SMU Data Science Review

Volume 1

2018

Yelp's Review Filtering Algorithm

Yao Yao
Southern Methodist University, yao.y89@gmail.com

Ivelin Angelov
Southern Methodist University, iangelov@smu.edu

Jack Rasmus-Vorrath
Southern Methodist University, jrasmusvorrath@smu.edu

Mooyoung Lee
Southern Methodist University, mooyoungl@smu.edu

Daniel W. Engels
Southern Methodist University, dwe@smu.edu

Recommended Citation

Yao, Yao; Angelov, Ivelin; Rasmus-Vorrath, Jack; Lee, Mooyoung; and Engels, Daniel W. (2018) "Yelp's Review Filtering Algorithm," *SMU Data Science Review*: Vol. 1 : No. 3 , Article 3.
Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/3

Yelp's Review Filtering Algorithm

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

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

Abstract. In this paper, we present an analysis of features influencing Yelp's proprietary review filtering algorithm. Classifying or misclassifying reviews as *recommended* or *non-recommended* affects average ratings, consumer decisions, and ultimately, business revenue. Our analysis involves systematically sampling and scraping Yelp restaurant reviews. Features are extracted from review metadata and engineered from metrics and scores generated using text classifiers and sentiment analysis. The coefficients of a multivariate logistic regression model were interpreted as quantifications of the relative importance of features in classifying reviews as recommended or non-recommended. The model classified review recommendations with an accuracy of 78%. We found that reviews were most likely to be recommended when conveying an overall positive message written in a few moderately complex sentences expressing substantive detail with an informative range of varied sentiment. Other factors relating to patterns and frequency of platform use also bear strongly on review recommendations. Though not without important ethical implications, the findings are logically consistent with Yelp's efforts to facilitate, inform, and empower consumer decisions.

1 Introduction

Filtering user reviews has inherent biases. Yelp is a third-party online platform where users can search for, find, and review businesses. People seeking advice or businesses seeking feedback will find crowd-sourced 1-to-5 star ratings paired with user-written reviews. Contributed reviews vary in detail and opinion, and some of them are deceptive or disruptive. Using proprietary algorithms, Yelp classifies reviews as either *recommended* or *non-recommended*. Review recommendations are designed to improve Yelp's service of providing quality, reliable information to help consumers gain insight and make decisions[1]. Data on user characteristics and activity are gathered to filter out reviews and flag fraudulent accounts whose submissions appear deceptive, disruptive, or otherwise in violation of Yelp's terms of service [1].

The review filtering process is an essential part of Yelp's power to influence consumer decisions and impact business revenue. Non-recommended reviews are still accessible to the public but are not calculated towards the business's average rating,

[1] See https://www.yelp-support.com/Recommended_Reviews for information about Yelp's recommended reviews

and good Yelp ratings are good for business. Estimates indicate that each star increase in average rating corresponds to a revenue increase of between 5% and 9% [2]. However, Yelp's filtering algorithm can misclassify credible reviews as non-recommended and non-credible reviews as recommended. Adjustments to the filtering algorithm will change which reviews are recommended, thus affecting a business's average rating and, ultimately, that business's revenue[2].

In our analysis, we construct a multivariate logistic regression model to investigate the Yelp filtering system and to identify which features have the most influence on the classification of reviews as recommended or non-recommended. We systematically sample and scrape Yelp's restaurant reviews. Features are created from review metadata and metrics generated using textual classifiers and sentiment analysis. Feature values are scaled from 0 to 1, and sampling adjustments are made to account for the unbalanced number of recommended and non-recommended reviews. As measures of feature importance, the coefficients of the multivariate logistic regression model are interpreted as quantifications of the criteria according to which the Yelp review filtering system makes its recommendations.

Our model classifies reviews as either recommended or non-recommended. It agrees with Yelp’s classification 77.61% of the time and has an F1-Score of 76.79%. The coefficients of the model features suggest that recommended reviews are more likely to consist of an overall positive message written in a few moderately complex sentences expressing substantive detail with an informative range of varied sentiment. Rating a business much higher than the average rating is more likely to result in one’s review being flagged as non-recommended. Users who have more review submissions, a profile photo, or a larger number of friends on the platform are more likely to have their reviews recommended. Review recommendations are also influenced by a user having made recent submissions and by the level of sentence complexity. Furthermore, reviews are less likely to be recommended when posted on pages which already have a great number of them, as are reviews which have been edited by their authors after submission.

The features identified as important for filtering reviews are logically consistent with Yelp's efforts to provide quality, reliable information to consumers. Sentence complexity level is an indicator of reviewer investment in delivering textual quality. Recommending reviews from identifiable users with recent and more frequent activity promotes reliability of content. Variations in sentiment provide readers with a range of experiences from which to draw their own conclusions. Ultimately, use of the platform is made according to personal discretion, and Yelp's filtering algorithm serves to facilitate, inform, and empower those decisions.

The remainder of this paper is organized as follows: In Section 2, we provide background on Yelp, how it is used, the demographics of its users, and reasons for filtering reviews. Section 3 elaborates on Yelp's influence on businesses. Section 4 discusses the sampling procedure used to select the data input to our classification model. In section 5, we present our exploration of the attributes of the data. We discuss the workflow of feature creation, model selection, and analysis in Section 6. In Section 7, we present the Natural Language Processing (NLP) techniques used on the review text to generate features for our classifier. Section 8 contains our analysis

[2] See https://www.yelp-support.com/Posting_Reviews for Yelp's review posting tips.

of the mean differences and correlation values associated with the distinction between recommended and non-recommended reviews. In Section 9, we evaluate the features influencing Yelp's review filtering algorithm according to the signs and magnitudes of the model coefficients. We present guidelines for writing recommended reviews and make note of the insignificant features of our model in Section 10. In Section 11, we describe the ethics of Yelp's role in helping users make better informed decisions by filtering reviews. We draw the relevant conclusions in Section 12.

2 Yelp

Background information includes what motivates Yelp's development, how their business model is structured, as well as relevant financial statistics. We introduce how to use Yelp, the demographics of reviewers, and the star rating system. We also introduce how average ratings are calculated and Yelp's distinction between recommended and non-recommended reviews.

2.1 Introduction to Yelp

Headquartered in San Francisco, Yelp was founded in October 2004 by former PayPal employees Russel Simmons and Jeremy Stoppelman[3]. Yelp was designed to function as an online directory where people can solicit help and advice on finding the best local businesses [3].

Yelp strives to be a platform on which small and large businesses alike can be publicly ranked and evaluated on an even playing field. Many businesses contend that a conflict of interest results from the fact that Yelp's main source of income is advertising sales, suggesting that businesses could pay their way into showing up on more search results and on the pages of their competitors [4]. Yelp has denied any wrongdoing, pointing out that the review filtering algorithm applies to everyone in the same way. From its perspective, ads are a way for the website to make revenue while providing a free service accessible to everyone[4].

According to Yelp's 2017 financial report, net revenue grew 19% in 2016 to $846.8 million, of which advertising revenue constitutes $771.6 million [5]. The other $75.2 million includes revenue from other provided services such as food delivery, a waitlist app, and sponsored Wi-Fi [5]. Since 2016, paid advertising accounts grew 21% to 163,000 [5]; the average paid advertising account spends $4,730 a year.

2.2 How to Use Yelp

As a third-party online platform, Yelp enables users to search for, find, and voluntarily review businesses. Once registered, users can update their location, profile picture, and interests. As depicted in Figure 1, user reviews of businesses consist of a

[3] See https://www.yelp.com/about for information about Yelp.

[4] See https://www.yelp.com/extortion for Yelp's policies on advertising

rating on a scale of one to five-stars, posted pictures, and written feedback in the form of short summary titles and long detailed reviews. Users can receive nominations to Yelp's Elite Squad, members of which receive benefits for frequently writing quality reviews and visiting new establishments[5]. For nominations, users are encouraged to provide their real names and post profile pictures. Online Yelp interactions include networking with other local reviewers, as well as complimenting others' reviews.

