Mining Restaurant Reviews on Yelp

Thomas Cochran
Computer Science Post-Bacc
University of Colorado, Boulder
cochran@colorado.edu

Daniel Bae
Computer Science Post-Bacc
University of Colorado, Boulder
daniel.bae@colorado.edu

Patrick Conley
Computer Science Post-Bacc
University of Colorado, Boulder
patrick.conley@colorado.edu

ABSTRACT

The goal of this project is to identify interesting patterns in Las Vegas restaurant review data on Yelp that can improve business planning and service. In this project, we sought to answer the following questions:

1. What restaurant categories are frequently reviewed by Yelp users with low and high review counts?
2. Do restaurants with high or low average star reviews cluster around specific locations in the city?
3. Are there areas in the city where review sentiment is more negative or positive?
4. Can seasonal changes improve or impair restaurant review sentiment? If so, which restaurants are most affected by seasonal changes?
5. What are common text topics in Yelp low star and high star reviews?

We found that nightlife is the most frequently reviewed category across all review counts. Other frequently reviewed categories include American, Mexican, specialty food, and fast food.

We also found areas in the city where certain restaurant categories with mediocre star reviews (0 to 3 stars) and good star reviews (4 to 5 stars) form dense clusters.

Little variation was observed seasonally based on a subset of businesses in Las Vegas. What was observed is more consistent review sentiment as the review rating increased.

Some common topics seen throughout both high and low star reviews were: manager service, atmosphere, price, and quality/fresh food. Additionally we found that reviews by users with high review counts, yielded more interpretative topics.

INTRODUCTION

Yelp is a social networking and business review website with over 173 million monthly active users [6]. On yelp, users can submit reviews to a wide array of businesses ranging from plumbing services to fine dining. When reviewing a business, users can post a star review from 0 to 5, and a text review describing their experience. These reviews are particularly important to restaurants which account for 18% of all reviewed busineses on Yelp, making it the second largest reviewed category [7].

Since restaurants are highly represented on Yelp, it is critical that they take into considertion their reputation on the platform. This dynamic presents an opportunity for data mining.

Restaurants equipped with knowledge of their reviewer demographic have a unique advantage over those that do not. For example, mining interesting patterns from Yelp datasets in certain

cities can be useful when deciding what food to serve, where to open a new branch, or gauging local competing restaurants.

Due to the usefulness of this data, it is often proprietary. However, Yelp has released an open dataset containing a subset of their review data for educational purposes [8]. Using this dataset, this project seeks to identify interesting patterns that may be beneficial to restaurant owners, both current and new, in the city of Las Vegas, NV.

Question 1 - Overview:

Our first question is: “What restaurant categories are frequently reviewed by Yelp users with low and high review counts“.

Users with high review counts use the app more frequently than users with low review counts. Identifying different preferences between these users could be useful because veteran users might offer higher quality reviews than infrequent reviewers. Additionally, sporadic users may review restaurants only when motivated by exceptionally positive or negative experiences.

Overall, this information could allow restaurants to gauge what categories are popular in the city, and this information can guide decisions on what food to serve, or what types of restaurants to open. We approach this question by grouping users by their review counts, then plot the frequency of reviews given to different types of restaurants.

Question 2 - Overview:

Our second question is: “Do restaurants with high or low average star reviews cluster around specific locations in the city?“.

The average star review for a restaurant on Yelp is a simple metric for rating a user’s experience at a restaurant. When a business on Yelp receives a review, it gives the user an option to assign a star review from 0 to 5. The average of these star reviews is listed by the restaurant, and is often used as a filter when users search for restaurants on Yelp.

Identifying the locations of high or low average star review restaurants can be useful for business owners. For example, suppose sushi bars cluster around certain streets or blocks. Opening a sushi restaurants near these areas could be challenging because of the competition. However, suppose all local sushi bars have mediocre star ratings (e.g. 0 to 3). If this is the case, then the quality of the competitors is low, but the public opinion of sushi bars in the area could be poor. Given this information, it might be advantageous to not open in the area.

We approach this question by assigning restaurant types to each restaurant (e.g. Mexican, American, Sushi, etc.). We then group each restaurant by their type and average star review, plot their locations and identify clusters.

Question 3 - Overview:

Our third question is: “Are there areas where review sentiment is more negative or positive?“.

On Yelp, every user has the option to submit a text review describing their experience. This text data is useful because it can be parsed and then used as training data to classify the sentiment of reviewers.

Identifying streets or blocks where review sentiment may be more negative or more positive can be beneficial to restaurant owners. If review sentiment tends to be more negative in certain areas, then the business owner can choose to avoid these areas, or be more attentive to common issues in an effort to improve the quality of service.

When location data was used to plot business review sentiment there was little observable patterns. The distribution of review sentiment did not show clear locations with more or less favorable review sentiment. In addition, the distribution of businesses by ratings and sentiment covered the entire Las Vegas metro area, suggesting no region contains a greater amount of favorable sentiment and business rating when assessed across all business categories. While there were regions observed with higher rated businesses in specific categories, the visualization of all categories in the city suggests no regions were more favorable.

