25].

2.3 Filtering Reviews

Yelp filters for recommended reviews to promote quality and reliable information to help consumers gain insight and make decisions [1]. Not everyone uses Yelp to solicit information or submits advice-driven reviews based on personal consumer experience. The responses of those people who participate can be driven by alternative incentives including bribery or complaint. Deceptive reviews are purposefully misleading and disruptive reviews contain unrelated content or non-understandable language. Besides defamation court rulings, users can report inappropriate submitted content if it breaches Yelp's terms of service.

Yelp does not disclose how their filtering algorithm works, which would reveal information on how to manipulate the system [1]. Yelp uses a myriad of approaches to filter reviews to evaluate whether a review is authentic and based on first hand experiences [26]. Reviews are filtered out based on user activity to reduce fraudulent accounts that submit deceptive, disruptive, and paid reviews [27].

Yelp has extensive methods where they cross examine review patterns left on different Yelp pages to uncover if businesses are trying to pay for more stars and better reviews [26]. As part of a sting operation, Yelp has found and filed lawsuits against 19 third-party websites for writing fraudulent reviews [26]. It was determined that a jewelry store in San Diego was willing to pay customers $5 per review, where they would pay $30 per 50 words upwards to $200 per review [27]. A court ruling forced a jewelry store in Massachusetts to pay $34,500 in damages for writing a negative review for a rival business [28].

Yelp has become a platform of protest or support for businesses' political views, where reviews motivated by current events are manually or algorithmically removed [29][30]. The Yelp algorithm seeks to filter out purchased reviews from third-party sources, reviews made by those affiliated with business owners, politically-motivated reviews, and reviews with unrelated content or non-understandable language [1]. Further analysis will be made on how Yelp achieves this filtering process through the comparison of recommended and not recommended reviews.

3 Yelp Dataset Collection

The official Yelp dataset challenge does not include not recommended reviews to conduct a study on their filtering algorithm. Promotional datasets may inherit unwarranted biases and an external audit through sampling allows for a better observational study. Gathering all 155 million reviews for every business documented on Yelp is not feasible because of the search limit and changes in the dynamic ordering of search results [5][9].

Yelp's dynamic ordering of results based on reviews and average rating also creates duplicates and skipped observations in the systematic scraping process. Yelp created individual web pages per local business, but obscure businesses are less likely to be reviewed. In addition, some cities have a low adoption rate for using Yelp. For metropolitan areas, over 5000 businesses exist yet only the first 1000 are available per searched city. To sample a dataset that maintains consistency with the total dataset, a two-stage sampling design with clustering and stratification was applied. A Python-activated Selenium browser programmatically scraped for Yelp's recommended and not recommended reviews.

3.1 Sampling Procedure

Yelp lists the various cities that adopts Yelp as a review platform [31][Table A]. Searching for cities on Yelp lists businesses by category, where restaurants have the most consistent participation for reviewers per demographic for every city size [9]. To preserve a similar distribution of participation for every city, only restaurant data was gathered. Since the review content would eventually undergo natural language processing, only restaurants in US cities were evaluated for English reviews. Python script was fed into a Selenium browser to mimic user behavior when scraping for Yelp data [32].

Two-stage sampling with clustering and stratification is a proportional method that preserves certain sampled attributes to represent the larger dataset by projection [33]. For cluster sampling, the population is separated into city subgroups, called clusters, where a sample of cities is drawn in blue [Figure 1]. For stratified sampling, the population is then separated into restaurant subgroups, called strata, where separate restaurant samples are drawn from each city subgroup in red [Figure 1]. The two-stage sampled dataset contains a proportionate number of each restaurant represented by city to facilitate restaurant comparisons between different cities [33]. Random sampling reduces the chances of recording duplicate and skipped observations compared to systematic scraping yet the dataset still underwent the removal of duplicates.

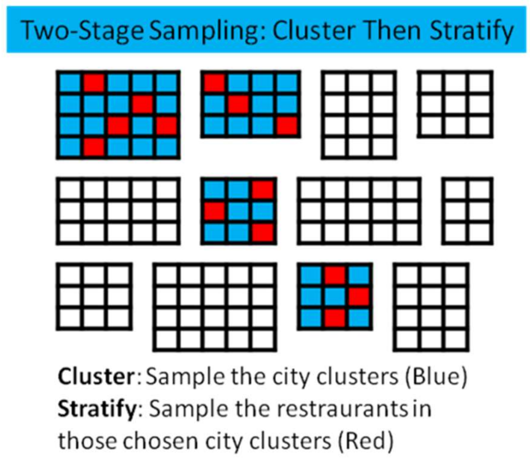

Figure 1. Two-stage sampling for clustering and stratification preserves equal probability selection in order to maintain consistency with the total dataset

Cut-off sampling is the assumption that reviews from popular restaurants can extrapolate for the lack of reviews for obscure restaurants [33]. The assumption may distort the sampled population by either over- or under-representing certain aspects of the dataset yet Yelp's search limit requires cut-off sampling [33]. The order in which restaurants rank based on reviews and average rating results in more reviews by popularity. Since more data exists for popular restaurants, the assumptions of cut-off sampling makes it practical to sample the available restaurants within Yelp's 1000 result limit.

3.2 Dataset Projection and Balancing

Using 676 restaurants from 157 cities, the two-stage sampling procedure with cut-off scraped 300,428 recommended and 47,389 not recommended reviews [31][Table A]. After cleaning for missing values and duplicates to compensate for dynamic search results, 224,604 reviews were recommended and 26,824 reviews were not recommended. Using equal probability projection, our dataset represents 198 million recommended reviews and 24 million not recommended reviews, which is more than the total 155 million reviews [9]. In our collected dataset, the ratio is 89% recommended and 11% not recommended, where the removed reviews are not accessible.

The dataset has to be pre-balanced with equal observations in both categories so that the outcomes of prediction does not have an initial 89% advantage in choosing correctly. Pre-balancing the dataset to 50:50 also forces the model to not weigh the larger number of observations with higher accuracy. Even with the option for the model to post-hoc balance, recall and F1-scores are subpar when predicting for the lesser number of observations.

Two papers that used models to find which Yelp features had more weight did not pre-balance and had inflated accuracy scores of 90% and above [34][35]. We are skeptical of their results because they only used Yelp's metadata for their model, where model has to be simple to lure out the weights of features. Any complicated model-fitting for accuracy have hidden layers and feature interactions that convolute the weights of feature influence. We chose multivariate logistic regression because the prediction of recommended and not recommended is binary and straight forward in explaining which features influence Yelp's review filtering algorithm the most.

4 Collected Yelp Dataset

Scraping does have its disadvantages in not being able to fully access Yelp's internal metadata such as page visitation information [36]. Scraping is an external audit where data has to be labeled, merged, and combined. Review compliments are manually labeled by other users and is more towards the ordering of already recommended reviews rather than the filtering process itself [4]. Compliments are only allowed for recommended reviews and data for this study are only compared if they exist for both recommended and not recommended reviews [4]. Therefore, individual's user-page metadata per review was not scraped. Most of the later user metadata additions were added in 2013 [37], 9 years after Yelp's first inception, and would not affect older reviews from being filtered using the newer features.

4.1 Data File

The scraped dataset contains restraurants.csv [Table 1] and reviews.csv [Table 2]. Restaurant data are combined with reviews by restaurant ID to create extra features for the analysis. Booleans and strings are processed and enumerated for the multivariate logistic regression model.

For the restaurant dataset, strings include restaurant name, address, city, and Yelp link [Table 1]. Floats and integers include average rating, number of reviews restaurant, number of restaurants, restaurant ID, and restaurant listing order. Restaurant city and internal restaurant links created by Yelp are irrelevant to a balanced dataset, where there are equal number of recommended and not recommended reviews that refer to the same city and links. The same restaurant name can exist in multiple addresses and are enumerated by word count to prove that the sampling procedure had an equal stratification probability along with restaurant listing order in choosing restaurants randomly per city. Due to outliers and asymmetry in the data distribution, number of reviews restaurant, number of restaurants are logarithmically transformed to fix skewness.

Table 1. Restaurants.csv contains restaurant data, which are bound with reviews by Restaurant ID to create extra features.

Category	Data Type	Description	Example

Name	String	Restaurant name	Garaje
Address	String	Full address	475 3rd St San Francisco, CA 94107
City	String	City hub	San Francisco
Average Rating	Float	Rounded to half-stars	4.5
Number of Reviews Restaurant	Integer	Number of reviews	1354
Number of Restaurants	Integer	Number of restaurants in city hub	4829
Restaurant Link	String	Yelp link	https://www.yelp.com/biz/garaje-san-francisco
Restaurant Listing Order	Integer	Yelp restaurant listing order	2
Restaurant ID	Integer	Bind with Reviews.csv	0

For the reviews dataset, strings include date, location, text, and username [Table 2]. Integers include number of friends, number of photos, use rating, restaurant ID, number of reviews user, and recommended. The profile picture Boolean can be enumerated by 1 as true and 0 as false. The month-day-year format of date is enumerated by subtracting the number of days since Yelp's inception into number of days published. Date also has an string indicating if the review was updated, which could be enumerated with the same Boolean method into text has been edited. Multiple users may have the same first and last initial and would be difficult to create factors to cross examine and therefore the column is dropped. Recommended is the predictor integer that the multivariate logistic regression solves based on the categories enumerated into features.