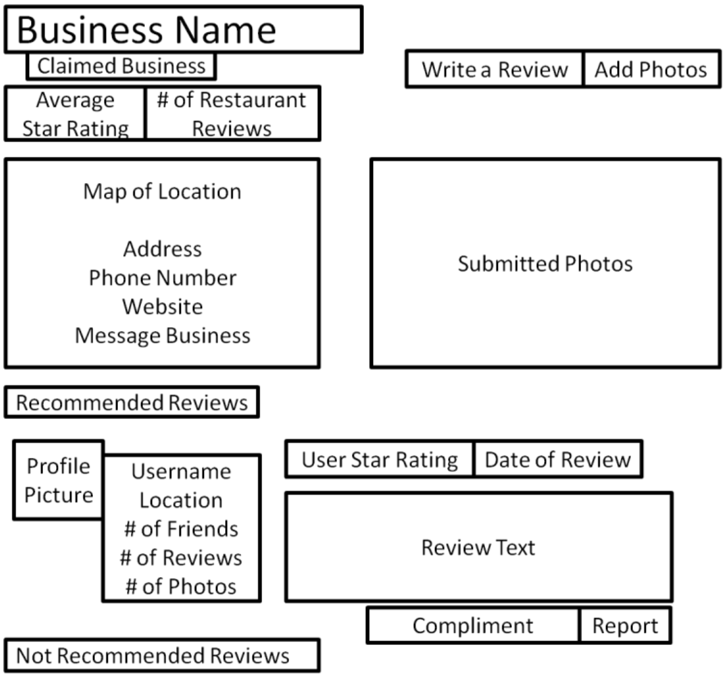

Figure 1. Layout of the Yelp website for a given business page, to which users for users contribute by posting star ratings, pictures, and review text

2.3 Demographics of Reviewers

From its inception in 2004 until March 2018, Yelp accumulated over 155 million reviews, of which 72% are classified as recommended and 21% are classified as non-recommended. The remaining 7% of reviews have been removed for breaching Yelp's terms of service[6]. As of March 2018, Yelp's metrics indicate that on a per monthly basis the Yelp app averages 30 million unique visitors, the mobile website averages 70 million unique visitors, and the desktop website averages 74 million unique visitors. 79% of searches and 65% of reviews are on mobile devices. The rating distribution of all reviews is depicted in Figure 2, which shows that 48% are five-star,

[5] See https://www.yelp.com/elite for information about Yelp's Elite Squad.

[6] See https://www.yelp.com/factsheet for Yelp's factsheet for more detailed graphics.

20% are four-star, 9% are three-star, 7% are two-star, and 16% are one-star ratings. The top-3 reviewed business categories are shopping at 21%, restaurants at 17%, and home and local services at 14%. The top-represented US demographics among Yelp reviewers are 35-54 year-olds (37%), college graduates (59%), and people having an annual income greater than $100K (49.6%).

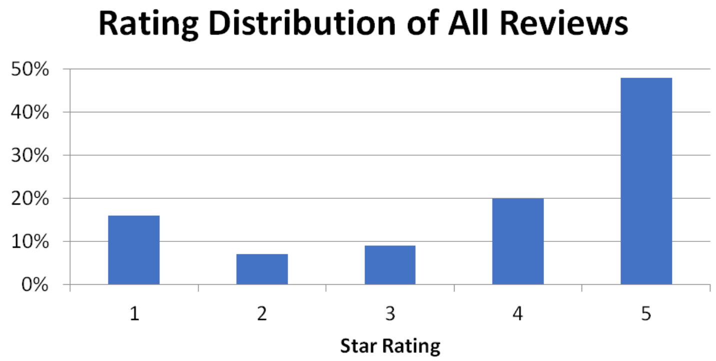

Figure 2. Distribution of star ratings across the total population of recommended and non-recommended reviews on Yelp

2.4 Yelp's Recommended Reviews

User reviews vary in detail and opinion. Importantly, some reviews are deceptive, i.e., purposefully misleading or written to artificially inflate or deflate a business's rating. Others are disruptive, i.e., containing unrelated content or unintelligible language. Yelp classifies reviews as either recommended or non-recommended by using proprietary algorithms. In this way, reviews are filtered to facilitate Yelp's service of providing quality, reliable information to help consumers gain insight and make decisions. Information on user characteristics and activity is gathered to flag fraudulent accounts submitting reviews which appear deceptive, disruptive, or otherwise in violation of Yelp's terms of service [1]. Users are encouraged to report such violations if they are found. Reviews classified as non-recommended are still publicly accessible but are not calculated towards the business's average rating. As mentioned above, estimates indicate that each star increase in average rating can correspond to a 5% to 9% increase in a business's revenue [2].

Yelp does not disclose the details of the review filtering system to discourage intentional manipulation of the ratings. Evaluation criteria for filtering a review includes whether it is deemed authentic and based on first-hand experiences [6]. Fraudulent accounts are suspended for deceptive or disruptive behavior [7]. It is noteworthy that Yelp has also become a platform of protest or support for businesses' political views. However, such politically motivated reviews are manually or algorithmically removed [8][9].

In summary, Yelp's review recommendation algorithm serves to filter out reviews which are either designed to mislead users, solicited of third-party sources, written by those affiliated with business owners, motivated by political interests, or filled with unrelated content or unintelligible language.

3 Yelp's Impact on Businesses

Yelp has the power to influence consumer decisions and impact business revenue based on the outcome of the review filtering process. However, Yelp's filtering algorithm can misclassify credible reviews as non-recommended and non-credible reviews as recommended. Adjustments to the filtering algorithm will change which reviews are recommended, thus affecting a business's average rating and, ultimately, that business's revenue.

Businesses can claim their pages on Yelp, which allows them to add menu items, offer discounts, directly respond to reviews publicly 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 on Yelp. For their sponsored advertisement to show at the top of search results or on their competitors' Yelp profiles, businesses are required to have at least an average rating of three-stars [11]. In this context it is worth noting that Harvard Business School found no significant correlation between advertising and better ratings and reviews on Yelp [12].

To keep businesses from tampering, exploiting, or otherwise unfairly intervening in the review process, Yelp takes preventative measures to protect reviewer data [13][14][15]. Significantly, Yelp receives an average of six subpoenas per month from businesses inquiring about their reviews, some of which have led to defamation cases [16][17]. In such cases, businesses have attempted to leverage disparagement clauses to sue or fine customers who post negative reviews [18]. Recent legal proceedings invoked a 2014 California State law protecting the user from business scrutiny. Yelp views such protection as essential to the site's integrity, ensuring that users may share reviews without fear of legal action [19].

For Yelp, preserving site integrity also entails examining review patterns left on different pages to discover businesses using illicit means to outperform competitors on the platform and acquire better reviews [6]. As part of a sting operation, Yelp found and filed lawsuits against 19 third-party websites for participating in writing fraudulent reviews [6]. In one case, it was revealed that a jewelry store in San Diego was willing to pay customers by the word for favorable reviews, in sums of up to $200 per review [7]. Another case resulted in a court ruling forcing a jewelry store in Massachusetts to pay $34,500 in damages for writing a negative review about a rival business [20].

Yelp itself has also been the target of legal action taken by businesses appearing on its platform [21]. However, extortion allegations have resulted in court rulings citing insufficient evidence to support the assertion that Yelp was manipulating review recommendations in exchange for ad revenue [22]. According to a 2014 ruling by the 9th circuit court of appeals, businesses paying for Yelp advertising does not meet the legal definition of extortion [23].

4 Yelp Dataset Collection

Yelp provides an open data challenge which invites the public to discover new insights from their data to benefit the platform as well as the businesses and consumers who use it[7]. However, the official dataset provided by Yelp does not include non-recommended reviews with which to conduct a study of their filtering algorithm. Moreover, promotional datasets of this kind may inherit undocumented biases distorting or failing to capture characteristics of the population of interest, and an external analysis applying careful sampling procedures allows for a more controlled observational study. At the same time, gathering millions of reviews across every business documented on Yelp is not feasible due to search limitations and ongoing changes in the ordering of search results.

Yelp's dynamic ordering of results creates duplicates and skipped observations when performing systematic scraping, i.e., the downloading of online information using a custom program. Scraping is made particularly difficult with respect to less frequently reviewed businesses in cities with a low adoption rate of Yelp's application. For some metropolitan areas, over 5,000 businesses exist, yet only the first 1,000 are available per searched city. In the interest of obtaining representative data, a two-stage cut-off non-probability sampling design is used. A Python-activated Selenium browser is used to programmatically scrape Yelp's recommended and non-recommended reviews[8].

4.1 Sampling Procedure

Yelp lists the various cities that have adopted it as a review platform[9]. When searching by city, Yelp lists businesses by category. Amongst businesses, restaurant pages were the most frequently reviewed across cities of every size. To facilitate statistical inference from a nation-wide population of reviewers, only restaurant data was gathered. Moreover, only English written reviews of restaurants located in US cities were included. The Python script and Selenium browser used in the scraping process are designed to mimic user searching behavior[8].

The two-stage cut-off non-probability sampling procedure applied to the data preserves certain attributes of the distribution to better represent the population [24]. In the first step, the data is collected from cities that Yelp has identified as having the highest rate of adopting its application. Figure 3 depicts the sampling procedure. Sampling from cities with a higher total number of restaurant reviews facilitates balancing the proportionately fewer number of non-recommended reviews before the analysis is performed. These high-adoption cities are discretized by number of restaurant reviews into five bins, to which a proportionate number of sampled restaurants is allocated. The highest bin receives five samples, and the lowest receives one. Within each bin, a random number generator is used to set a sampling interval

[7] See https://www.yelp.com/dataset/challenge for information about Yelp's dataset challenge.