Question 4 - Overview:

The review data set contains timestamps for each review and most businesses contain years of reviews. When plotted by month, the review sentiment provides a distribution of sentiment for that month when using data over the course of multiple years. Interestingly, there was little variation in the review sentiment throughout the year. Review sentiment was consistent for each user assigned rating and no major patterns were observed.

Question 5 - Overview:

Our fifth question “What are common text topics in Yelp low star and high star reviews”, seeks to extract distinct and unique topics in our reviews.

The idea behind binning reviews by low star and high star reviews, is to find both positive and negative constructive feedback. Additionally, by binning reviews by frequent and infrequent users, is to evaluate if frequent Yelp leave higher quality feedback.

A topic can thought of as an abstract label we’ve assigned to a specific collection of words [11]. To illustrate topics more concretely, let’s use a simple example. Suppose we looked at 1 star reviews 2 topics that could’ve developed are:

Topic #1: {‘Rude’, ’Manager’, ‘Service’, ’Never’}.

Topic #2: {‘Fresh,’, ‘Bad, ’Disgusting’, ‘Taste’}.

A review tagged with Topic #1, suggests that "poor customer service" was likely context of the review. While a review tagged with Topic #2, suggests that "product quality" is more likely context of the review. In general, the motivation is to find specific and actionable feedback, based on top and bottom rated reviews

RELATED WORK

1. Inferring Future Business Attention

This article uses an older version of our dataset and focuses on businesses in Phoenix, AZ. Its goal is to predict the popularity of a business in the future by creating a model that will predict the number of reviews a business will receive within the next six months.

2. Yelp Text Mining and Sentiment Analysis

This article uses the Yelp API to collect its restaurant data. It selects 17 burger restaurants in the Bay Area, and performs basic data exploration and cleaning tasks, such as removing unnecessary review text strings, converting review text into bag-of-words, and performing lexicon-based sentiment analysis on reviews.

3. Identifying Restaurant Features

This article uses an older version of our dataset and performs sentiment analysis of review text by creating their own classifier with a support vector machine (SVM). The authors did several preprocessing tasks that we will do, such as filtering out restaurants businesses, looking at the distribution of cuisine types, and label review

sentiment. Interestingly, the study applies a word score to label the degree of positivity or negativity of the words used in each review.

4. Predicting Restaurant Closure

This article uses a Yelp dataset consisting of 3327 restaurants in Phoenix, AZ and constructs a model to predict whether a restaurant will close within the next year. The author also uses the model to rank feature importance among all restaurants to highlight some potentially important aspects that may keep a restaurant open. The top two features are (1) whether the restaurant is a chain, and (2) the number of reviews relative to restaurants nearby.

DATA SET

The Yelp Open Dataset contains over 8 million user reviews on over 160 thousand businesses in 8 metropolitan areas in 5 json files.

1. yelp_academic_dataset_business.json

Business data including geolocation, business category, star review, and operation hours.

2. yelp_academic_dataset_checkin.json

Business identification and the dates that users have reportedly checked in.

3. yelp_academic_dataset_review.json

Full review text data paired with user identification and business identification.

4. yelp_academic_dataset_tip.json

Full text for user tips, which are shorter reviews paired with dates, and business identification.

5. yelp_academic_dataset_user.json

User information including user identification, friends list, join date, and review count.

For detailed documentation covering every attribute in this dataset is provided by Yelp, see reference [9].

MAIN TECHNIQUES APPLIED

1. General data cleaning techniques:

Each of the initial json files used were cleaned in different ways which resulted in new sets of cleaned json files. These cleaning steps were performed primarily by loading, cleaning, and exporting the initial jsons in Python using the pandas package.

The initial json file containing business data was filtered to only contain businesses in Las Vegas, then filtered again to only contain business categories listing ‘Restaurant’ and ‘Food’.

The initial reviews json file containing review data was filtered to only contain text reviews for our filtered restaurants in Las Vegas. The review text was cleaned to eliminate extraneous characters and prepared for sentiment analysis.

The initial users json file containing user account data was filtered to contain users who reviewed our filtered restaurants in Las Vegas.

Redundant attributes were dropped, determined by Pearson correlation analysis.

After we cleaned and pre-processed our five initial json files, we integrated the data into an Sqlite relational database.

2. Apriori pruning business categories:

We used the apriori algorithm implementation provided by the machine learning extensions (mlxtend) python library. For detailed documentation, see reference [10].

Restaurants in our Yelp dataset can be identified by looking up the business category of ‘Restaurant’. However, business categories are often used liberally on Yelp. As a result, many businesses categorized as a restaurant are not necessarily a restaurant. For example, if we look

at the most frequent and least frequent business categories paired with the ‘Restaurant’ business category:

Top 5 Restaurant Categories

	categories	count
0	Nightlife	1168
1	Bars	1080
2	Fast Food	1063
3	American (Traditional)	959
4	Mexican	943

Bottom 5 Restaurant Categories

	categories	count
613	Bounce House Rentals	1
614	Airsoft	1
615	Furniture Rental	1
616	Yelp Events	1
617	Towing	1

We find that ‘Restaurant’ is paired with extraneous business categories such as ‘Furniture Rental’ and ‘Towing’. If we look at the ‘Towing’ business that also calls itself a restaurant, we find that it lists itself with the following additional business categories:

Transmission Repair, Pizza, Financial Services, Auto Parts & Supplies, Auto Insurance, Towing, Oil Change Stations, Insurance, Restaurants, Windshield Installation & Repair, Auto Repair, Auto Glass Services, Automotive, Body Shops

Businesses like this pose a challenge for our project, since we want to focus exclusively on restaurants that only serve food, particularly those that specialize in different types of cuisines. In order to accomplish this, we need a method to select, among all businesses categorized as a ‘Restaurant’, only those businesses with the most frequent business category associations.