Table 2. Reviews.csv contains full review text data, which is bound with restaruants by Restaurant ID to analyze how Yelp recommends reviews.

Category	Data Type	Description	Example
Date	String	Date formatted MM-DD-YYYY, Also shows updated review	3/9/2016
Number of Friends	Integer	Number of user's friends, max at 5000	22
Has Profile Picture	Boolean	True or false for profile picture	True
Location	String	City, State of user location	San Diego, CA
Number of Photos	Integer	Number of total photos taken	122
User Rating	Integer	Rating from 1 to 5	5
Restaurant ID	Integer	Bind with Restaurant.csv	0
Number of Reviews by User	Integer	Number of reviews that the user made	7
Text	String	Review text	Great place to hang
Username	String	First name, last initial	Alex, B.
Recommended	Integer	0 for false, 1 for true	1

4.2 Adding Features

Multivariate logistic regression requires all data to be enumerated for the prediction model. Review text is cleaned for special formatting and special characters and is converted into the number of sentences by punctuation, number of words by letter clusters, and number of words without stop words, which is the removal of common semantic words. The difference between user to average rating is another feature created called user to average rating. To enumerate user location and restaurant location, the distance in miles between user to restaurant was obtained from the Google Maps API, which is an interface to programmatically query and retrieve information from websites [38]. Number of sentences, number of words, number of words no stop words, number of friends, number of photos, and number of reviews user are also logarithmically transformed to fix asymmetry in the data distribution. Recommended ratio is the ratio of recommended reviews per restaurant ID.

Table 3. Attributes of features created from merging review with restaurant data, as a result of various enumarated conversions. Asterisk (*) denotes data before logarithmic transformation.

Category	Data Type	Description	Example
Number of Days Published*	Float	Difference in days between review submission and October 1, 2004	525
Has Been Edited	Integer	0 for false, 1 for true	0
Number of Friends*	Float	Number of user's friends, max at 5000	22
Has Profile Picture	Integer	0 for false, 1 for true	1
User to Restaurant Distance*	Float	Distance between user and restaurant location in miles	522
Number of Photos of User*	Float	Number of total photos taken by user	122
User Rating	Integer	Rating from 1 to 5	5
Number of Reviews User*	Float	Number of reviews that the user made	7
Word Length of Text*	Float	Word length of review text	4
Word Length of Text Without Stopwords*	Float	Word length of review text with no stopwords	3
Sentence Length of Text*	Float	Sentence length of review text	1
Recommended	Integer	0 for false, 1 for true	1
Recommended Ratio	Float	Number of recommended reviews divided by total reviews	0.9212
Word Length of Restaurant Name	Float	Word length of restaurant name	1
Word Length of Restaurant Address*	Float	Word length of restaurant address	7
Average Rating	Float	Rounded to half-stars	4.5
User to Average Rating	Float	User rating subtracted by average restaurant rating	0.5
Number of Reviews Restaurant*	Float	Number of reviews of restaurant	1354
Number of Restaurants in City*	Float	Number of restaurants in city hub	4829
Restaurant Listing Order	Integer	Yelp restaurant listing order	2

The Yelp data set could be appropriated to represent all review datasets and how an algorithm could filter reviews based on text and enumerated metadata. The influence of a filtering system can be contextualized for broader data sets such as movie, music, shopping, and search results. Data similarity in sparse, contextualized text submitted by users are applied to text classifier datasets and sentiment natural language processing [39][40][41].

The generalization of the dataset is the application of review filtering systems for volunteered feedback. The scope of the implications would be how filtering systems could benefit and adversely affect certain businesses and the user experience if the system is left unregulated.

5 Methods and Experiments

A correlation and prediction model compares scaled numerical features from review metadata and text analysis. Coefficients from correlation and multivariate logistic regression are compared to analyze which features have the most influence on Yelp's algorithm for filtering reviews.

5.1 Feature Creation and Prediction Modeling

Two-stage sampling preserves equal probability selection to maintain scraped data consistency with the total dataset [33]. Balancing the dataset to 50:50 for recommended and not recommended reviews by restaurant represents how the Yelp algorithm filters reviews per business page. Features are created from metadata and text analysis, which range from word count to complex natural language processing. All features are converted into numerical and logarithmically transformed as necessary to reduce the asymmetry in the data distribution. Natural log transformation is a method to keep data integrity while solving for data normality in an even distribution. All features are scaled from 0 to 1 so that direct comparisons are made for the model's coefficients. Multivariate logistic regression is the model that uses features to predict the binary outcomes of recommended and not recommended reviews. The prediction model produces feature coefficients to predict recommended and not recommended reviews.

For any binary prediction, there are true positives, true negatives, false positives, and false negatives. True positives and negatives accurately predict labels while false positives and negatives are misclassifications. Different from accuracy [Eq. 1], precision is used to optimize for model prediction [Eq. 2]. Model evaluations for predicting true positives also include recall [Eq. 3] and F1-Score [Eq. 4], which is the harmonic mean between precision and recall. These metrics are used throughout.

$$Accuracy = \frac{True\ \ Positives + True\ \ Negatives}{Total\ \ Observations} \quad [Eq. 1]$$

$$Precision = \frac{True\ \ Positives}{True\ \ Positives + False\ \ Positives} \quad [Eq. 2]$$

$$Recall = \frac{True\ \ Positives}{True\ \ Positives + False\ \ Negatives} \quad [Eq. 3]$$

$$F1\ \ Score = \frac{2 * True\ \ Positives}{2 * True\ \ Positives + False\ \ Positives + False\ \ Negatives} \quad [Eq. 4]$$

5.2 Analysis of Feature Importance

Features created and analyzed to verify Yelp's efforts to promote quality and reliable information to help consumers gain insight and make decisions [1]. The determining features that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients. Some features are insignificant towards prediction and are removed from the full model in the reduced model for multivariate logistic prediction. Significance is determined by a p-value threshold of 0.05, which is the hypothesis test statistic to determine probable evidence towards an observation. The full and reduced model are explored to evaluate which features have a significant role in determining what causes reviews to become recommended or not recommended. The reduced model evaluated which features have a significant role in determining what causes reviews to become recommended or not recommended. A guideline on how to submit a recommended review is the result of the investigation.

6 Text Processing

Features are extracted from the review text with the bag of words model, naive Bayes text classifiers, and sentiment natural language processing (NLP). The bag of words model is a method that process word count and frequencies without checking for grammar or word order [42]. Naive Bayes is the matching the probabilities of frequent words occurring in labeled text classifiers with the probabilities of those in the review text [43]. Sentiment NLP is how words arranged in a certain order can affect the tonality of a sentence [44].

6.1 Bag of Words Model

Features from review text are processed as a bag of words model using readability indexes, which is a semantic metric to statistically solve for the difficulty of understanding text [45]. The total number of characters, word count from letter clusters, and sentences based on punctuation are used to solve for the automated readability index of review text [Eq. 5][45]. Age and grade level readability by score of the Automated Readability Index are listed in Table 4 [45]. For the Flesch–Kincaid Grade Level Formula [46], the total number of syllables are extracted from the Google dictionary API to directly solve for the grade level readability of review text [Eq. 6][47]. The Google dictionary API was also used to find the percentage of words spelled correctly for the review text [47].

$$Automated\ Readability\ Index = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43 \quad [Eq. 5]$$

$$Flesch\text{–}Kincaid\ Grade\ Level\ Formula = 0.39\left(\frac{words}{sentences}\right) + 11.8\left(\frac{syllables}{words}\right) \quad [Eq. 6]$$

Table 4. The Automated Readability Index by score is based on age and grade level readability [45].

Score	Age	Grade Level
1	5-6	Kindergarten
2	6-7	First Grade

3	7-8	Second Grade
4	8-9	Third Grade
5	9-10	Fourth Grade
6	10-11	Fifth Grade
7	11-12	Sixth Grade
8	12-13	Seventh Grade
9	13-14	Eighth Grade
10	14-15	Ninth Grade
11	15-16	Tenth Grade
12	16-17	Eleventh grade
13	17-18	Twelfth grade
14	18-22	College

6.2 Naive Bayes Text Classifiers

Naive Bayes text classifiers is a continuation of the bag of words assumption, where it does not account for grammar or word order [43]. Naive Bayes stems from the Bayes Theorem equation and is the assumption that probability calculations are isolated events without cause or effect. The word frequencies of a sentence are tallied or vectorized to calculate the probability of that sentence is of a certain label [Table 5]. Given that the trained text is labeled positive, the probability is calculated by finding which words has a larger difference in occurrence for positive and not positive labels.