[8] See https://www.seleniumhq.org/projects/ide for information about the Selenium browser.

[9] See https://www.yelp.com/locations for a list of cities that adopt Yelp. These cities are listed in Table A of the Appendix.

with which the specified number of restaurants are drawn from the total listed in that city. In the second step, reviews of the selected restaurants in these cities are randomly sampled from the maximum 1000 accessible to our web-scraping application. A down-sampling procedure is used on those which are selected to ensure an equal number of recommended and non-recommended reviews [24]. As random sampling of systematically scraped data may still introduce duplicate reviews, the data set also underwent a manual post-processing step to correct for these errors.

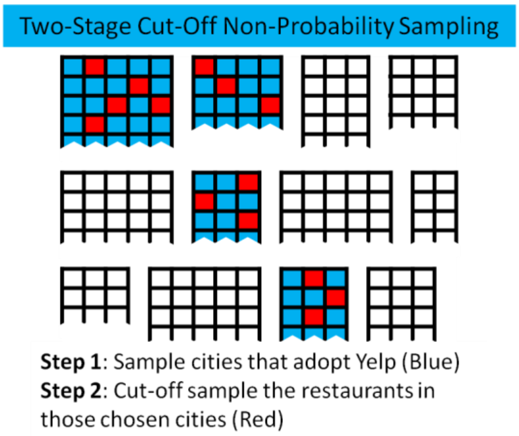
Two-Stage Cut-Off Non-Probability Sampling
Step 1: Sample cities that adopt Yelp (Blue)
Step 2: Cut-off sample the restaurants in those chosen cities (Red)

Figure 3. Two-stage cut-off non-probability sampling procedure.

A cut-off sampling method is used under the assumption that inferences drawn from popular restaurants with many reviews can be applied to more obscure, less frequently reviewed restaurants [24]. Although this assumption may result in over- or under-representing certain features of the population, Yelp's search limit requires that one use a cut-off sampling approach [24]. Restaurants are ranked according to average rating and how frequently they are reviewed, meaning more popular restaurants are listed first. Since the listing is limited to the first 1000 results, cut-off sampling is a practical way to make use of these readily available records.

4.2 Dataset Projection and Balancing

From 676 restaurants and 157 cities, the two-stage cut-off sampling procedure scraped 300,428 recommended and 47,389 not-recommended reviews[10]. After cleaning missing values and duplicates, the dataset contained 224,604 recommended reviews and 26,824 non-recommended reviews. The resulting data set thus consists of 89% recommended reviews and 11% non-recommended reviews. Reviews removed by Yelp for whatever reason were not accessible and therefore not represented in the sample.

[10] See https://github.com/post2web/capstone for our Yelp dataset and the code used to perform the analysis.

The resulting dataset is balanced to have equal observations in both categories so as not to distort the outcomes of classification. Balancing the dataset ensures that accuracy metrics are truly representative of classifier performance. Post-hoc balancing of the model is methodologically suspect, and still results in poor F1 and recall scores in classifying the less frequent non-recommended reviews. As the focus of this study concerns the relative importance of features influencing the classification of reviews, the modeling approach chosen prioritizes such interpretability. Toward this end, the analysis of the balanced dataset applies multivariate binomial logistic regression to the task of classifying recommended and non-recommended reviews.

5 Collected Yelp Dataset

Scraped data is labeled, merged, and combined. Compliments given to reviews by other users can only be posted to recommended reviews, and data features are only included in this study if they exist for both recommended and non-recommended categories**Error! Bookmark not defined.**. Such review metadata is therefore not scraped. As many of these features were first added in 2013 [25], nine years into Yelp's existence, they have no bearing on the filtering of older reviews. Other features are not accessible to the scraping procedure, including internal metadata such as page visitation information.

5.1 Data File

The scraped dataset contains both restaurant information and review data. The restaurant attributes are summarized in Table 1 and the review attributes are summarized in Table 2.

In Table 1, string values include restaurant name, address, city, and Yelp link. Float and integer values include average rating, number of reviews (of the restaurant), number of restaurants (of that name in a given city), restaurant ID, and restaurant listing order. In addition to internal links created by Yelp, the city in which a restaurant is situated is extraneous information, as the sampling procedure ensures equal numbers of recommended and non-recommended reviews in each city. Duplicates of the same restaurant at multiple addresses are filtered out. The number of reviews and the number of restaurants (of that name in a given city) are logarithmically transformed to facilitate multivariate model fitting.

Table 1. The Restaurants.csv file contains restaurant data, which are merged with reviews by Restaurant ID to create extra features[10].

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	Merge with Reviews.csv	0

In the Reviews dataset, string values include date, location, text, and username, as identified in Table 2. Integer values include number of friends, number of photos, user rating, restaurant ID, number of reviews by user, and the binary target variable indicating whether a review was recommended. A binary value is also used to indicate the presence of a profile picture. The month-day-year date format is transformed to the number of days after Yelp's inception that the review was published. The Date variable also includes a string value indicating whether the review was updated, which is converted to a binary value. As multiple users may have the same first names and last initial, the user name column is dropped.

Table 2. Reviews.csv contains full review text data, which is merged with restaurants by Restaurant ID[10].

Category	Data Type	Description	Example
Date	String	Date formatted M-D-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

5.2 Adding Features

Multivariate logistic regression requires quantification of all data passed to the classifier. Review text is formatted and cleaned of special characters before being converted into the number of sentences, the number of words, and the word count excluding common 'stop words', which contain no informative semantic content. The difference between user rating and the average rating of the restaurant is also quantified. The distance in miles between user and restaurant is obtained using the Google Maps API[11]. Number of sentences, number of words, word count excluding stop words, number of friends, number of photos, and number of reviews per user are all logarithmically transformed to facilitate model fitting. The recommended ratio feature captures recommended-to-total reviews per restaurant ID.

Table 3. Data features created by merging review with restaurant data. An asterisk (*) denotes data values 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 Stop-words*	Float	Word length of review text with no stop words	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

Analysis of the effects of a review filtration system on providers and consumers of goods and services can be extended to other domains, such as movies, music, shopping, and search results. Classifiers relying on user metadata, textual sentiment analysis, and other natural language processing techniques encounter similar challenges in analyzing the filtering process[12,13,14]. The broader implications of such analyses concern how review filtering systems work to the benefit or detriment of the providers and consumers who make use of them.

[11] See https://cloud.google.com/maps-platform for the Google geo-location API.

[12] See https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge for the text classification dataset.

[13] See http://myleott.com/op-spam.html for the spam opinion corpus dataset.

[14] See https://nlp.stanford.edu/software for information about Stanford's NLP software.

6 Multivariate Logistic Regression and Metrics

Our binary classification model uses scaled numerical features derived from metadata and textual characteristics of reviews. Multivariate logistic regression quantifies the log-odds of the probability of an event (i.e., recommended or non-recommended) as a linear combination of predictor variables input as features to the model. Coefficients of the multivariate logistic regression classifier are evaluated to determine which features have the most influence on Yelp's review filtering system.

6.1 Metrics of Binary Prediction

The results of any binary classification consist of true positives, true negatives, false positives, and false negatives. True positives and negatives accurately predict labels while false positives and negatives are misclassifications. In addition to accuracy (1), precision is used as a measure of model performance, quantifying how good the classifier is at only identifying recommended reviews as such (meaning, fewer false positives) (2). Metrics of model performance also include recall, which quantifies how good the classifier is at correctly identifying all the reviews in the 'recommended' category (meaning, fewer false negatives) (3). F1-Score (4) is also used as a weighted accuracy metric consisting of the harmonic mean of precision and recall.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Observations} \tag{1}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

$$F1\ Score = \frac{2 * True\ Positives}{2 * True\ Positives + False\ Positives + False\ Negatives} \tag{4}$$

6.2 Evaluating Feature Importance

The sign and magnitude of the model coefficients are interpreted to determine the relative importance of features with respect to the classification of reviews as recommended or non-recommended. Some features' contributions to the classifier are insignificant and therefore removed when paring down from a full model to a reduced one. Feature significance is determined by a p-value threshold of alpha = 0.05, below which the null hypothesis of no contribution to the model is deemed improbable. Both the full and reduced model are evaluated with respect to which features are the most significant and influential in classifying recommended vs. non-recommended reviews.

In providing insight into the review filtration system, the evaluation of feature importance provides guidelines on how to submit recommended reviews.

7 Text Processing of Restaurant Reviews

Features are extracted from the review text using natural language processing techniques, including sentiment analysis and a Bag-of-Words based Naïve Bayes text classifier. A Bag-of-Words approach processes word frequencies without respect to grammar, spelling, or word order [26]. Applying the Bag-of-Words approach, the Naïve Bayes method uses labeled text documents to classify unlabeled documents according to the probabilities of words occurring in documents of a particular class [27]. Sentiment analysis is used to identify the tonality of a sentence [28].