Our solution is to filter by category ‘Restaurant’, then use the apriori algorithm to mine frequent business category associations and then filter restaurants by these frequent associations. For example, among all businesses with a ‘Restaurant’ tag, the following association may have a high lift:

American (Traditional) => Burgers

If this is the case, we would keep restaurants that contain this business category association. After running the apriori algorithm, we found that with a 1% confidence threshold and 3% lift we could eliminate many restaurant categories with uncommon associations. Many of the associations found were also consistent with what we expect for certain types of restaurants, for example:

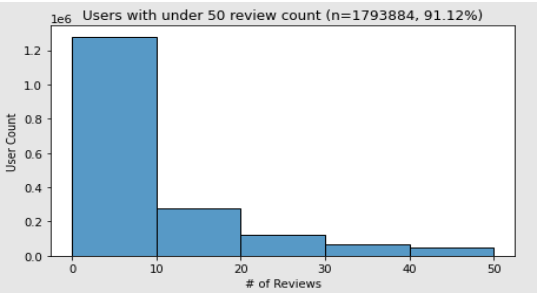
Association rules (High to Low Lift)

	antecedents	consequents	lift
186	(Nightlife, Sports Bars)	(American (New))	4.350748
349	(American (New))	(Bars, Sports Bars, Nightlife)	4.350748
10	(Sports Bars)	(American (New))	4.350748
11	(American (New))	(Sports Bars)	4.350748
75	(Breakfast & Brunch)	(Cafes)	4.294391
74	(Cafes)	(Breakfast & Brunch)	4.294391

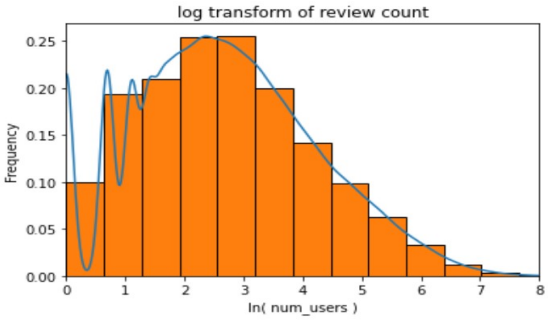
In the above association rules, we expect cafes to be associated with breakfast, and sports bars to be associated with American food.

3. Review count log transform and binning:

We planned on grouping Yelp users by their review counts, such as: low, medium, or high review count. However, the vast majority users submit between 0 and 10 total reviews:

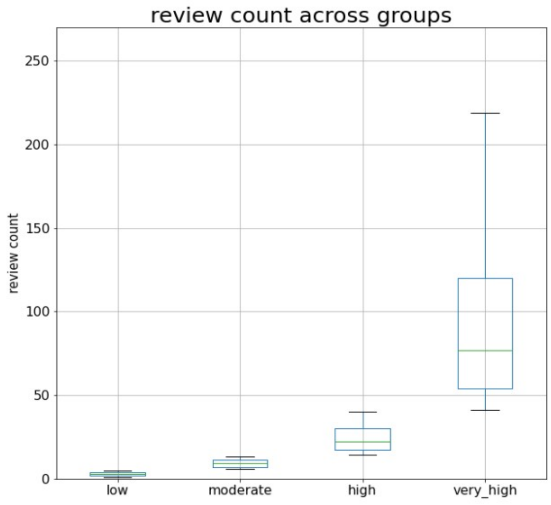


Since review counts are highly skewed, we did a logarithmic transformation:



This allowed us to separate Yelp reviewers into quartiles: low (<Q1), medium-high (Q3-Q1) and

very high (>Q3) review counts. We then inspected the distribution of review counts within each group using box plots:



Using this method, the distribution of review counts in each group is more normal, however we still eliminated outliers in the very high review count group whose review count exceeded:

$$MAX = Q3 + (1.5 \times IQR)$$

4. Clustering-based classification

We utilized apriori pruning to eliminate fringe restaurants that have very liberal use of business categories. However, we still have restaurants that are not labeled as a specific type. Each restaurant is described by a long list of business categories. For example, a restaurant serving Mexican food may list its categories as:

[Restaurant, Food, Tacos, Mexican, Burritos, Fast Food].

This poses a new challenge for our project, because some of our questions require grouping restaurants into single restaurant types. Ideally, the above restaurant would be assigned the restaurant type of 'Mexican', allowing it to be grouped with other Mexican restaurants.

We approached this problem by classifying each restaurant using these restaurant category lists as follows.

4.1 Term frequency - reverse document frequency (TF-IDF)

We approached this problem first by transforming our category lists into vectors, then assigning weights to each term using an information retrieval technique called term frequency-inverse document frequency (TF-IDF).