Table 5. Vectorizing the word frequency of a sentence and calculating the probability that the sentence is labelled positive.

Trained Text	Positive Label	Word Vectors	This	Place	Is	Good	The	Bad
This place is good.	1		1	1	1	1	0	0
The place is good.	1		0	1	1	1	1	0
This place is bad.	0		1	1	1	0	0	1
The place is bad.	0		0	1	1	0	1	1
p(label=1)	0.5	p(Word|1)	0.5	1	1	1	0.5	0
p(label=0)	0.5	p(Word|0)	0.5	1	1	0	0.5	1

The Bayes Theorem equation solves for the relative proportion of probabilities for the Naive Bayes classifier [43]. A trained text classifier uses word vectors to predict for spam with 90% probability yet misclassify for not spam with 20% probability [Figure 2]. The relative proportions of spam and not spam show that spam is classified correctly with 90% * 25% = 22.5% probability yet spam is classified incorrectly with 20% * 75% = 15% probability. Adjusted for the relative proportion size of both labels, the probability that a trained word vector can predict for spam is 60% [Eq. 7] [43]. For the simplified example, the Bayes Theorem equation is equivalent to that of precision [Eq. 2] and becomes complex as more words are vectorized.

$$p(Spam|Trained\ \ Words) = \frac{p(Trained\ \ Words|Spam)p(Spam)}{p(Trained\ \ Words)}$$

$$= \frac{90\% * 25\%}{75\% * 20\% + 90\% * 25\%} = 60\% \ \ [Eq. 7]$$

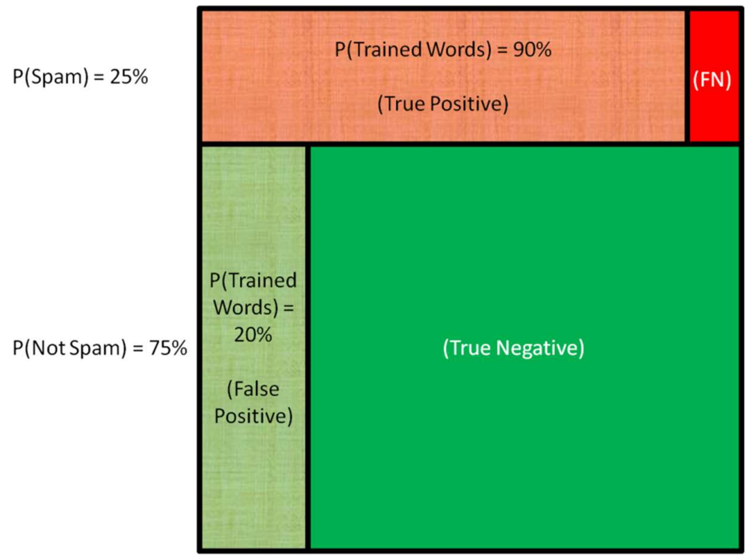

Figure 2. A trained text classifier uses word vectors to predict for spam with 90% probability yet misclassify for not spam with 20% probability. The relative proportions of spam and not spam show that spam is classified correctly with 90% * 25% = 22.5% probability yet spam is classified incorrectly with 20% * 75% = 15% probability. FN is false negative.

6.3 Deceptive Opinion

The Deceptive Opinion Spam Corpus includes labeled reviews from online communities to evaluate truthful and deceptive text [48]. Deceptive reviews come from unreliable sources and often misguide consumers. Truthful reviews for the Corpus dataset are scraped from TripAdvisor, Expedia, Hotels.com, Orbitz, and Priceline [Table 6][48]. The review content are similar to that of Yelp in which users contribute opinionated reviews with similar community guidelines and functionality. Deceptive reviews are scraped from Amazon Mechanical Turk, which is a platform where users write reviews to earn money. None of the other mentioned review platforms have a monetary system, which alters the incentives to submit a misguided review for monetary gains [48].

Table 6. The Deceptive Opinion Spam Corpus includes labeled reviews from online communities to evaluate truthful and deceptive text [48].

Review	Label	Quality	Data Origin
Positive	Truthful	400	TripAdvisor
Positive	Deceptive	400	Amazon Mechanical Turk
Negative	Truthful	400	Expedia, Hotels.com, Orbitz, Priceline
Negative	Deceptive	400	Amazon Mechanical Turk

For the purposes of Naive Bayes text classification, only deceptive and truthful labels are used to train word vectors known as deceptive. Stratified k-fold is a sampling process in which the observations of the full Corpus dataset is arranged randomly by 66% training and 33% testing to optimize the prediction model with multiple iterations for the highest precision. For the Corpus dataset, the text classifier precision for deceptive reviews is 88% while recall and F1-score are also 88%.

6.4 Extreme Comments

Yelp will comments if they breach their terms of service [9]. Extreme text classifiers can test if the Yelp algorithm filters out extreme content prior to manual removal. Conversation AI is founded by Google to filter out online harassment in comments, where some have multiple labels [40]. The comments are from Google services such as YouTube, Blogger, Google Maps, and Google+ and are similar to that of Yelp in terms of online platform interactions and community feedback [49]. The labeled dataset contains 159,574 observations where 15,294 are toxic, 1,595 are severely toxic, 8,449 are obscene, 478 has threats, 7,877 have insults, and 1,405 have identity hate [40]. Using the same text classifier procedure to create deceptive score, the precision, recall, and F1-score results are shown in Table 7. Precision, recall, and F1-Score are higher for validation due to the limited observations of some labels.

Table 7. Text classifiers for extreme comments include toxicicity, severe toxicicity, obscenity, threats, insults, and identity hate.

Classifier	Precision	Recall	F1-Score
Toxic	0.96	0.96	0.96
Severely Toxic	0.99	0.99	0.99
Obscene	0.98	0.98	0.98
Threat	1	1	1
Insult	0.97	0.97	0.97
Identity Hate	0.99	0.99	0.99

6.5 Sentiment NLP

The Stanford NLP system architecture for detecting sentiment analysis takes word order into consideration when detecting the overall tonality of a sentence [Table 8] [50]. Words from a sentence are identified as individual strings. Compound and complex sentences are split into fragments by punctuation. Words are identified as nouns, verbs, adjectives, and adverbs. Word families are identified by root word, suffix, and prefix analysis. Proper nouns are identified. Grammar rules are applied to identify the logic behind the sentence composition. Gender is identified and pronouns are then linked to nouns. Using definitions, words are labeled as very positive, positive, neutral, negative, or very negative.

Table 8. Execution flow of the Stanford NLP system architecture for sentence sentiment analysis [50].

Procedure	Description
Tokenization	Classifying words as individual strings
Sentence Splitting	Sentences are split into fragments by punctuation
Parts of Speech Tagging	Identify words as nouns, verbs, adjectives, and adverbs
Morphological Analysis	Identify word families, root words, suffixes, and prefixes
Named Entity Recognition	Identify proper nouns
Syntactic Parsing	Apply grammar rules to identify the logic in sentence composition
Coreference Resolution	Identify gender and link pronouns to nouns
Sentiment Annotation	By word definition, label as very positive, positive, neutral, negative, or very negative

A recursive tree structure fragments and uses grammar rules to find the tonality of nested phrases stemming to individual words [Figure 3][44]. A recurrent neural tensor network (RNTN) has hidden layers that feed grammar-guided prediction outcomes into the tree hierarchy system, which dictates word order [51] Solving for tree hierarchy of related phrases and word fragment structures creates the sentiment label of the sentence. From the example, a comma splits the sentence in two branches, where it starts negative but is overall positive [Figure 3][50]. The RNTN from Stanford research efforts is 87.6% accurate when labeling positive and negative sentences [44].

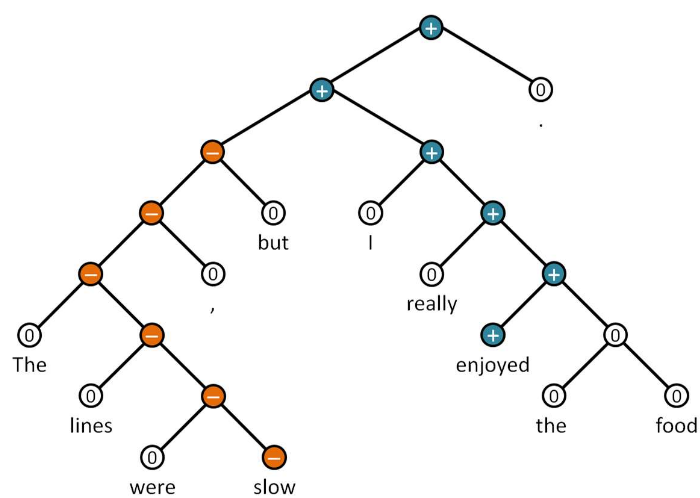

Figure 3. A recursive tree structure fragments and uses grammar rules to find the tonality of nested phrases stemming to individual words to find the sentiment label of the sentence [44]. A comma splits the sentence in two branches, where it starts negative but is overall positive [50].