7.1 Readability and Spelling Model

Additional features are created using readability indexes, which measure the difficulty of understanding text. The total numbers of syllables, characters, words, and sentences are used to generate the readability index of review text (5). Age and grade-level readability are listed by Automated Readability Index (ARI) score in Table 4[15]. According to the Flesch–Kincaid Grade Level Formula[16], the total number of syllables is also extracted using the Google dictionary API in determining the grade-level readability of review text (6)[29]. The Google dictionary API was likewise used to find the percentage of words spelled correctly in the review text [29].

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

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

[15] See http://www.readabilityformulas.com/automated-readability-index.php

[16] See http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php

Table 4. The Automated Readability Index score is based on age group and grade-level [30].

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

7.2 Naïve Bayes Text Classifiers

A model feature encoding whether a review is deceptive or truthful is created using a Naïve Bayes text classifier. The Bag-of-Words approach applied by this classifier does not account for grammar and word order of the text [27]. The Naïve Bayes method is based on the Bayes Theorem, which describes the probability of an event in terms of prior knowledge of conditions, relating conditional and marginal probabilities. Table 5 shows how the word frequencies of a text document, i.e., a given restaurant review, are vectorized to calculate the probability of the document belonging to a certain class, i.e., deceptive or truthful. Probabilities of class belonging are calculated according to class differences in probabilities of word occurrence.

Table 5. Vectorizing the word frequency of a document and calculating the probability that the document 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

As depicted in Figure 4, the pre-trained Naïve Bayes text classifier uses the word vectors derived from new data to classify documents according to the word occurrence probabilities on which it was trained. To use the classical example of spam detection, , the conditional probability P(A|B) that a given text document, i.e. review (B), is spam, i.e., deceptive (A), is equal to the conditional probability P(B|A), scaled by the marginal probability of P(A) divided by P(B) (7)[27].

$$p(Spam|Text\ Document) = \frac{p(Text\ Document|Spam)p(Spam)}{p(Text\ Document)} \quad (7)$$

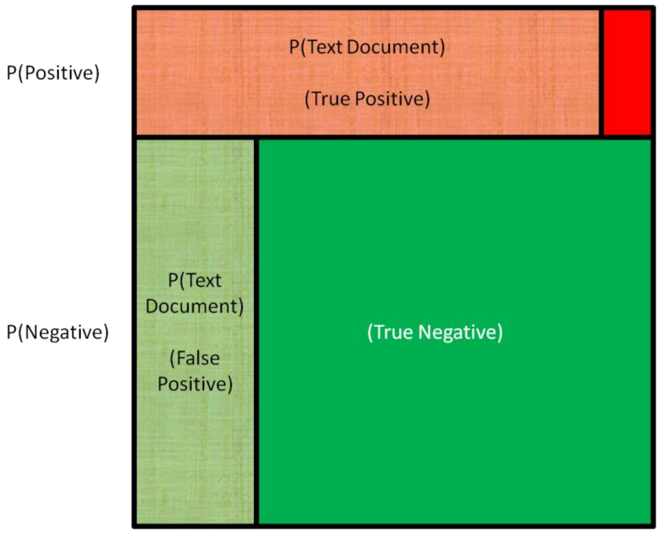

Figure 4. A trained text classifier attempts to correctly classify the presence (true positives) or absence (true negatives) of the target variable.

7.3 Deceptive Opinion

As no ground-truth labels of deceptive and truthful restaurant reviews were available, the Naïve Bayes classifier used to extract the deceptive score feature of our model is trained on the Deceptive Opinion Spam Corpus, which includes labeled reviews from other online communities and applications [30]. As shown in Table 6, truthful reviews in the Corpus dataset are scraped from TripAdvisor, Expedia, Hotels.com, Orbitz, and Priceline [30]. The user terms and guidelines applied to review content on these sites are similar to those applied to reviews on Yelp. Deceptive reviews are generated using the Amazon Mechanical Turk service, which is a platform on which users are compensated for the reviews they write [30].

Table 6. The Deceptive Opinion Spam Corpus includes labeled reviews on which classifiers of truthful and deceptive text can be trained [30].

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

The Naïve Bayes classifier uses these ground truth labels to generate a probabilistic score of a review's deceptiveness. Stratified three-fold cross-validation is applied to the training procedure on the Spam Corpus data set, randomly allocating a third of the data for testing during each iteration. As applied to this data set, the precision, recall, and F1-scores of the classifier are all 88%. After cross-validation and testing, the

Naïve Bayes classifier is trained on all of the Spam Corpus data before it is applied to the restaurant review data during feature engineering.

7.4 Extreme Comments

Yelp will flag reviews as non-recommended or remove them if they breach its terms of service[6]. To incorporate this effect into our model, the Naïve Bayes classifier described above is trained on labeled data exhibiting features that violate these terms. Sponsored by Google's Conversation AI team, the Toxic Comment Classification Kaggle data set provides these labeled data on which to train. The comments and reviews exhibiting these 'toxic' features are taken from Google services such as YouTube, Blogger, Google Maps, and Google+[17]. As above, the Naïve Bayes classifier is first trained and three-fold cross-validated during testing on the labeled data, generating the precision, recall, and F1-score results shown in Table 7. The classifier is then trained on the Kaggle data set in its entirety before it is applied to Yelp's restaurant reviews to feature engineer probabilistic scores.

Table 7. Features engineered using the extreme comments text classifier include: toxic, severely toxic, obscene, threats, insults, and identity hate.

Classifier	Observations	Precision	Recall	F1-Score
Toxic	15,294	0.96	0.96	0.96
Severely Toxic	1,595	0.99	0.99	0.99
Obscene	8,449	0.98	0.98	0.98
Threat	478	1.0	1.0	1.0
Insult	7,877	0.97	0.97	0.97
Identity Hate	1,405	0.99	0.99	0.99
Total	**159,574**			

7.5 Sentiment NLP

Features encoding text sentiment are generated using the Stanford NLP system architecture. Unlike the Bag-of-Words model, this architecture takes word order into consideration when classifying sentiment at the sentence level [31]. The work flow of the Stanford NLP system architecture is summarized in Table 8. Sentences are first discretized into individual word strings, or 'tokens'. Compound and complex sentences are split into clauses by punctuation. Parts of speech tagging identifies words as nouns, verbs, adjectives, or adverbs. Word families are identified by root word, suffix, and prefix analysis. Proper nouns are identified. Grammar rules are applied to identify the logic of sentence composition. Gender is identified, and pronouns are then linked to nouns. Using built-in definitions, words are labeled as very positive, positive, neutral, negative, or very negative.

[17] See https://conversationai.github.io for information about conversational AI.

Table 8. Work flow of the Stanford NLP system architecture for sentence sentiment analysis [31].

Procedure	Description
Tokenization	Discretize words into individual tokens
Sentence Splitting	Split sentences into clauses 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 of 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

Figure 3 shows how a recursive tree structure uses grammar rules and discretizes text into words and nested phrases to classify overall sentence sentiment [28]. To generate labels of sentence-level sentiment, the hidden layers of a recurrent neural tensor network (RNTN) encode grammar, word order, and other hierarchical linguistic information[18]. Such hierarchy is exhibited in Figure 5. A comma splits the sentence into two branches; although the first branch is negative, the overall sentiment of the sentence is positive [31]. Developed by Socher, et. al. at Stanford University, the RNTN architecture is 87.6% accurate in labeling positive and negative sentence sentiment, as measured using benchmark data derived from movie reviews [28].

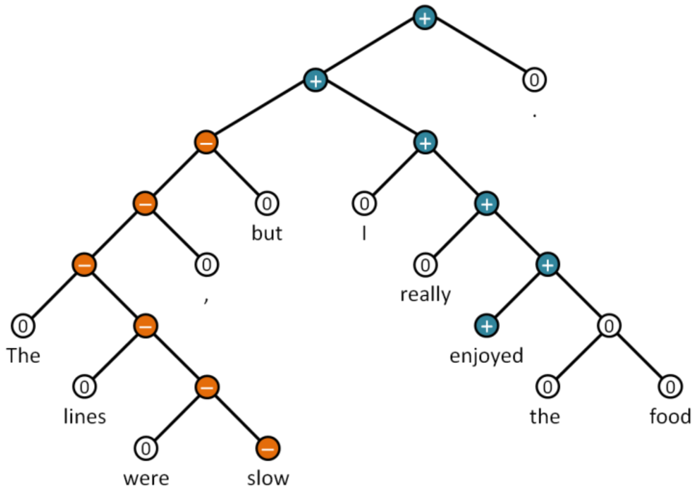

Figure 5. A recursive tree structure uses grammar rules and discretizes text into words and nested phrases to classify the data [28]. A comma splits the example sentence into two branches. Although the first branch is negative, the overall sentence sentiment is positive. [31].