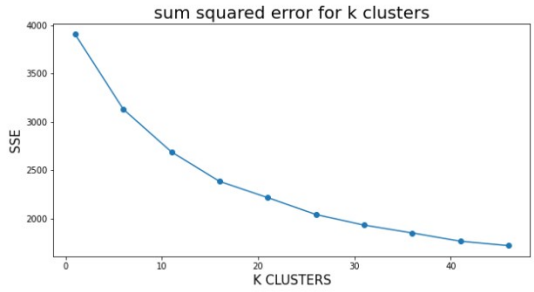
This method extracts information from text by assigning weights to words using the product of two statistics. The first is term frequency, which assigns larger weights to more frequent words. The second is inverse document frequency, which assigns larger weights to words that occur less frequently across all documents.

We thought this would be ideal for extracting information from our restaurant category lists if we treat each category as a document. For example, most of our categories have 'Restaurant' and 'Food'. Since these are common terms among all categories, the inverse document frequency would be low and therefore these terms would be assigned smaller weights. However, the categories 'Mexican' or 'Tacos' would be frequent among mexican restaurants, but infrequent among all other non-mexican restaurants. This would allow 'Mexican' or 'Tacos' to be assigned larger weights.

4.2 K-Means clustering tf-idf scores:

After vectorizing and weighing terms in our restaurant category lists, we clustered terms using their TF-IDF weights and K-Means. In order to determine the starting number of k-centroids, we graphed the sum-squared error (SSE) over a range of 'k'. Since SSE is the sum squared difference between each point being clustered and its mean, a smaller SSE yields

tighter clusters. Therefore, in our plot of SSE over a range of ‘k’, we selected ‘k’ where SSE approaches a minima:



From this graph, we selected an initial value of k=30 and assigned category labels to each of our restaurants using K-means as shown below:

	categories	restaurant_type
0	Mexican, Restaurants, Fast Food	mexican
1	Burgers, Restaurants, American (Traditional), ...	burgers
2	Fast Food, Restaurants	fastfood
3	Specialty Food, Health Markets, Food, Shopping...	specialtyfood
4	Pizza, Salad, Burgers, Restaurants	pizza
...	...	...
4269	American (New), Karaoke, Restaurants, Lounges,...	american
4270	Salad, Sushi Bars, Japanese, Restaurants, Asia...	specialtyfood
4271	Delis, Restaurants, Sandwiches, Food, Pizza	sandwiches
4272	Pizza, Italian, Restaurants	pizza
4273	Chinese, Restaurants	chinese

In the table above, each row is a restaurant in our dataset. The business categories are listed under the categories column, and the labeled restaurant types are listed under the restaurant_type column. These labels allow us to continue our investigations since we can now group each restaurant by the food it is predicted to serve given its long business category list.

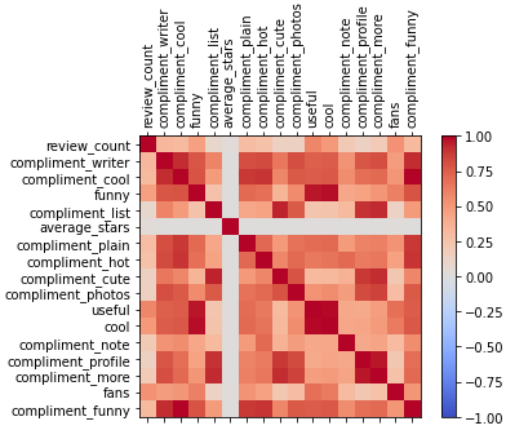
5. Density-based clustering:

We identified clusters of restaurants with low or high average star review at different locations around the city using density-based spatial clustering of applications with noise (DBSCAN).

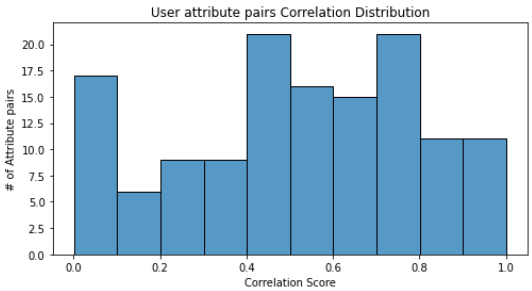
We chose DBSCAN because we intended to discover dense clusters of restaurants along small strips or blocks of the city. This task is well suited for DBSCAN because, with a low epsilon value, we are able to reduce the distance between neighboring businesses within clusters. This allowed us to limit the distances between restaurant in each cluster to distances that are comparble to blocks or small strips. It also made it easy to detect outliers.

6. Reducing redundant user attributes

Users in the Yelp dataset contained 22 total attributes, to which many of them seemed quite correlated to each other. In order to select the attributes that we wanted to keep; we first normalized the data with Z-scores. Using the Z-scores allowed us to compare different attributes to each other, on a standard scale. Then we calculated the correlation matrix (using Pearson Coefficient):



In order to selectively drop redundant (or highly correlated) attributes; a correlation threshold needed to be defined. Looking at the distribution of Pearson correlation coefficients:



We decided to use a threshold of a 25% Pearson coefficient, as we were selecting for more independent attributes. As a result, only 7 attributes were kept out of the 17 that we were indifferent towards.