6.6 Text Features Added

Table 9 shows all the text features created from the bag of words model, naive Bayes text classifiers, and sentiment NLP. Since every sentence in a review has a sentiment tally, the total sentiment is calculated by giving a weight for each category [Eq. 8] [44]. Average sentiment ranges from 1 to 5 and is total sentiment divided by the number of sentences in the review [44]. Average sentiment to user rating is how differently the user rated and wrote the review. Sentiment to average rating is how differently the user wrote the review from the average rating of the restaurant. Each sentiment category tallied also has to be divided by the total number of sentences. Most of the added features are logarithmically transformed to fix asymmetry in the data distribution. Text sentiment to user rating is an intermediate feature created to validate the process of comparing average sentiment score and user rating due to both being on the same 1 to 5 scale. Text sentiment to user rating is removed from the later prediction models because it is redundant to the information expressed in average sentiment rating and user rating.

$$Total\ \ Sentiment = 1*(Very\ \ Negative) + 2*(Negative) + 3*(Neutral) + 4*(Positive) + 5*(Very\ \ Positive)\ \ [Eq.8]$$

Table 9. Attributes of enumerated features created from bag of words model, naive Bayes text classifiers, and sentiment NLP. Asterisk (*) denotes data before logarithmic transformation.

Category	Data Type	Description	Example
Text Readability AR Score*	Float	The Automated Readability score is based on age and grade level	6
Text Readability FK Score*	Float	The Flesch–Kincaid Formula directly solves for grade level for readability	5
Text Spelling Score*	Float	Percentage of review spelled correctly	1
Text Deceptive Score	Float	Probability that review is deceptive	0.2
Text Toxic Score*	Float	Probability that review is toxic	0.11
Text Severely Toxic Score*	Float	Probability that review is severely toxic	0.04
Text Obscene Score*	Float	Probability that review is obscene	0.03
Text Threat Score*	Float	Probability that review has threats	1e-5
Text Insult Score*	Float	Probability that review has insults	0.01
Text Identity Hate Score*	Float	Probability that review has identity hate	0
Text Very Negative*	Float	Percent of sentences that are very negative	0
Text Negative*	Float	Percent of sentences that are negative	2
Text Neutral*	Float	Percent of sentences that are neutral	3
Text Positive*	Float	Percent of sentences that are positive	1
Text Very Positive*	Float	Percent of sentences that are very positive	0

Text Total Sentiment*	Float	Sentiment score weighted by 1 to 5 from their respective categories	17
Text Average Sentiment	Float	Total sentiment divided by sentences	2.833
Sentiment To User Rating	Float	Average user sentiment subtracted by user rating	-2.167
Sentiment To Average Rating	Float	Average user sentiment subtracted by average restaurant rating	-1.667

7 Data Exploration

Features from review metadata and processed text are transformed to adjust for data distribution asymmetry and scaled from 0 to 1 for direct coefficient comparisons. Recommended and not recommended are balanced 50:50 by restaurant where each label has 26,824 observations for an even prediction model. Taking the mean of features for each recommendation label evaluates their magnitude and differences. Pearson's correlation coefficient, which measures the linear relationship between two variables [52], adjusts for remaining feature distribution asymmetry and evaluates how each correspond to recommended reviews.

7.1 Mean Differences

Evaluating for mean differences by magnitude show how features correspond to recommended and not recommended reviews [Figure 4]. Features with mean differences higher than 0.1 for recommended reviews include having a profile picture, number of friends, number of reviews user made, number of photos by user, number of words in text without stopwords, number of words in text, percent of negative sentences, and total number of sentences. Other features with mean differences that are higher for recommended reviews include text total sentiment, user to restaurant distance, text deceptive score, text readability Flesch–Kincaid score, text Automated Readability score, percent of neutral sentences, edited reviews, percent of very negative sentences, number of days after October 2004 the review was published, and text threat score.

Features with mean differences that are higher for not recommended reviews include percentage of positive sentences, text average sentiment, text sentiment to restaurant average rating, user rating, percentage of very positive sentences, user rating to average rating, text sentiment to user rating, text toxic score, text insult score, text obscene score, text severely toxic score, text spelling score, text identity hate score.

Features that show no difference in mean between recommended and not recommended reviews include number of restaurants in city, number of words in restaurant name, number of words of restaurant address, number of reviews for restaurant, review recommended ratio of restaurant, and restaurant listing order.

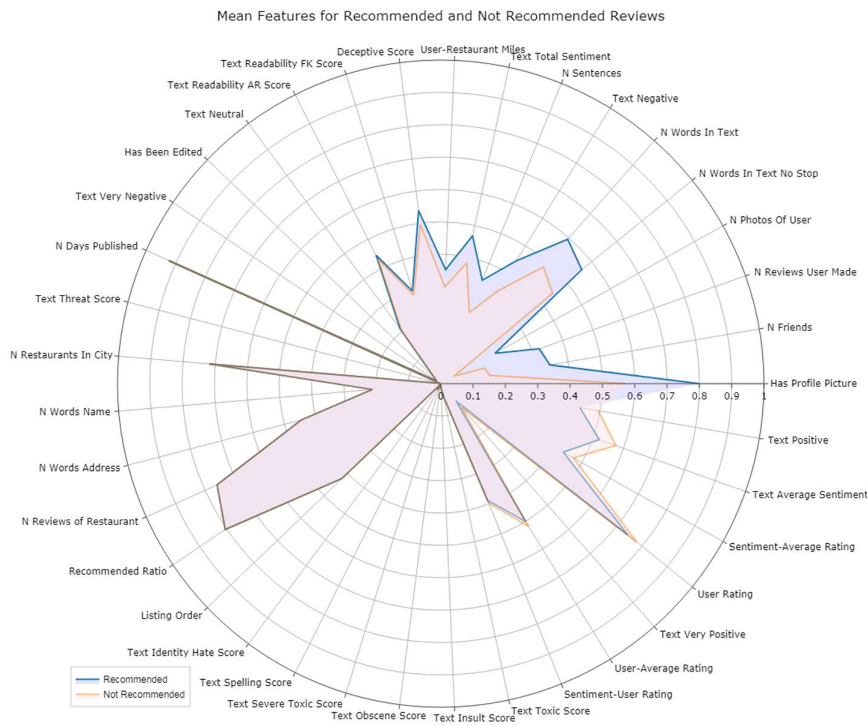

Figure 4. Evaluating for mean differences by magnitude show how features correspond to recommended and not recommended reviews.

7.2 Correlation Coefficients

Pearson's correlation coefficient adjusts for remaining feature distribution asymmetry and evaluates how each feature correspond to recommended reviews [Figure 5][52]. Features with correlation coefficients higher than 0.3 for recommended reviews include number of reviews user made, number of photos by user, number of friends, number of words in text without stopwords, number of words in text, number of sentences, and text total sentiment. Other features with correlation coefficients higher for recommended reviews include having a profile picture, percentage of negative sentences, text deceptive score, text readability Flesch–Kincaid score, user to restaurant distance, text Automated Readability score, percentage of very negative sentences, percentage of neutral sentences, edited reviews, and number of days after October 2004 the review was published.

Features with correlation coefficients higher than 0.1 for not recommended reviews include text average sentiment and text sentiment to average rating. Other features with correlation coefficients higher for not recommended reviews include percentage of positive sentences, percentage of very positive sentences, user to average restaurant rating, user rating, text severely toxic score, text sentiment to user rating, text insult score, text toxic score, text obscene score, text identity hate score, and text spelling score.

Features that show no correlation for recommended reviews include text threat score, number of words in restaurant name, number of reviews of restaurant, review recommended ratio, number of restaurants in city, restaurant listing order, and number of words in address.