[18] See ttps://skymind.ai/wiki/recursive-neural-tensor-network for more information.

7.6 Text Features Added

Table 9 shows all the textual features engineered using Naïve Bayes classification and sentiment analysis. Since every sentence in a review is assigned a sentiment score, the total sentiment is calculated as a weighted sum (8)[28]. Ranging from 1 to 5, very negative to very positive, average sentiment is then calculated by dividing total sentiment by the number of sentences in the review [28]. Average sentiment to user rating encodes the difference between review sentiment and user rating. Sentiment to average rating encodes the difference between review sentiment and the average rating of the restaurant. Each sentiment category is also quantified as the sum of all sentences exhibiting that feature divided by the total number of sentences. As indicated below, most of the features are logarithmically transformed to account for asymmetry in the data distribution. Sentiment to user rating is a feature created to validate the use of a 1 to 5 scale in quantifying sentiment. During the modeling process, this feature is removed to reduce collinearity with the text average sentiment and user rating features.

$$Total\ Sentiment = 1 * (Very\ Negative) + 2 * (Negative) + 3 * (Neutral) + 4 * (Positive) + 5 * (Very\ Positive) \quad \textbf{(8)}$$

Table 9. Attributes of features engineered using Naïve Bayes text classifiers and sentiment analysis. 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 encodes readability by grade level	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	Weighted overall sentiment score from 1 to 5	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

8 Data Exploration

To facilitate exploratory data analysis, features from review metadata and processed text are transformed to adjust for distributional asymmetry and scaled from 0 to 1. As a preliminary metric of differences between the two classes of the target variable, the means of all data features are calculated across an equal number of 26,824 observations of both recommended and non-recommended reviews. Pearson's correlation coefficients are also used to quantify the linear relation between features and the binary target variable, for which a value of one represents recommended, and a value of zero represents non-recommended reviews [32].

8.1 Mean Differences

As shown in Figure 6, calculating the magnitude of differences in mean feature values of the two classes of the target variable helps visualize what distinguishes recommended and non-recommended reviews. Relative to non-recommended reviews, features of recommended reviews having mean differences larger than 0.1 include: the presence of a profile picture, a user's number of friends, the number of reviews the user has made, the number of user photos, the number of words in the review text with and without stop words, the percent of sentences with negative sentiment, and the total number of sentences. Other features of recommended reviews having marginally larger mean differences include: the total sentiment of the text, the user-to-restaurant distance, the deceptive score of the text, the Flesch–Kincaid text readability score, the Automated Readability score of the text, the percent of sentences with neutral sentiment, whether or not the review has been edited, the percent of sentences with very negative sentiment, the number of days after October 2004 that the review was published, and the threat score of the text.

Relative to recommended reviews, features of non-recommended reviews having larger mean differences include: the percentage of sentences with positive sentiment, the average sentiment of the text, the text sentiment to restaurant average rating, the user rating, the percentage of sentences with very positive sentiment, the user rating to average rating, the text sentiment to user rating, the toxic score of the text, as well as its insult score, obscene score, severely toxic score, spelling score, and identity hate score.

Features showing no mean difference between recommended and non-recommended reviews include: the number of restaurants in the city, the number of words in the restaurant's name, the number of words in the restaurant's address, the number of reviews of the restaurant, the recommended to non-recommended review ratio of the restaurant, and the restaurant's order in the Yelp listing.

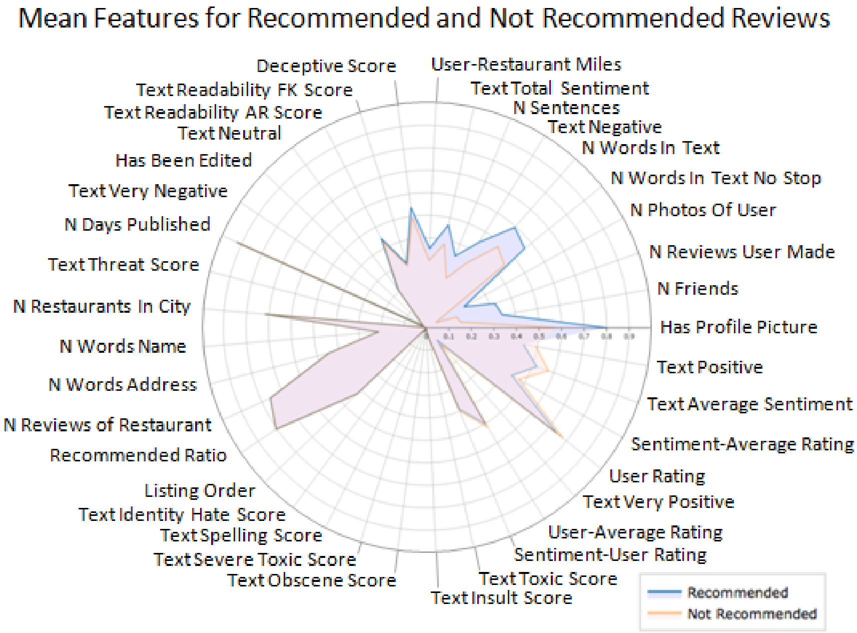

Figure 6. The magnitude of mean differences shows the divergence in feature values across recommended and non-recommended reviews.

8.2 Correlation Coefficients

Figure 7 shows Pearson's correlation coefficients, which quantify the linear relationship between feature values and the binary target variable, where a value of one represents recommended reviews and zero represents non-recommended reviews [32]. Showing stronger correlation with the recommended class, features with positive coefficients greater than 0.3 include: the number of reviews the user made, the number of photos by user, the user’s number of friends, the number of words in the text with and without stop-words, the number of sentences in the text, and the total sentiment of the text. Other features with marginally positive correlation coefficients include: the presence of a profile picture, the percentage of sentences with negative sentiment, the deceptive score of the text, the Flesch–Kincaid text readability score, the user-to-restaurant distance, the Automated Readability score of the text, the percentage of sentences with very negative sentiment, the percentage of sentences with neutral sentiment, whether or not the review has been edited, and the number of days after October 2004 that the review was published.

Showing stronger correlation with the non-recommended class, features with negative coefficients less than -0.1 include: the average sentiment of the text, and the text sentiment to average rating. Other features with marginally negative correlation coefficients include: the percentage of sentences with positive sentiment, the percentage of sentences with very positive sentiment, the user to average restaurant rating, the user rating, the severely toxic score of the text, the text sentiment to user rating, the insult score of the text, as well as its toxic score, obscene score, identity hate score, and spelling score.

Features showing no correlation with either class of reviews include: the threat score of the text, the number of words in the restaurant’s name, the number of reviews of the restaurant, the recommended to non-recommended review ratio, the number of restaurants in the city, the restaurant’s order in the Yelp listing, and the number of words in the restaurant’s address.

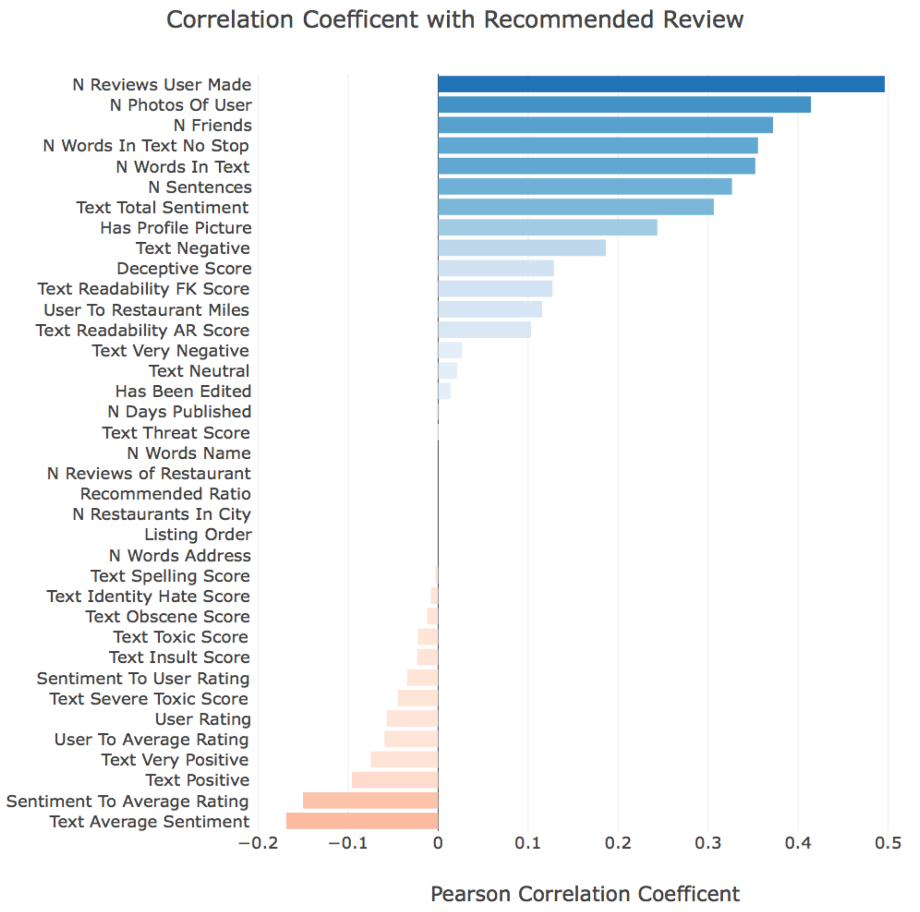

Figure 7. Pearson's correlation coefficient quantifies the linear relationship between feature values and the binary target variable, where a value of one represents recommended reviews, and zero represents non-recommended ones.