The goal is to use this well-rounded set of attributes, as a basis to perform unsupervised clustering methods. More concretely, to classify or profile user behaviors per their frequency, as described above in section 2 “Review count frequency of users”.

7. Review Topics Sentiment Analysis

Several python libraries were used for this analysis, such as sklearn, nltk and textblob. These libraries support quick tokenization of strings as well as removal of ‘stopwords’ or words that are known to convey little information regarding sentiment.

7.1 Text Pre-processing

In order to build a distinct set of topics, we needed the remove noisy words that build topics.

First we tokenized the text to seperate each distinct word in the documents. These tokens serve as bag-of-words that are unique. And will serve as the basis to build K topics.

Then we apply another filter to our tokens by removing English stop words. Some examples of English stop words : "the", "is", and "are".

Lastly, to standardize the words across documents we applied lemmatization to distill words from some tense to the present form (ex. Looking -> Look)

7.2 Modeling topics with LDA

Latent Dirichlet allocation (LDA) is unsupervised modeling technique to determine the major topics throughout a collection of documents. LDA assumes every document’s topic is a probability distribution of topics. In other words, every document contains a mixed bag of several topics.

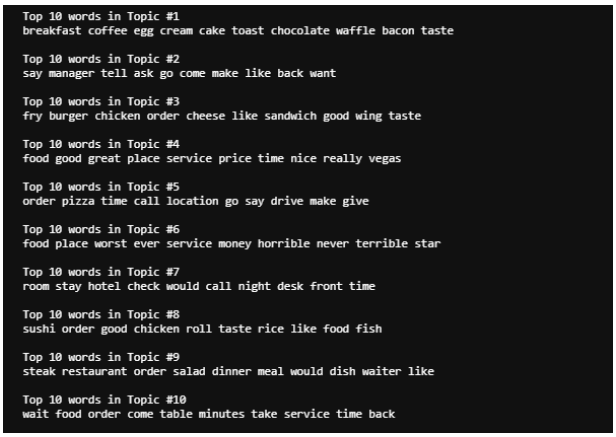
Then we used LDA to scan all the words and assigns each document term to a random topic out of K topics. As we scan more documents, our topics become more and more distinct.

We assign words to topics based off:

1. How often a word appears in the document
2. How often a word occurs in our topic

Some hyper-parameters we chose were: 10 topics (K), and 10 words in a topic.

Topics: Good star reviews (>=4 stars) by *Low* review count users



Topics: Good star reviews (>=4 stars) by *High* review count users

7.3 Modeling topics with NMF

Non-negative matrix factorization (NMF) is an multivariate algorithm we used in conjunction with our TF-IDF matrix. LDA's are favorable when we have coherent topics. Where coherent topics are topics that have a high intra-similairty of words. The topics that are more incoherent are typically more suited for NMF applications [12].

Given that a single restaurant reviews can cover a wide range of topics (ex. service, food quality, location, price, etc.) The incoherant topics developed in our NMF model appear to be more interpretable as the model is more flexible to a range of topics.

Topics: Good star reviews (>=4 stars) by *Low* review count users

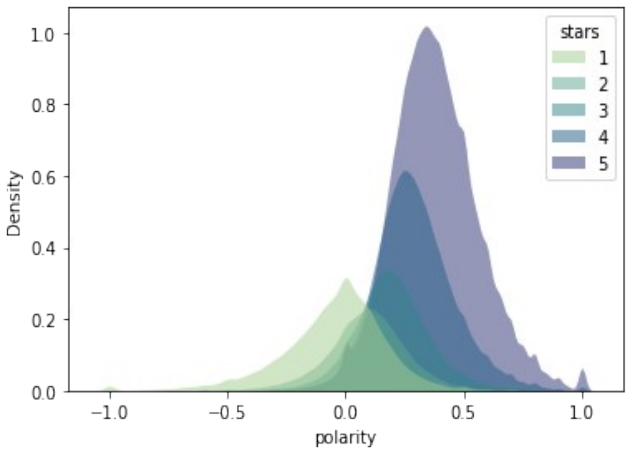
```
Top 10 words in Topic #1
come say ask table tell drink manager time want make

Top 10 words in Topic #2
order time wrong delivery drive minutes location phone pick forget

Top 10 words in Topic #3
room hotel stay check desk clean book night smoke smell

Top 10 words in Topic #4
chicken fry burger taste like sauce cheese cook flavor meat

Top 10 words in Topic #5
pizza wing delivery cheese crust slice pepperoni pizzas cold deliver

Top 10 words in Topic #6
wait minutes seat line long hour table time mins finally

Top 10 words in Topic #7
food service horrible terrible customer worst slow cold come time

Top 10 words in Topic #8
great friendly atmosphere service staff love food definitely delicious amaze

Top 10 words in Topic #9
place sushi roll recommend love like vegas really star better

Top 10 words in Topic #10
good price really pretty food nice portion friendly little overall
```