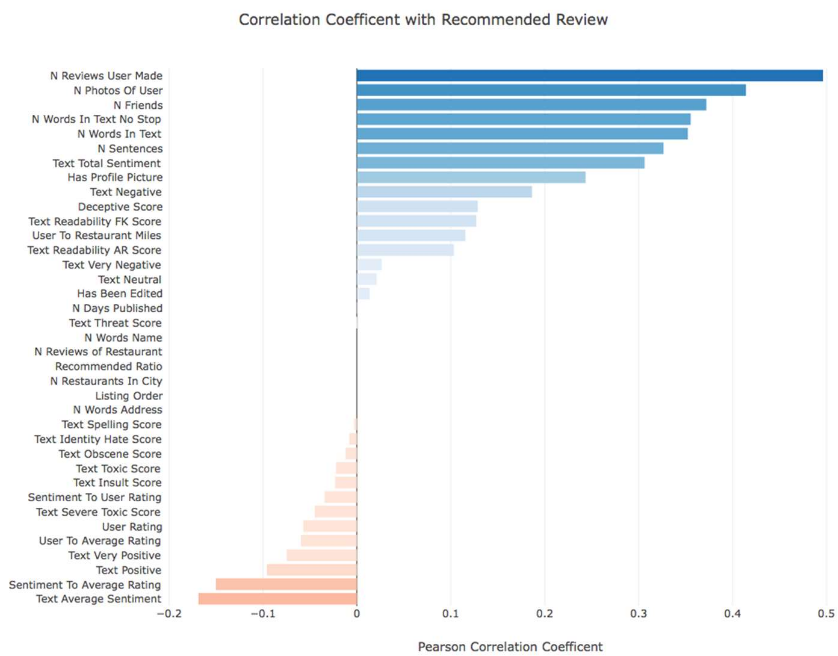

Figure 5. Pearson's correlation coefficient adjusts for remaining feature distribution asymmetry and evaluates how each feature correspond to recommended reviews.

8 Results

The multivariate logistic model produces feature coefficients to predict recommended and not recommended reviews [53]. The determining features that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients. Some features are insignificant towards prediction and are removed from the full model in the reduced model for multivariate logistic prediction. Significance is determined by a p-value threshold of 0.05, which is the hypothesis test statistic to determine probable evidence towards an observation [53]. The full and reduced model are explored to evaluate which features have a significant role in determining what causes reviews to become recommended or not recommended.

8.1 Full Model

The determining features that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients in the full model [Figure 6]. The full model for binary prediction has a 77.56% accuracy score, a 79.75% precision score, a 74.14% recall score, and a 76.84% F1-Score. The predicted results for the full model shows that balancing the dataset predicts for both observations with similar accuracy [Table 8].

Table 8. The predicted results for the full model shows that balancing the dataset predicts for both observations with similar accuracy.

	Predicted Not Recommended	Predicted Recommended
Actual Not Recommended	21647	5018
Actual Recommended	6897	19768

Features with coefficients higher than 10 for predicting recommended reviews include text sentiment to average restaurant rating, user rating, and text total sentiment. Features with coefficients higher than 1 for predicting recommended reviews include the number of reviews the user made, and number of days after October 2004 the review was published, number of words in text without stopwords, text threat score, number of words in text, percentage of very negative sentences, number of photos of user, number of friends, and text Automated Readability score.

Features with coefficients higher than 10 for predicting not recommended reviews include user rating to average rating, text average sentiment, and number of sentences. Features with coefficients higher than 1 for predicting not recommended reviews include text severe toxic score, percentage of very positive sentences, percentage of positive sentences, and text identity hate score.

Other features with low magnitudes for coefficients predicting recommended reviews include user to restaurant distance, text spelling score, deceptive score, percentage of very negative sentences, having a profile picture, text toxic score, number of words of restaurant address, and text readability Flesch–Kincaid score. Other features with low magnitudes for coefficients predicting not recommended reviews include recommended ratio of restaurant reviews, edited review, percentage of neutral sentences, number of reviews of restaurant, number of restaurants in city, text obscene score, number of words in restaurant name, text insult score, and Yelp restaurant listing order.

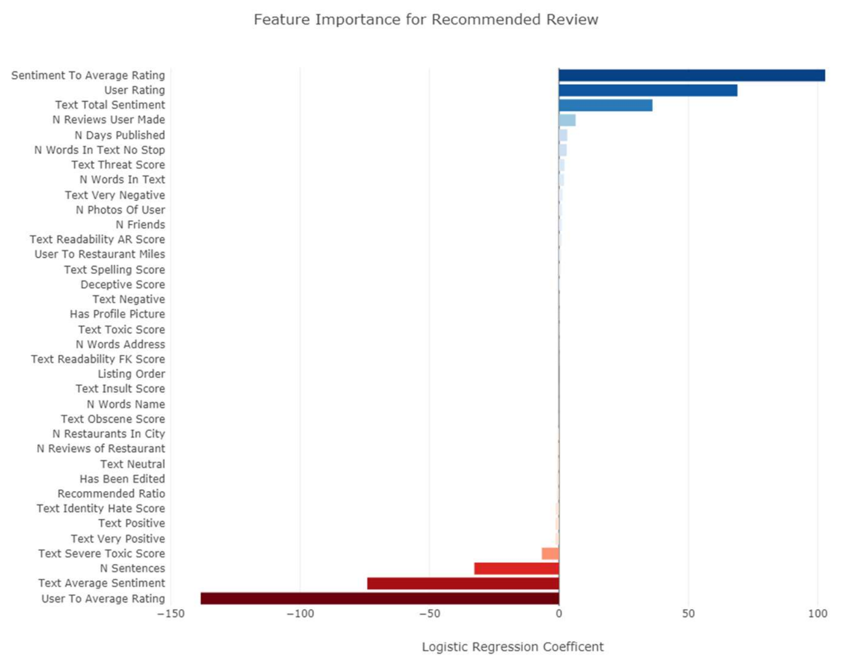

Figure 6. The determining features that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients in the full model.

8.2 Reduced Model

The significant features by that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients in the reduced model [Figure 7]. The reduced model for binary prediction has a 77.61% accuracy score, a 79.71% precision score, a 74.07% recall score, and a 76.79% F1-Score. The predicted results for the reduced model shows that balancing the dataset predicts for both observations with similar accuracy [Table 9].

Table 9. The predicted results for the reduced model shows that balancing the dataset predicts for both observations with similar accuracy.

	Predicted Not Recommended	Predicted Recommended
Actual Not Recommended	21639	5026
Actual Recommended	6913	19752

Features with coefficients higher than 10 for predicting recommended reviews include text sentiment to average restaurant rating, user rating, and text total sentiment. Features with coefficients higher than 1 for predicting recommended reviews include the number of reviews the user made, and number of days after October 2004 the review was published, number of words in text without stopwords, number of words in text, number of photos of user, percentage of very negative sentences, number of friends, and text Automated Readability score.

Features with coefficients higher than 10 for predicting not recommended reviews include user rating to average rating, text average sentiment, and number of sentences. Features with coefficients higher than 1 for predicting not recommended reviews include percentage of positive sentences and percentage of very positive sentences.

Other features with low magnitudes for coefficients predicting recommended reviews include user to restaurant distance, text spelling score, deceptive score, and having a profile picture. Other features with low magnitudes for coefficients predicting not recommended reviews include percentage of neutral sentences, recommended ratio of restaurant reviews, edited review, number of reviews of restaurant, and number of restaurants in city.

The features that are no longer significant in the reduced model include percentage of negative sentences, text readability Flesch–Kincaid score, text identity hate score, text insult score, text threat score, text obscene score, text severe toxic score, text toxic score, number of words of restaurant name, number of words of restaurant address, and Yelp listing order.

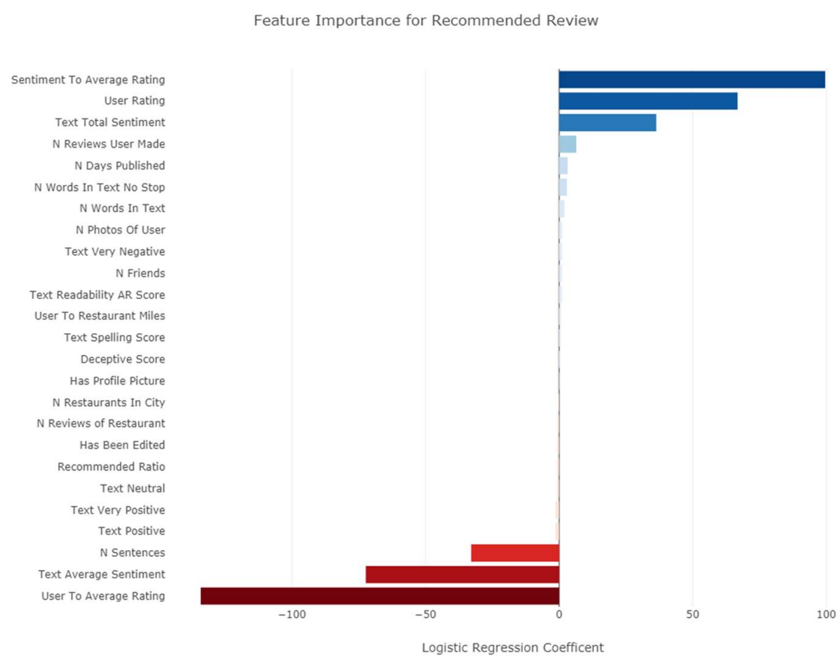

Figure 7. The significant features by that influence Yelp's review filtering algorithm is based on the magnitude of those coefficients in the reduced model.