9 Results

Encoding the log of the odds ratio of belonging to the class of recommended reviews, the coefficients of the multivariate logistic regression model quantify the relationship of the features and the target being classified [33]. Features with greater magnitude have greater impact on the odds of belonging to one or the other class of the binary target variable. Features with negligible coefficient values are removed to produce a more parsimonious and interpretable reduced model. Feature significance is determined according to a p-value threshold setting alpha equal to 0.05, below which it is statistically improbable that the feature's contribution to the model is insignificant [33]. The full and reduced model are thus interpreted to infer which features contribute most toward reviews being recommended or non-recommended.

9.1 Full Model

The first pass of the modeling procedure produces the horizontal bar chart in Figure 8, which shows the primary features influencing the classification of recommended and non-recommended reviews. The full binary classification model has a 77.56% accuracy score, a 79.75% precision score, a 74.14% recall score, and a 76.84% F1-Score. As shown in Table 8, the values of the confusion matrix resulting from the full model indicate balanced accuracy metrics for both classes of the target variable.

Table 8. The results of the full model indicate that balanced sampling of the binary target variable facilitates classifying both recommended and non-recommended reviews with similar accuracy.

	Classified as Non-Recommended	Classified as Recommended
Actual Non-Recommended	21647	5018
Actual Recommended	6897	19768

Indicating higher odds of belonging to the recommended class of the target variable, features with positive coefficient values greater than 10 include: the text sentiment to average restaurant rating, the user rating, and the total sentiment of the text. Features with positive coefficient values greater than 1 include: the number of reviews the user made, the number of days after October 2004 that the review was published, the number of words in the text with and without stop words, the threat score of the text, the percentage of sentences with very negative sentiment, the number of user photos, the user's number of friends, and the Automated Readability score of the text.

Indicating higher odds of belonging to the non-recommended class of the target variable, features with negative coefficients less than -10 include: the user rating to average rating, the average sentiment of the text, and the number of sentences in the text. Features with negative coefficients less than -1 include: the severe toxic score of the text, the percentage of sentences with very positive sentiment, the percentage of sentences with positive sentiment, and the identity hate score of the text.

Other features with marginally positive coefficient values include: the user to restaurant distance, the spelling score of the text, the deceptive score of the text, the percentage of sentences with very negative sentiment, the presence of a user profile picture, the toxic score of the text, the number of words in the restaurant's address, and the Flesch–Kincaid readability score of the text. Other features with marginally negative coefficient values include: the recommended to non-recommended review ratio, whether the review has been edited, the percentage of sentences with neutral sentiment, the number of reviews, the number of restaurants in the city, the obscene score of the text, the number of words in the restaurant's name, the insult score of the text, and the restaurant's order in the Yelp listing.

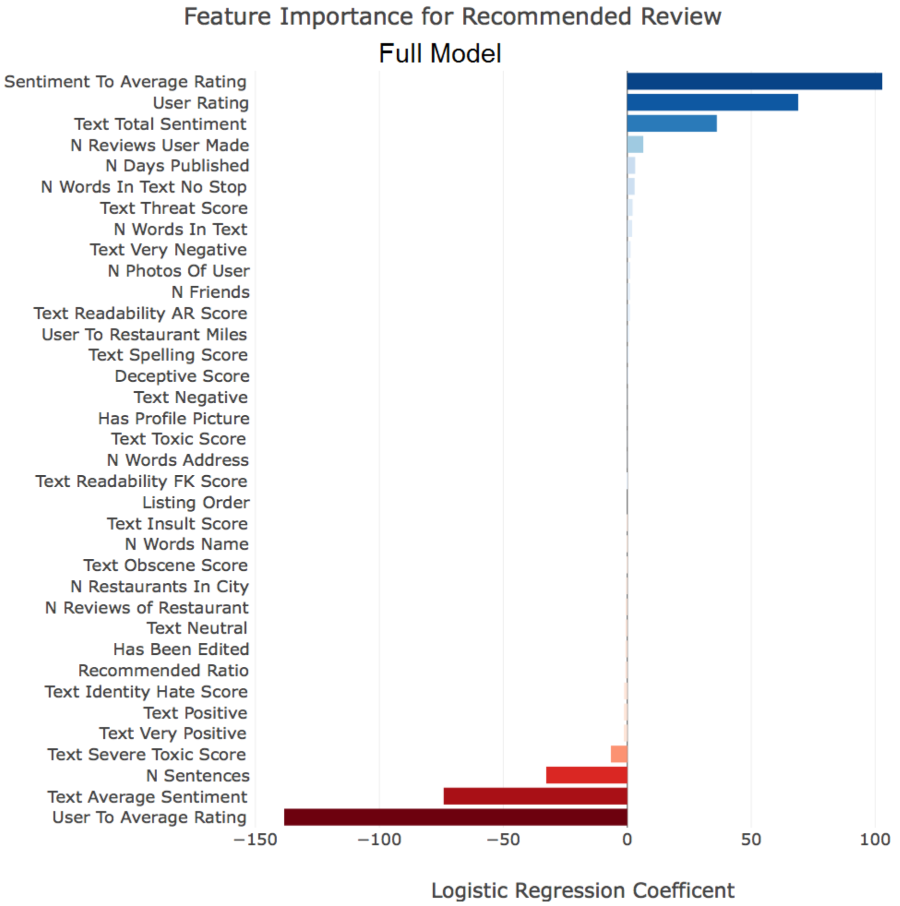

Figure 8. The primary features of the full model for classifying recommended and non-recommended reviews.

9.2 Reduced Model

As shown in Figure 9, paring down the statistically insignificant features identified in the first pass produced a more interpretable reduced model. The reduced binary classification model has a 77.61% accuracy score, a 79.71% precision score, a 74.07% recall score, and a 76.79% F1-Score. As before, the values of the confusion matrix shown in Table 9 indicate balanced accuracy metrics for both classes of the target variable.

Table 9. The results of the reduced model indicate that balanced sampling of the binary target variable facilitates classifying both recommended and non-recommended reviews with similar accuracy.

	Classified as Non-Recommended	Classified as Recommended
Actual Non-Recommended	21639	5026
Actual Recommended	6913	19752

Indicating higher odds of belonging to the recommended class of the target variable, features with positive coefficients greater than 10 include: text sentiment to average restaurant rating, the user rating, and the total sentiment of the text. Features with positive coefficients greater than 1 include: the number of reviews the user made, the number of days after October 2004 that the review was published, the number of words in the text with and without stop words, the number of user photos, the percentage of sentences with very negative sentiment, the number of friends the user has, and the Automated Readability score of the text.

Indicating higher odds of belonging to the non-recommended class, features with negative coefficients less than -10 include: user rating to average rating, the average sentiment of the text, and the number of sentences. Features with negative coefficients less than -1 include: the percentage of sentences with positive sentiment, and the percentage of sentences with very positive sentiment.

Other features with marginally positive coefficient values include: the user to restaurant distance, the spelling score of the text, the deceptive score of the text, and the presence of a user profile picture. Other features with marginally negative coefficient values include: the percentage of sentences with neutral sentiment, the recommended to non-recommended review ratio, whether the review has been edited, the number of reviews, and the number of restaurants in the city.

Statistically insignificant features removed from the reduced model include: the percentage of sentences with negative sentiment, the Flesch–Kincaid readability score of the text, the number of words in the restaurant's name, the number of words in its address, the restaurant's order in the Yelp listing, the identity hate score of the text, the insult score of the text, as well as its threat score, obscene score, toxic score, and severe toxic score.