Topics: Good star reviews (>=4 stars) by *High* review count users

```
Top 10 words in Topic #1
like order steak taste really come dish salad restaurant sauce

Top 10 words in Topic #2
room stay hotel check strip clean pool casino nice night

Top 10 words in Topic #3
burger fry burgers shake cheese onion good truffle ring shack

Top 10 words in Topic #4
wait minutes order table come time drink food seat line

Top 10 words in Topic #5
buffet crab line legs selection station seafood vegas desserts worth

Top 10 words in Topic #6
good food service price place pretty really nice vegas better

Top 10 words in Topic #7
breakfast chicken waffle sandwich egg bacon fry toast hash portion

Top 10 words in Topic #8
pizza slice crust pepperoni secret cheese place pizzas white good

Top 10 words in Topic #9
great service place drink food love atmosphere vegas friendly staff

Top 10 words in Topic #10
sushi roll fish ayce tuna rice fresh sashimi quality nigiri
```

7.4 Generating Sentiment Scores from Tokenized Review Text

Both nltk and textblob are well developed natural language processing libraries that provide a variety of text analysis. Both libraries support generation of classification models but also contain pre-built models for analysis. Textblob contains a built-in corpus of english words with sentiment scores (positivity, subjectivity) that allow for sentiment analysis on tokenized text without the need for building a unique model [15].



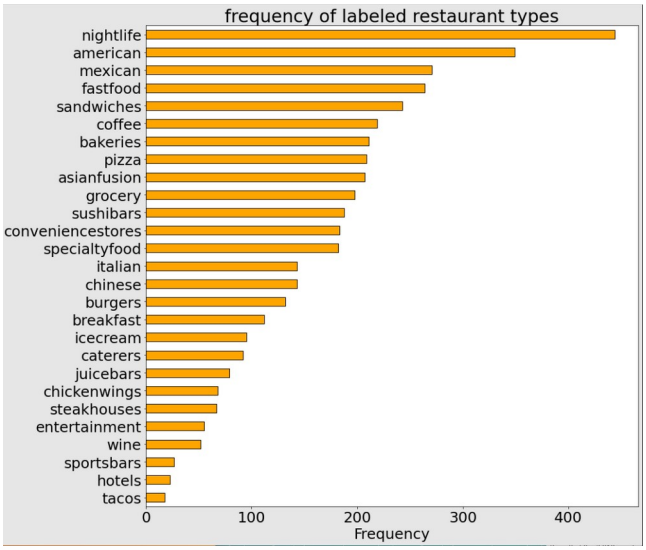
Application of this model to tokenized review data from our subset of users/business allows for scoring review text and further analysis. A kernel density estimate of the distribution in review sentiment from this generic modeling shows distinct distributions of review sentiment for each user assigned rating, suggesting the built in scoring is sufficient for this analysis.

KEY RESULTS

1. Review frequencies differ among users with low and high review counts:

Using the business category lists, each restaurant was assigned a restaurant type. Nightlife is the most frequently reviewed

restaurant type, followed by American, Mexican, Fast Food, and sandwiches:

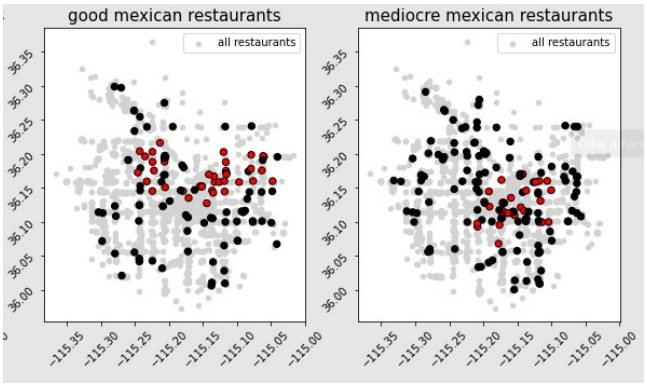


American restaurants are reviewed very frequently by users across each group. Low review count users tend to review fast food restaurants more frequently, while high review count users review asian fusion restaurants, stakehouses, and bakeries more frequently.

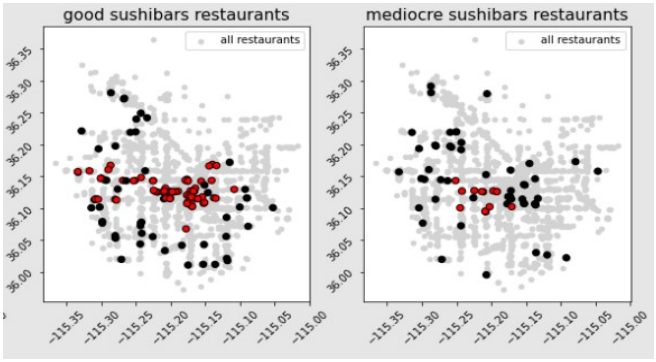
2. Restaurant types of varying quality form clusters around different locations of the city:

Note: Clusters identified by DBSCAN are labeled red, while outliers are labeled black. All other restaurant types are labeled gray.