9 Analysis

The reduced model evaluated which features have a significant role in determining what causes reviews to become recommended or not recommended. A guideline on how to submit a recommended review is the result of the investigation. Insignificant features from the reduced model validates the randomized sampling procedure and the application of text classifiers towards removing extreme comments.

9.1 Guideline For Recommended Review

To submit a review that is more likely to be recommended, compose an overall positive message in multiple sentences that express variations in sentiment. Rating a business higher than the average rating or having too many sentences would result in a not recommended review. Users that have a larger number of friends, reviews, and photos submitted also increases their likelihood to be recommended. Recommended reviews also are based on recent submission, higher text readability, and less stop words. Reviews are less recommended if many reviews already exist per business and if the review is edited.

9.2 Insignificant Features Interpretation

The reduced model's insignificant features show that the two-stage sampling procedure had equal probability in scraping restaurants with randomized number of words in restaurant name, number of words in restaurant address, and Yelp listing order [33]. Yelp is not necessarily a platform for extreme comments, yet their filtering algorithm does not filter text based on identity hate score, insult score, threat score, obscene score, severe toxic score, and toxic score. Deceptive score has a low magnitude for a review to be recommended, which should be opposite for removing deceptive reviews. Since removed reviews are inaccessible, text classifiers are still useful for removing comments that violate the terms of use [40].

10 Ethics

Yelp's role in helping society make better informed decisions can be met with unease with their undisclosed filtering algorithm for recommending reviews [23]. This study brings a clearer understanding of the significant features that influence Yelp's filtering algorithm. Greater transparency and a guideline for writing recommended reviews will help Yelp gain more users on its platform submitting and receiving advice [13].

10.1 Yelp's Role

Yelp serves as an online platform for users to solicit information and advice from the general public. When people solicit information from friends and family, there is a chance for that information to be useful, important, misguided, or wrong. Yelp is a reflection of society because all of its reviews are crowd sourced [54]. Yelp pools its information towards the general consensus, which makes its information less likely to be wrong [54]. Yet, Yelp realizes that some of the information collected is not useful or irrelevant in terms of helping their user base and filters reviews as recommended and not recommended [6]. Yelp strives to succinctly highlight the useful information while filtering out deceitful and wrong information [1].

Yelp serves to collect, organize, and abridge information so the end users can make their decisions [1]. The ultimate decision is ultimately dependent on the end user, including the decision to use Yelp to obtain information and advice. Yelp strives to filter for good information because informed decisions lead to better experiences by the users. Free speech is not censored because not recommended reviews are still accessible [23]. Although business revenue is affected by information on Yelp, the filtered reviews and average rating only serves as a justified reflection of the collective experience [3][54]. Yelp is not at fault because the filtering mechanism has no malicious intent but to abridge information it gathers from society based on the motive to help users make better informed decisions [23][54].

11 Conclusions

The features important for filtering reviews are logical towards Yelp's efforts to promote quality and reliable information to help consumers gain insight and make decisions [1]. Yelp serves to collect, organize, and abridge information so the end users can make their decisions. The ultimate decision is ultimately dependent on the end user, including the decision to use Yelp to obtain information and advice.

Generic 5 star responses are prevented with Yelp's filtering algorithm. Advice with a myriad of sentimental information while rating critically is what Yelp wants to promote in order for people to tell the truth rather than having inflated reviews and ratings. Quality of text is promoted by reviews with higher readability and less stop words. Reliability of content is promoted by recent reviews from users with more activity and submitted data. Insight is gained by the variation of sentimental context of the collective experience. Ultimately, consumers' decisions are based on personal discretion and Yelp's filtering algorithm only help to create more informed decisions. The filtered reviews and average rating only serves as a justified reflection of the collective experience [54].

12 Future Study

A future study is to create a different text classifier for every sampled business page to further analyze Yelp's filtering algorithm. Matching a new review with the existing word bank of submitted reviews per page reduces redundant information. Filtering out reviews with redundant messages also reduces the business owners' incentive of telling customers key words of what to write as reviews and for the recommended reviews to be genuine. Creating a text classifier for every sampled user page also filters out users that write redundant reviews that are less genuine for every business page. We postulate that every business and user page on Yelp operates on a different word bank of submitted reviews, which is why creating an unified logistic regression model that satisfies Yelp's overall filtering algorithm less accurate yet still meaningful for finding overall significant features. 12 Future Study

A future study is to create a different text classifier for every sampled business page to further analyze Yelp's filtering algorithm. Matching a new review with the existing word bank of submitted reviews per page reduces redundant information. Filtering out reviews with redundant messages also reduces the business owners' incentive of telling customers key words of what to write as reviews and for the recommended reviews to be genuine. Creating a text classifier for every sampled user page also filters out users that write redundant reviews that are less genuine for every business page. We postulate that every business and user page on Yelp operates on a different word bank of submitted reviews, which is why creating an unified logistic regression model that satisfies Yelp's overall filtering algorithm less accurate yet still meaningful for finding overall significant features. (edited)

References

1. Yelp Support Center. (2018). Recommended Reviews | Support Center | Yelp. [online] Available at: https://www.yelp-support.com/Recommended_Reviews?l=en_US [Accessed 3 Mar. 2018].
2. Pimentel, B. (2013). Yelp steps up campaign against fake reviews. [online] MarketWatch. Available at: https://www.marketwatch.com/story/yelp-steps-up-campaign-against-fake-reviews-2013-09-18 [Accessed 19 May 2018].
3. Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.com. [online] Hbs.edu. Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=41233 [Accessed 19 May 2018].
4. Yelp Support Center. (2018). Posting Reviews & Tips | Support Center | Yelp. [online] Available at: https://www.yelp-support.com/Posting_Reviews?l=en_US [Accessed 3 Mar. 2018].
5. About Yelp. (2018). About Us | Yelp. [online] https://www.yelp.com/about [Accessed 3 Mar. 2018].
6. YouTube. (2018). Inside Yelp: How the Company Got Its Name. [online] Available at: https://youtu.be/Lnw3hqsGsMM [Accessed 19 May 2018].
7. Yelp. (2018). Yelp Elite Squad. [online] Available at: https://www.yelp.com/elite [Accessed 19 May 2018].
8. SEC Database. (2018). Yelp! Inc, Form 8-K, Current Report. [online] Available at: http://edgar.secdatabase.com/897/115752318000240/filing-main.htm [Accessed 19 May 2018].
9. Yelp. (2018). Factsheet | Yelp. [online] Available at: https://www.yelp.com/factsheet [Accessed 3 Mar. 2018].
10. Yelp. (2018). Our Gift to Business Owners: A Yelp App Just for You! - Yelp. [online] Available at: https://www.yelpblog.com/2014/12/our-gift-to-business-owners-a-yelp-app-just-for-you [Accessed 19 May 2018].
11. Yelp Support Center. (2018). Yelp for Business Owners | Support Center | Yelp. [online] Available at: https://www.yelp-support.com/Yelp_for_Business_Owners?l=en_US [Accessed 3 Mar. 2018].
12. Chafkin, M. (2010). You've Been Yelped. [online] Inc.com. Available at: https://www.inc.com/magazine/20100201/youve-been-yelped_pagen_3.html [Accessed 19 May 2018].
13. Eaton, K. (2010). Yelp Tweaks Its System for Transparency–and Lawsuit-Dodging. [online] Fast Company. Available at: https://www.fastcompany.com/1608067/yelp-tweaks-its-system-transparency-and-lawsuit-dodging [Accessed 19 May 2018].
14. Chang, A. (2010). Yelp makes two major changes in the way reviews are posted. [online] LA Times. Available at: http://articles.latimes.com/2010/apr/06/business/la-fi-yelp7-2010apr07 [Accessed 19 May 2018].
15. Banks, S. (2013). Turning a critical eye on Yelp. [online] LA Times. Available at: http://articles.latimes.com/2013/apr/20/local/la-me-banks-yelp-20130420 [Accessed 19 May 2018].
16. Hansell, S. (2008). Why Yelp Works. [online] Bits Blog. Available at: https://bits.blogs.nytimes.com/2008/05/12/why-yelp-works/ [Accessed 19 May 2018].
17. Yelp. (2018). Yelp Does Not Extort Local Businesses or Manipulate Ratings.. [online] Available at: https://www.yelp.com/extortion [Accessed 19 May 2018].
18. Post, D. (2015). Exposing Anonymous Yelp Reviewers. [online] Washington Post. Available at: https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/04/23/exposing-anonymous-yelp-reviewers/?noredirect=on&utm_term=.0bb9bc99d0c1 [Accessed 19 May 2018].