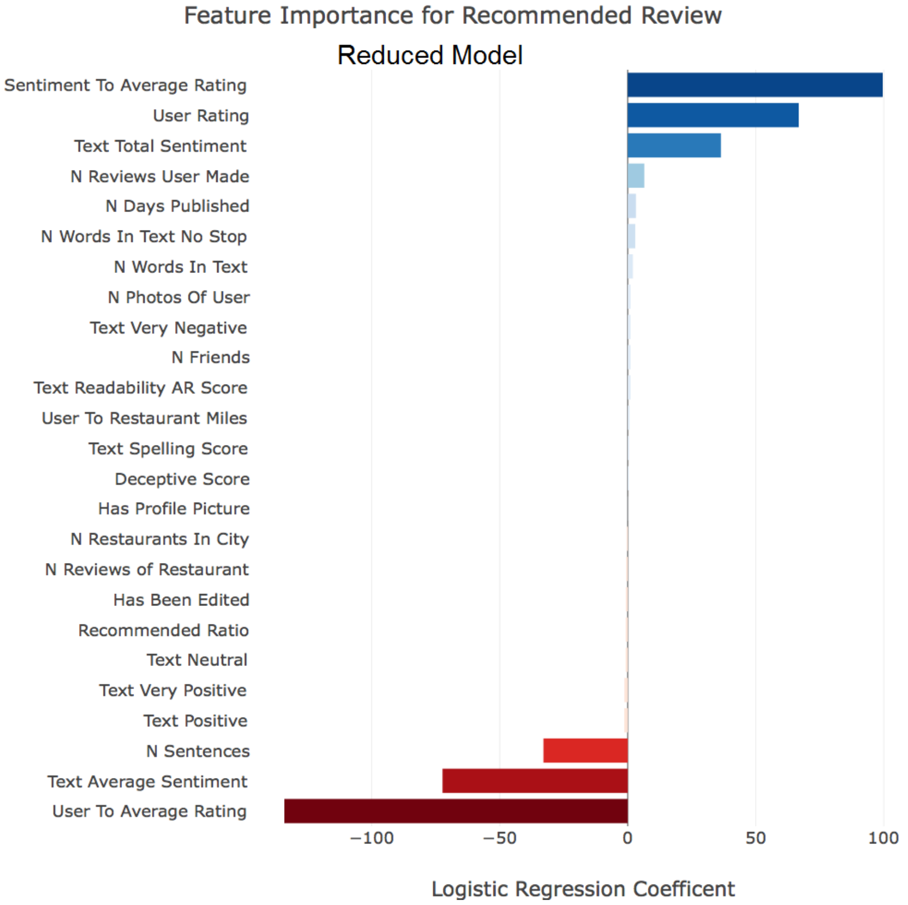

Figure 9. Statistically significant features influencing the classification of recommended and non-recommended reviews filtered by Yelp's algorithm.

10 Analysis

The coefficient values of the reduced model quantify the impact of the observed features on reviews being classified as recommended or non-recommended. This encoded information can be interpreted as a guideline for users interested in submitting recommended reviews, or for those keen on ensuring that their reviews are not flagged as non-recommended.

10.1 Guideline For Recommended Reviews

As summarized in Table 10, many of the feature coefficients of the reduced model make intuitive sense. Having submitted numerous reviews in the past (*N reviews user made*) indicates a frequent, more experienced user of the platform who is likely to know and care more about producing a well-written, credible, and informative review. The coefficient value of *user rating* indicates a general tendency to confirm a shared positive (rather than negative) experience, suggesting that reviews recommending restaurants are more likely to receive recommendations themselves. Although the meanings of other feature coefficients are less transparent, they remain interpretable. While the strongest positive coefficient (*sentiment to average rating)* appears at odds with the strongest negative coefficient (*user to average rating*), it suggests the relative importance of the review text. When it comes to recommended reviews, words are more powerful than stars. A convincing account of one's good experience is more likely to sway matters in favor of recommending the review; a five-star rating for a restaurant that has otherwise received abysmal ratings flags suspicious behavior. The discrepancy between *text total sentiment* and *text average sentiment* is related to that between *N words in text (no stop)* and *N sentences*. The higher-valued weighting of sentences with very positive sentiment (*text total sentiment*) is offset insofar as reviews providing a more balanced account of the user's positive and negative experiences are favored (*text average sentiment*). Thoroughly written reviews with concrete, descriptive details (*N words in text no stop*) can be especially informative to users trying to get a sense of what their experience may be like. Without substantive content to flesh out one's review, a series of flatly written statements, complimentary or critical (*N sentences*), is unrelatable to the reader and non-informative. More users making use of genuinely informative reviews aligns with higher odds of those being recommended in more recent history (*N days published*).

Table 10. For a review to be recommended, the feature coefficients of the model classifier suggest that users do the following:

Feature	Guideline for having a Recommended Review
User to Average Rating	Rate Critically
N Sentences	Write concisely
Sentiment to Average Rating	Write an overall positive message
Text Average Sentiment	Express variations of positive and negative sentences
N Reviews of Restaurant	Submit for businesses with less reviews
Text AR Readability Score	Write with mild complexity
N words in Text (No Stop)	Write with less common words
N Friends	Accumulate friends
N Reviews User Made	Accumulate total reviews
N Days Published	Accumulate recent reviews
Has Been Edited	Update reviews less

10.2 Insignificant Features

Although Yelp is not broadly used or known as a platform for expressing extreme comments, it is worth mentioning that none of the textual features encoding this information—identity hate score, insult score, threat score, obscene score, and (severe) toxic score—proved statistically significant in the reduced model.

11 Yelp's Ethical Role in Recommending Reviews

Four principle ethical implications of Yelp's online review platform may be considered with respect to the code of ethics promulgated by the Institute of Electrical and Electronics Engineers (IEEE)[19].

The crowdsourced nature of the information Yelp provides entails that its users also share the responsibility of upholding the IEEE principle of being "honest and realistic in stating claims or estimates based on available data." Co-responsibility is implicit in using an online platform on which users solicit information and advice from the general public. Crowdsourcing and digitally publishing public opinion can be as informative or misinformative as the users by whom such knowledge is provided. The usefulness of crowdsourcing derives from the same virtue of common sense with which one judiciously extracts meaningful information from the general consensus. Realizing the strengths and limitations of this knowledge base, Yelp applies its review filtering algorithm to attune readers to what it has gleaned from vast troves of data on user attributes and patterns of behavior [3]. In applying the algorithm, Yelp endeavors to highlight the useful information and filter out what appears deceptive and suspect. Significantly, there is no censoring of free speech on Yelp's platform; non-recommended reviews are still accessible to users [21].

Insofar as illicit users of its platform are able to disseminate misinformation for monetary gain, Yelp is charged with the particular administrative responsibility of intervening to maintain integrity of a service that "reject[s] bribery in all its forms" [1]. Its filtering algorithm relies on user activity and metadata to flag fraudulent accounts whose reviews appear purposely deceptive or disruptive [1]. Suspicious activity and evidently false, misleading, or nonsensical reviews written to artificially inflate or deflate a business's rating will be filtered as non-recommended. Users are involved in the administrative process in reporting inappropriate content that breaches Yelp's terms of service.

As host and primary administrator of its platform, Yelp is under obligation to hold itself to the same ethical standard of "[avoiding] real or perceived conflicts of interest whenever possible, and disclos[ing] them to affected parties when they do exist." Insofar as reviews influence consumer decisions and impact business revenue, Yelp must negotiate the challenge of maintaining the impartiality of its user-driven service while generating profit from hosted advertisements, some of which may be for the same businesses being reviewed [34]. Although a lack of ground-truth knowledge regarding reviewer motivation prevents disclosure of whether filtering resulted from a conflict of interest, the motivation for business marketing on Yelp is clear, as are the reasons for disclosing the terms on which it is conducted.

To protect both users and businesses, and to preserve platform openness and fairness, the same principle of impartiality demands that Yelp "[avoid] injuring

[19] See https://www.ieee.org/about/corporate/governance/p7-8.html for IEEE's code of ethics.

others, their property, reputation, or employment by false or malicious action." Non-recommended reviews include those which appear non-credible for expressly damaging the reputation of a business without reasonable justification. Where appropriate, a willingness to engage in legal proceedings indicates Yelp’s investment in protecting the interests of all parties involved on its platform.

12 Conclusions

The features identified as influential in classifying recommended and non-recommended reviews are logically consistent with Yelp's efforts to provide quality, reliable information to help consumers make informed decisions. Yelp’s platform serves to collect, organize, and summarize information toward that end.

The results of the analysis of the features driving the classification of reviews cohere with the purposes which the Yelp application serves. Generic, unqualified praise or criticism are equally uninformative. Substance and descriptive detail concerning the good and bad facilitate informed decision making. Well-founded justifications of opinions deviating from the norm can still be convincing and receive recommendation, insofar as they contribute to a wealth of different perspectives with which one can more readily make one’s own judgment. The unique power of crowdsourcing such information entails that user and platform credibility both appreciate as breadth and frequency of use increase. Insight can be gleaned from the full range of the collective user experience. Though consumer decisions are ultimately discretionary, Yelp's review filtering algorithm is designed to facilitate, inform, and empower them.