Mexican restaurants with a low average star review between 0 and 3 (mediocre) form clusters closer to the center of the city:

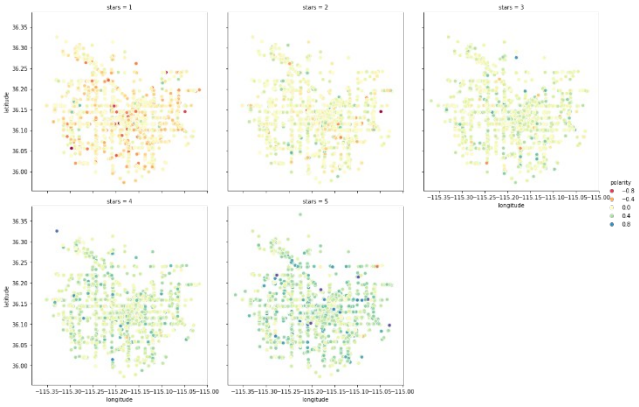


In contrast, sushi bars with high average star reviews (good) form clusters closer to the center of the city near the casinos:



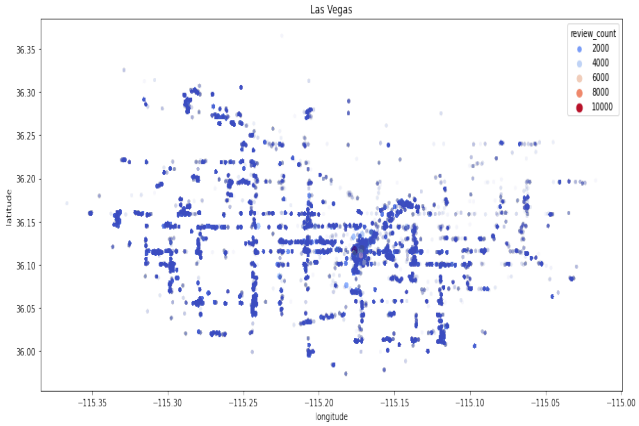
3. Variation in sentiment based on location

When plotting sentiment scores by business location, the distribution of highly rated business seems equivalent throughout the city. When subdividing by user ratings, again there is no clear distinction that one region has significantly higher user sentiment over another. The sentiment scores are colored by polarity and there is a clear distinction between the determined sentiment and the assigned user rating suggesting the sentiment scoring is accurate. However, there is no noticeable region of the city that attracts more favorable business when viewed using every business category.



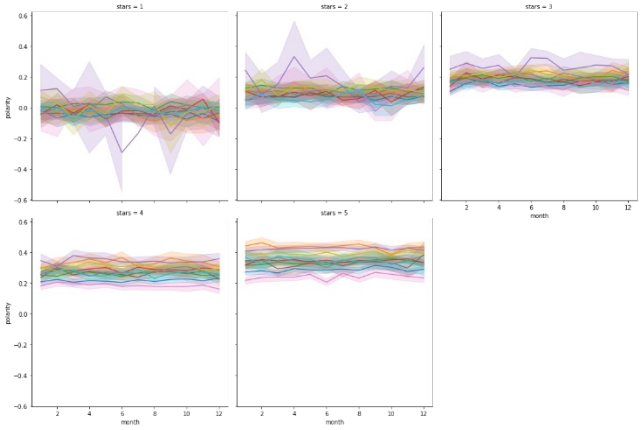
When considering total review count, there is obviously a concentration of businesses with more reviews in the city center. Looking further, it

appears even businesses on the outskirts of Las Vegas receive a similar amount of reviews when compared to businesses closer to the city center.



4. Variation in sentiment based on season

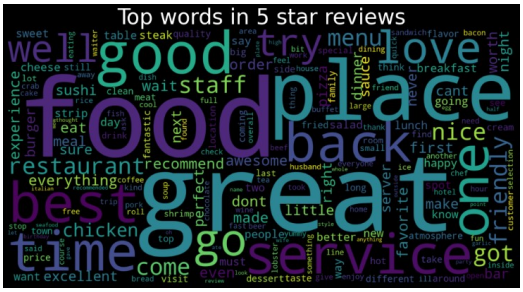
When plotting sentiment scores by month, the multiple years of collected data allows for assessment of the distribution of the sentiment at each month. Interestingly, there was little variation observed throughout the year. Below is a plot of the 10 businesses with the greatest number of reviews and their subsequent sentiment scores at each month. The shaded regions bordering each line represent the distribution of sentiment scores over the years of data collected.

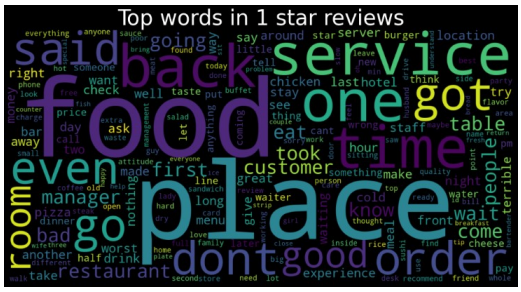


More variation was seen in the 1 and 2 star reviews but only for a single business. Overall, there is little change over the course of the year in review sentiment.