19. Loten, A. (2014). Yelp Regularly Gets Subpoenas About Users. [online] Wallstreet Journal. Available at: https://www.wsj.com/articles/yelp-regularly-receives-subpoenas-about-users-1396467536 [Accessed 19 May 2018].
20. Zetter, K. (2010). Yelp Accused of Extortion. [online] Wired. Available at: https://www.wired.com/2010/02/yelp-sued-for-alleged-extortion/ [Accessed 19 May 2018].
21. Sherman, E. (2014). Yelp battles to keep local reviewers anonymous. [online] CBS News. Available at: https://www.cbsnews.com/news/yelp-battles-to-keep-local-reviewers-anonymous [Accessed 19 May 2018].
22. Luca, M. and Zervas, G. (2015). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. [online] Harvard Business School. Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=45151 [Accessed 19 May 2018].
23. O'Brien, S. (2015). Yelp: You can trust our reviews. [online] CNN Money. Available at: http://money.cnn.com/2015/01/07/technology/ftc-yelp-reviews [Accessed 19 May 2018].
24. Egelko, B. (2014). Yelp can manipulate ratings, court rules. [online] SFGate. Available at: https://www.sfgate.com/news/article/Yelp-can-give-paying-clients-better-ratings-5731200.php [Accessed 19 May 2018].
25. CBS Local San Francisco. (2014). Court Sides With San Francisco-Based Yelp In Lawsuit From Small Business Owners. [online] Available at: http://sanfrancisco.cbslocal.com/2014/09/04/court-sides-with-san-francisco-based-yelp-in-lawsuit-from-small-business-owners-9th-circuit-court-of-appeals-online-reviews [Accessed 19 May 2018].
26. Roberts, D. (2013). Yelp's Fake Review Problem. [online] Fortune. Available at: http://fortune.com/2013/09/26/yelps-fake-review-problem [Accessed 19 May 2018].
27. Streitfeld, D. (2012). Yelp Tries to Halt Deceptive Reviews. [online] NYTimes. Available at: https://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html [Accessed 19 May 2018].
28. Cotter, S. (2017). Quincy Center Jeweler Wins Suit Over Yelp Review. [online] Patriot Ledger. Available at: http://www.patriotledger.com/news/20170405/quincy-center-jeweler-wins-suit-over-yelp-review [Accessed 19 May 2018].
29. McKeever, A. (2015). Why Yelp Emerged as a Site for Social Protest. [online] Eater. Available at: https://www.eater.com/2015/5/19/8588185/yelp-social-protest-trolling-memories-pizza [Accessed 19 May 2018].
30. Donnelly, G. (2018). Racist Rant by New York Lawyer Tanks His Law Firm's Yelp Rating. [online] Fortune. Available at: http://fortune.com/2018/05/17/racist-lawyer-yelp [Accessed 19 May 2018].
31. Yelp. (2018). More Places to Yelp. [online] Available at: https://www.yelp.com/locations [Accessed 19 May 2018].
32. Selenium. (2018). Selenium IDE. [online] Available at: https://www.seleniumhq.org/projects/ide/ [Accessed 19 May 2018].
33. Blair, E. and Blair, J. (2015). Applied Survey Sampling. Los Angeles: Sage Publications, Inc.
34. Mukherjee, S., Dutta, S. and Weikum, G. (2017). Credible Review Detection with Limited Information using Consistency Analysis. [PDF] Cornell University Library, arXiv:1705.02668. Ithaca, NY.
35. Mukherjee, A., Venkataraman, V., Liu, B. and Glance, N. (2013). What Yelp Fake Review Filter Might Be Doing?. In: Seventh International AAAI Conference on Weblogs and Social Media. [online] Association for the Advancement of Artificial Intelligence. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006 [Accessed 3 Mar. 2018].
36. Yelp. (2018). Yelp Dataset JSON. [online] Available at: https://www.yelp.com/dataset/documentation/json [Accessed 3 Mar. 2018].

37. Youngblade, R. (2013). Compliments: They're free. Give them! [online] Available at: https://www.yelpblog.com/2013/03/compliments-theyre-free-give-them [Accessed 3 Mar. 2018].
38. Google Maps Platform. (2018). Google Maps Platform - Geo-location APIs | Google Maps Platform | Google Cloud [online] Available at: https://cloud.google.com/maps-platform [Accessed 3 Mar. 2018].
39. Ott, M. (2018). Deceptive Opinion Spam Corpus v1.4. [online] Myle Ott. Available at: http://myleott.com/op-spam.html [Accessed 3 Mar. 2018].
40. Kaggle. (2018). Toxic Comment Classification Challenge. [online] Kaggle. Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge [Accessed 7 Jul. 2018].
41. The Stanford NLP Group. (2018). Software - The Stanford Natural Language Processing Group. [online] Available at: https://nlp.stanford.edu/software/ [Accessed 19 May 2018].
42. Brownlee, J. (2018). A Gentle Introduction to the Bag-of-Words Model. [online] Available at: https://machinelearningmastery.com/gentle-introduction-bag-words-model/ [Accessed 7 Jul. 2018].
43. Deshpande, M. (2018). Text Classification Tutorial with Naive Bayes. [online] Available at: https://pythonmachinelearning.pro/text-classification-tutorial-with-naive-bayes/ [Accessed 7 Jul. 2018].
44. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. [online] Available at: https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf [Accessed 7 Jul. 2018].
45. Readability Formulas. (2018). The Automated Readability Index (ARI). [online] Available at: http://www.readabilityformulas.com/automated-readability-index.php [Accessed 7 Jul. 2018].
46. Readability Formulas. (2018). The Flesch Grade Level Readability Formula. [online] Available at: http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php [Accessed 7 Jul. 2018].
47. Yee, J. (2018). Google's Dictionary API (Unofficial) in Python. [online] Available at: http://www.lleess.com/2013/03/googles-unofficial-dictionary-api-in.html [Accessed 7 Jul. 2018].
48. Ott, M., Choi, Y., Cardie, C. and Hancock, J. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. [online] Association for Computational Linguistics. Available at: http://myleott.com/op_spamACL2011.pdf [Accessed 3 Mar. 2018].
49. Conversation AI. (2018). Conversation AI. [online] Available at: https://conversationai.github.io [Accessed 7 Jul. 2018].
50. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. [online] Available at: https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf [Accessed 7 Jul. 2018].
51. A.I. Wiki. (2018). Recursive Neural Tensor Network. [online] Available at: https://skymind.ai/wiki/recursive-neural-tensor-network [Accessed 7 Jul. 2018].
52. Lane, D. (2018) Values of the Pearson Correlation. [online] Available at: http://onlinestatbook.com/2/describing_bivariate_data/pearson.html [Accessed 7 Jul. 2018].
53. McDonald, J. (2014). Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.
54. CBS This Morning. (2016). Yelp CEO on site's popularity and pitfalls. [online] Available at: https://www.youtube.com/watch?v=1mlzdXwbtZo [Accessed 7 Jul. 2018].

Appendix

Table A. Sampling design of city clusters with restaurant stratification methods are scraped from the list of featured Yelp cities [31].

	City Cluster	Tot. Rest.	Strat.	Rec.	N. Rec.	Est. Rec.	Est. N. Rec.
1	Phoenix, AZ	2200	5	1888	420	830,720	184,800
2	Scottsdale, AZ	792	4	1006	140	199,188	27,720
3	Tempe, AZ	4685	5	1577	178	1,477,649	166,786
4	Tucson, AZ	1510	5	667	112	201,434	33,824
5	Alameda, CA	1831	5	785	72	287,467	26,366
6	Albany, CA	247	2	72	7	8,892	865
7	Alhambra, CA	3624	5	2072	252	1,501,786	182,650
8	Anaheim, CA	4528	5	2720	457	2,463,232	413,859
9	Belmont, CA	438	3	626	55	91,396	8,030
10	Berkeley, CA	2043	5	4669	532	1,907,753	217,375
11	Beverly Hills, CA	5510	5	2328	334	2,565,456	368,068
12	Big Sur, CA	11	1	294	25	3,234	275
13	Burbank, CA	2978	5	4056	516	2,415,754	307,330
14	Concord, CA	1390	5	965	118	268,270	32,804
15	Costa Mesa, CA	2934	5	1854	242	1,087,927	142,006
16	Culver City, CA	5060	5	4936	460	4,995,232	465,520
17	Cupertino, CA	1653	5	671	67	221,833	22,150
18	Daly City, CA	2238	5	1285	192	575,166	85,939
19	Davis, CA	194	1	946	132	183,524	25,608
20	Dublin, CA	555	3	259	12	47,915	2,220
21	Emeryville, CA	1439	5	1915	186	551,137	53,531
22	Foster City, CA	319	2	1264	159	201,608	25,361
23	Fremont, CA	3308	5	1526	265	1,009,602	175,324
24	Glendale, CA	11942	5	2732	313	6,525,109	747,569
25	Hayward, CA	3596	5	802	62	576,798	44,590
26	Healdsburg, CA	112	1	150	18	16,800	2,016
27	Huntington Beach, CA	7567	5	2076	151	3,141,818	228,523
28	Irvine, CA	6394	5	4612	477	5,897,826	609,988
30	Livermore, CA	786	4	511	49	100,412	9,629
31	Long Beach, CA	8395	5	4890	516	8,210,310	866,364
32	Los Altos, CA	1295	5	2190	290	567,210	75,110
33	Los Angeles, CA	9494	5	8086	1260	15,353,697	2,392,488
34	Los Gatos, CA	1060	5	514	86	108,968	18,232
35	Marina del Rey, CA	1086	5	1613	171	350,344	37,141
36	Menlo Park, CA	1221	5	3906	380	953,845	92,796
37	Mill Valley, CA	499	3	510	74	84,830	12,309
38	Millbrae, CA	468	3	2360	226	368,160	35,256
39	Milpitas, CA	1460	5	791	55	230,972	16,060
40	Monterey, CA	519	3	4393	589	759,989	101,897
41	Mountain View, CA	4700	5	783	44	736,020	41,360
42	Napa, CA	480	3	1208	135	193,280	21,600
43	Newark, CA	807	5	470	23	75,858	3,712
44	Newport Beach, CA	5421	5	5706	1339	6,186,445	1,451,744
45	Oakland, CA	7906	5	2717	268	4,296,120	423,762
46	Orange County, CA	4603	5	1215	139	1,118,529	127,963
47	Palo Alto, CA	1155	5	1229	108	283,899	24,948
49	Pasadena, CA	2275	5	1286	161	585,130	73,255
50	Pleasanton, CA	438	3	1033	118	150,818	17,228
51	Redondo Beach, CA	2230	5	2649	310	1,181,454	138,260
52	Redwood City, CA	3030	5	1712	120	1,037,472	72,720
53	Sacramento, CA	2191	5	1502	131	658,176	57,404