References

1. B. Pimentel, "Yelp steps up campaign against fake reviews," Marketwatch, Dow Jones Media Group, Sept. 18, 2013. [Online]. Available: https://www.marketwatch.com/story/yelp-steps-up-campaign-against-fake-reviews-2013-09-18. [Accessed: May 19, 2018].
2. M. Luca, "Reviews, Reputation, and Revenue: The Case of Yelp.com," Harvard Business School Working Paper, Harvard Business Publishing, No. 12-016, Sept. 2011. (Revised March 2016) [Online]. Available: https://www.hbs.edu/faculty/Pages/item.aspx?num=41233. [Accessed: May 19, 2018].
3. J. Stoppelman, D. Holloway, and B. Porter, "Inside Yelp: How the Company Got Its Name," Youtube, Feb. 22, 2013. [Online]. Available: https://youtu.be/Lnw3hqsGsMM. [Accessed: May 19, 2018].
4. S. Hansell, "Why Yelp Works," Bits Blog, New York Times Company, Dec. 5, 2008. [Online]. Available: https://bits.blogs.nytimes.com/2008/05/12/why-yelp-works. [Accessed: May 19, 2018].
5. A. Dalglish, "Yelp! Inc, Form 8-K, Current Report," United States Securities and Exchange Commission, Feb. 7, 2018. [Online]. Available: http://edgar.secdatabase.com/897/115752318000240/filing-main.htm. [Accessed: May 19, 2018].
6. D. Roberts, "Yelp's Fake Review Problem," Fortune, Meredith Corporation, Sept. 26, 2013. [Online]. Available: http://fortune.com/2013/09/26/yelps-fake-review-problem. [Accessed: May 19, 2018].
7. D. Streitfeld, "Yelp Tries to Halt Deceptive Reviews," New York Times Company, Oct. 18, 2012. [Online]. Available: https://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html. [Accessed: May 19, 2018].
8. A. McKeever, "Why Yelp Emerged as a Site for Social Protest," Eater, Vox Media, May 19, 2015. [Online]. Available: https://www.eater.com/2015/5/19/8588185/yelp-social-protest-trolling-memories-pizza. [Accessed: May 19, 2018].
9. G. Donnelly, "Racist Rant by New York Lawyer Tanks His Law Firm's Yelp Rating," Fortune, Meredith Corporation, June 17, 2018. [Online]. Available: http://fortune.com/2018/05/17/racist-lawyer-yelp. [Accessed: May 19, 2018].
10. Yelp PR Blog, "Our Gift to Business Owners: A Yelp App Just for You!" Yelp Blog, Yelp Inc., Dec. 15, 2014. [Online]. Available: https://www.yelpblog.com/2014/12/our-gift-to-business-owners-a-yelp-app-just-for-you. [Accessed: May 19, 2018].
11. M. Chafkin, "You've Been Yelped," Inc.com, Mansueto Ventures LLC, Feb. 1, 2010. [Online]. Available: https://www.inc.com/magazine/20100201/youve-been-yelped_pagen_3.html. [Accessed: May 19, 2018].
12. M. Luca, and G. Zervas, "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," Harvard Business School Working Paper, Harvard Business Publishing, May 2015. [Online]. Available: https://www.hbs.edu/faculty/Pages/item.aspx?num=45151. [Accessed: May 19, 2018].
13. K. Eaton, "Yelp Tweaks Its System for Transparency–and Lawsuit-Dodging," Fast Company, Mansueto Ventures, LLC, Apr. 6, 2010. [Online]. Available: https://www.fastcompany.com/1608067/yelp-tweaks-its-system-transparency-and-lawsuit-dodging. [Accessed: May 19, 2018].
14. A. Chang, "Yelp makes two major changes in the way reviews are posted," Los Angeles Times, Tribune Media, Tronc, Apr. 6, 2010. [Online]. Available: http://articles.latimes.com/2010/apr/06/business/la-fi-yelp7-2010apr07. [Accessed: May 19, 2018].
15. S. Banks, "Turning a critical eye on Yelp." Los Angeles Times, Tribune Media, Tronc, Apr. 20, 2013. [Online]. Available: http://articles.latimes.com/2013/apr/20/local/la-me-banks-yelp-20130420. [Accessed: May 10, 2018].

16. D. Post, "Exposing Anonymous Yelp Reviewers," The Washington Post, Arc Publishing, Apr. 23, 2015. [Online]. Available: https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/04/23/exposing-anonymous-yelp-reviewers. [Accessed: May 19, 2018].
17. A. Loten, "Yelp Regularly Gets Subpoenas About Users" Wall Street Journal, Dow Jones & Company, Inc, Apr. 2, 2014. [Online]. Available: https://www.wsj.com/articles/yelp-regularly-receives-subpoenas-about-users-1396467536. [Accessed: May 19, 2018].
18. K. Zetter, "Yelp Accused of Extortion," Wired, Condé Nast, Feb 24, 2010. [Online]. Available: https://www.wired.com/2010/02/yelp-sued-for-alleged-extortion. [Accessed: May 19, 2018].
19. E. Sherman, "Yelp battles to keep local reviewers anonymous," CBS News, CBS Interactive Inc, Apr. 4, 2014. [Online]. Available: https://www.cbsnews.com/news/yelp-battles-to-keep-local-reviewers-anonymous. [Accessed: May 19, 2018].
20. S. Cotter, "Quincy Center Jeweler Wins Suit Over Yelp Review," Patriot Ledger, GateHouse Media, LLC, Apr. 5, 2017. [Online]. Available: http://www.patriotledger.com/news/20170405/quincy-center-jeweler-wins-suit-over-yelp-review. [Accessed: May 19, 2018].
21. S. O'Brien, "Yelp: You can trust our reviews," CNN Tech, Cable News Network, A WarnerMedia Company, Jan. 7, 2015. [Online]. Available: http://money.cnn.com/2015/01/07/technology/ftc-yelp-reviews. [Accessed: May 19, 2018].
22. B. Egelko, "Yelp can manipulate ratings, court rules," SFGate, Hearst Communications Inc., Sept. 4, 2014. [Online]. Available: https://www.sfgate.com/news/article/Yelp-can-give-paying-clients-better-ratings-5731200.php. [Accessed: May 19, 2018].
23. CBS Local San Francisco, "Court Sides With San Francisco-Based Yelp In Lawsuit From Small Business Owners," CBS SF Bay Area, CBS Broadcasting Inc, Sept. 4, 2014. [Online]. Available: https://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: May 19, 2018].
24. E. Blair and J. Blair, "Applied Survey Sampling." Los Angeles, CA: Sage Publications, Inc., 2014.
25. R. Youngblade, "Compliments: They're free. Give them!" Yelp Blog, Yelp Inc., Mar. 1, 2013. [Online]. Available: https://www.yelpblog.com/2013/03/compliments-theyre-free-give-them. [Accessed: Mar. 3, 2018].
26. J. Brownlee, "A Gentle Introduction to the Bag-of-Words Model," Machine Learning Mastery, Oct, 9, 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-bag-words-model. [Accessed: Jul. 7, 2018].
27. M. Deshpande, "Text Classification Tutorial with Naïve Bayes," Python Machine Learning, Zenva Pty Ltd. Sept. 24, 2017. [Online]. Available: https://pythonmachinelearning.pro/text-classification-tutorial-with-naive-bayes. [Accessed: Jul. 7, 2018].
28. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality," Empirical Methods in Natural Language Processing, pp.1631-1642, 2013. [Online]. Available: https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf. [Accessed: Jul.7,2018].
29. J. Yee, "Google's Dictionary API (Unofficial) in Python," Wu Mii, Shenzhen Ermu Technology Co., Ltd, Mar. 2013. [Online]. Available: http://www.lleess.com/2013/03/googles-unofficial-dictionary-api-in.html. [Accessed: Jul. 7, 2018].
30. M. Ott, Y. Choi, C. Cardie, and J. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," Association for Computational Linguistics, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011. [Online]. Available: http://myleott.com/op_spamACL2011.pdf. [Accessed: Mar. 3, 2018].

31. C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," Association for Computational Linguistics, pp.55-60, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.[Online]. Available: https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf. [Accessed: Jul. 7, 2018].
32. D. Lane, "Online Statistics Education: An Interactive Multimedia Course of Study. [ebook] Rice University, pp.170-174. Available: http://onlinestatbook.com/2/describing_bivariate_data/pearson.html. [Accessed: Jul. 7, 2018].
33. J. McDonald, "Handbook of Biological Statistics." 3rd ed. Baltimore, MD: Sparky House Publishing, 2014.
34. J. Stoppelman, "Yelp CEO on site's popularity and pitfalls," Youtube, Jan. 13, 2016. [Online]. Available: https://youtu.be/1mlzdXwbtZo. [Accessed: Jul. 7, 2018].

Appendix

Table A. Two-stage sampling design of the cities and restaurants used for data collection.

	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**