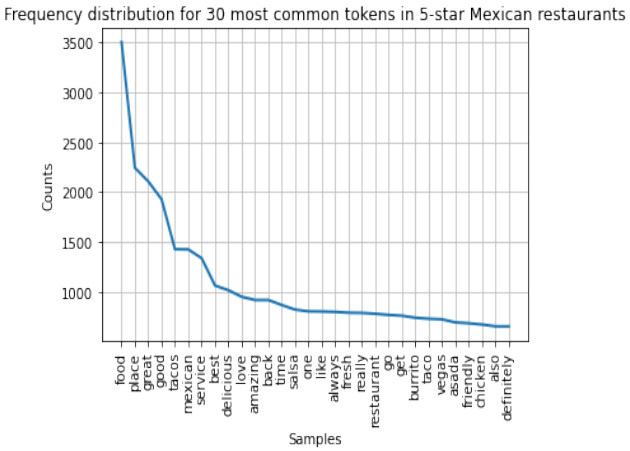
5. Common topics among review rating

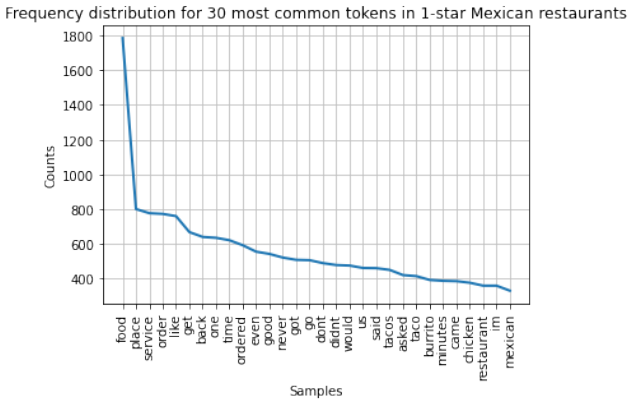
Using the tokenized review text, the top word across all categories is food. While this is not particularly interesting, it does highlight the most important aspect of good and poorly rated businesses.





Even when looking a single category, like Mexican restaurants, again the most common word across good and bad reviews is food.





A prospective restauranteurs should consider the most important aspect of their restaurant to be the food. A standout word among 1-star reviews is service, highlighting people tend to poorly rate restaurants if they receive poor service.

APPLICATIONS

What Las Vegas restaurants generate the most hype?

As expected, restaurants classified as 'nightlife' were the most frequently reviewed restaurants across all reviewer types. This was followed by American restaurants and Mexican restaurants. A restaurant opening in Las Vegas may use this data when deciding whether to open an Mexican restaurant versus an Italian restaurant, since Mexican restaurants appear to get more attention by reviewers on Yelp.

It is important to note, however, that more reviews may not necessarily be desirable. Fast food restaurants are also frequently reviewed, yet our DBSCAN results indicate that there exist several mediocre fast food restaurant clusters, and no good fast food restaurant clusters. Furthermore, the number of good fast food restaurant outliers were quite small.

Does restaurant location matter?

Our DBSCAN results indicate that for some restaurant categories, there do exist streets or blocks with dense clusters of restaurants that have either higher or lower average star reviews. However, this effect seems to be highly dependent on the restaurant type. Mexican restaurants and sushi bars have very clear clusters of both high and low average star review clusters. Yet there are no good clusters for fast food and coffee restaurants. Given these differences across restaurant types, applying this information would likely have to be done on a case-by-case basis. A business owner with a choice of neighborhood should not anticipate any effect on reviewer polarity in different neighborhoods.

Prospective restauranteurs should be able to assess their competition using the above location and review data. Perhaps they have several locations to choose from and one may be far from other highly rated restaurants of the same type. This may provide an advantage by lowering competition. It appears that people are always seeking good food, no matter the time or place, and they tend to leave particularly scathing reviews if there is poor service.

REFERENCES

[1] V. Hood, V. Hwang, J. King. Inferring future business attention. https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_InferringFuture.pdf

[2] Elva Xiao. 2018 Text Mining and Sentiment Analysis for Yelp Reviews of A Burger Chain. https://towardsdatascience.com/text-mining-and-sentiment-analysis-for-yelp-reviews-of-a-burger-chain-6d3bcfcab17b

[3] B. Yu, J. Zhou, Y. Zhang, Y. Cao. Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. https://arxiv.org/pdf/1709.08698.pdf

[4] Z. Zhang. Machine Learning and Visualization with Yelp Dataset. 2017. https://medium.com/@zhiwei_zhang/final-blog-642fb9c7e781

[5] Kaggle competition, 2019. Clustering with Kmeans, PCA, TSNE. https://www.kaggle.com/aussie84/clustering-with-kmeans-pca-tsne

[6] T. Zhu. App Critique-Yelp. 2019. https://medium.com/@theresazhu/app-critique-yelp-29a7df5e5bd6

[7] Fast Facts: An Introduction to Yelp Metrics as of December 31, 2020.https://www.yelp-press.com/company/fast-facts

[8] Yelp Open Dataset, 2020. https://www.yelp.com/dataset

[9] Yelp Open Dataset Documentation, 2020. https://www.yelp.com/dataset/documentation

[10] Machine learning extensions (mlxtend) library documentation. http://rasbt.github.io/mlxtend/

[11] Topic model https://en.wikipedia.org/wiki/Topic_model

[12] Ex-Twit: Explainable Twitter Mining on Health Data https://arxiv.org/pdf/1906.02132.pdf

[13] Natural Language Toolkit — NLTK 3.6.2 documentation

[14] TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation

[15] https://github.com/sloria/TextBlob/blob/dev/textblob/en/en-sentiment.xml