54	San Bruno, CA	586	3	185	7	36,137	1,367
55	San Carlos, CA	517	3	220	21	37,913	3,619
56	San Diego, CA	3887	5	13001	2916	10,106,977	2,266,898
57	San Francisco, CA	4873	5	7644	944	7,449,842	920,022
58	San Jose, CA	3253	5	1017	132	661,660	85,879
59	San Leandro, CA	1294	5	700	84	181,160	21,739
60	San Mateo, CA	1171	5	455	30	106,561	7,026
61	San Rafael, CA	1005	5	4432	977	890,832	196,377
62	Santa Barbara, CA	656	4	3109	581	509,876	95,284
63	Santa Clara, CA	2727	5	1372	167	748,289	91,082
64	Santa Cruz, CA	463	3	1541	265	237,828	40,898
65	Santa Monica, CA	1534	5	3566	433	1,094,049	132,844
66	Santa Rosa, CA	820	5	394	78	64,616	12,792
67	Sausalito, CA	144	1	1958	214	281,952	30,816
68	Sonoma, CA	119	1	2893	286	344,267	34,034
69	South Lake Tahoe, CA	237	2	3158	369	374,223	43,727
70	Stockton, CA	712	4	396	62	70,488	11,036
71	Studio City, CA	5165	5	552	114	570,216	117,762
72	Sunnyvale, CA	1820	5	1017	84	370,188	30,576
73	Torrance, CA	7067	5	3110	337	4,395,674	476,316
74	Union City, CA	2964	5	2557	243	1,515,790	144,050
75	Venice, CA	3311	5	1389	190	919,796	125,818
76	Walnut Creek, CA	2094	5	979	178	410,005	74,546
77	West Hollywood, CA	6204	5	4441	512	5,510,393	635,290
78	West Los Angeles, CA	1693	5	1666	157	564,108	53,160
79	Westwood, CA	5	1	25	6	125	30
80	Yountville, CA	29	1	1018	90	29,522	2,610
81	Boulder, CO	1072	5	1393	259	298,659	55,530
82	Denver, CO	3191	5	2038	367	1,300,652	234,219
83	Hartford, CT	910	5	1148	375	208,936	68,250
84	New Haven, CT	965	5	134	12	25,862	2,316
85	Washington, DC, DC	8095	5	11213	2176	18,153,847	3,522,944
86	Fort Lauderdale, FL	4850	5	3050	549	2,958,500	532,530
87	Gainesville, FL	552	3	71	25	13,064	4,600
88	Miami, FL	4108	5	3495	1390	2,871,492	1,142,024
89	Miami Beach, FL	6172	5	4302	708	5,310,389	873,955
90	Orlando, FL	2494	5	1578	419	787,106	208,997
91	Tampa, FL	2158	5	537	86	231,769	37,118
92	Atlanta, GA	3398	5	2944	530	2,000,742	360,188
93	Savannah, GA	929	5	401	56	74,506	10,405
94	Honolulu, HI	3031	5	7337	735	4,447,689	445,557
95	Lahaina, HI	355	2	4333	355	769,108	63,013
96	Iowa City, IA	310	2	338	82	52,390	12,710
97	Boise, ID	994	5	1143	435	227,228	86,478
98	Chicago, IL	6942	5	7671	1733	10,650,416	2,406,097
99	Evanston, IL	1289	5	779	184	200,826	47,435
100	Naperville, IL	2173	5	486	84	211,216	36,506
101	Schaumburg, IL	2548	5	582	94	296,587	47,902
102	Skokie, IL	2143	5	207	33	88,720	14,144
103	Bloomington, IN	318	2	48	18	7,632	2,862
104	Indianapolis, IN	1576	5	1530	223	482,256	70,290
105	Louisville, KY	1635	5	969	173	316,863	56,571
106	New Orleans, LA	2765	5	2953	305	1,633,009	168,665
108	Boston, MA	6078	5	1656	208	2,013,034	252,845
110	Brookline, MA	4188	5	2011	209	1,684,414	175,058
112	Somerville, MA	4125	5	1223	211	1,008,975	174,075
113	Baltimore, MD	4148	5	653	103	541,729	85,449
114	Ann Arbor, MI	781	4	187	35	36,512	6,834
115	Detroit, MI	2294	5	1470	486	674,436	222,977

116	Minneapolis, MN	2251	5	1643	543	739,679	244,459
117	Saint Paul, MN	1698	5	543	104	184,403	35,318
118	Kansas City, MO	1257	5	324	46	81,454	11,564
119	Saint Louis, MO	2000	5	1496	238	598,400	95,200
120	Charlotte, NC	1709	5	620	50	211,916	17,090
121	Durham, NC	836	5	213	17	35,614	2,842
122	Raleigh, NC	1297	5	171	49	44,357	12,711
123	Newark, NJ	3010	5	441	82	265,482	49,364
124	Princeton, NJ	1492	5	161	62	48,042	18,501
125	Albuquerque, NM	1668	5	899	130	299,906	43,368
126	Santa Fe, NM	435	3	1045	206	151,525	29,870
127	Las Vegas, NV	3893	5	4789	676	3,728,715	526,334
128	Reno, NV	1014	5	476	113	96,533	22,916
129	Brooklyn, NY	13063	5	463	41	1,209,634	107,117
131	New York, NY	24399	5	9466	1811	46,192,187	8,837,318
132	Flushing, NY	19167	5	957	116	3,668,564	444,674
133	Cincinnati, OH	1646	5	916	121	301,547	39,833
134	Cleveland, OH	1839	5	602	57	221,416	20,965
135	Columbus, OH	2182	5	526	80	229,546	34,912
136	Portland, OR	3717	5	5865	1602	4,360,041	1,190,927
137	Salem, OR	652	4	1082	331	176,366	53,953
138	Philadelphia, PA	5604	5	2248	253	2,519,558	283,562
139	Pittsburgh, PA	2215	5	2178	465	964,854	205,995
140	Providence, RI	1415	5	407	74	115,181	20,942
141	Charleston, SC	1431	5	2174	375	622,199	107,325
142	Memphis, TN	954	5	1889	537	360,421	102,460
143	Nashville, TN	1863	5	2272	320	846,547	119,232
144	Austin, TX	2584	5	4834	783	2,498,211	404,654
145	Dallas, TX	3249	5	2496	378	1,621,901	245,624
146	Houston, TX	3137	5	1133	191	710,844	119,833
147	San Antonio, TX	2623	5	610	50	320,006	26,230
148	Salt Lake City, UT	1726	5	2370	528	818,124	182,266
149	Alexandria, VA	6977	5	3848	697	5,369,499	972,594
150	Arlington, VA	4731	5	2493	250	2,358,877	236,550
151	Richmond, VA	1604	5	1397	260	448,158	83,408
152	Burlington, VT	357	2	1168	204	208,488	36,414
153	Bellevue, WA	4430	5	764	92	676,904	81,512
154	Redmond, WA	2514	5	1378	140	692,858	70,392
155	Seattle, WA	3568	5	1265	167	902,704	119,171
156	Madison, WI	1051	5	1536	378	322,867	79,456
157	Milwaukee, WI	1745	5	568	91	198,232	31,759
		Total	**676**	**300428**	**47389**	**265,329,274**	**43,165,092**