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

PROBLEM STATMENT AND MOTIVATION

The goal of this project is to identify text patterns in Yelp restaurant reviews that can facilitate improvements in food, service, and business planning. In our study, we will focus on two types of reviews in the city of Las Vegas, NV. First, we will explore reviews received by different categories of restaraunts. Second, we will explore reviews written by Yelp users with varying review counts.

1 Restaurant Reviews:

These reviews are received by restaurants of a certain category, e.g. American, Italian, etc. In this set of reviews, we aim to identify common issues associated with restaurants serving different types of food. There are three questions that we will focus on.

1. What are frequent features of review text for low rated (1-2 star) and high rated (4-5 star) restaurants?
2. Are there streets or blocks where restaurants are more likely or less likely to receive negative review sentiment?
3. Can seasonal changes improve or impair restaurant review sentiment? If so, which restaurants are most affected by seasonal changes?

In answering each of these questions, we hope to create a short profile of common issues faced by restaurants serving different kinds of food. An example of this profile is as follows.

Italian Restaurant Profile: In Las Vegas, reviews for Italian restaurants with 4-5 stars focus on "authenticity" and "service". Italian restaurants on Washington Ave receive more negative review sentiment during the Winter than the Summer, while Italian restaurants at Charleston Blvd are invariant to seasons.

2 Yelp User Reviews:

These reviews are generated by Yelp users with low ($n < 5$), infrequent ($n < 15$), and frequent ($n < 50$) review counts. In this set of reviews, we aim to discover patterns among different types of Yelp users that may be beneficial to restaurant businesses. We will focus on two questions.

1. Do low, infrequent, and frequent reviewers cluster around certain attributes such as: streets, average star review, or restaurant category?
2. How does review sentiment vary among low, infrequent, and frequent reviewers?

An example application of this data is as follows. If a certain block in Las Vegas has clusters of Yelp users who review more frequently with negative sentiment, then it may be beneficial to avoid opening in that area. Conversely, if a block has a cluster of frequent positive reviewers then it may be beneficial to open a restaurant in that location.

LITERATURE SURVEY

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.

PROPOSED WORK

Since our dataset is distributed for academic purposes by a large company, it is well structured and documented. As a result, much of our preprocessing work will involve reducing the dimensionality of our data and ensuring the review text is legible and ready to be processed.

1. Cleaning:

Remove non-English reviews and special characters. Remove business categories unrelated to restaurants. Remove users with no text reviews, remove redundant attributes using correlation analysis. Remove unused attributes. Remove restaurants that are closed.

2. Integration:

Create a schema for our relational database. Merge all relevant json files into a sqlite database. If time permits, upload and access the database via google cloud.

3. Processing derived data:

Restaurant Reviews:

Group restaurants based on star rating. Classify review text sentiment. Identify common words and phrases used in negative and positive reviews. Identify streets and dates with frequent negative and positive reviews.

Yelp User Reviews:

Separate users into different categories based on review count. Partition data from the top-down by splitting on selected attributes and cluster based on similarity.

The questions posed in this project differ from those in the listed literature. However, since each study involves restaurants using the Yelp dataset, there will be similarities in data cleaning and some processing procedures such as clustering or review sentiment analysis.

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 bu siness identification.

5. **yelp_academic_dataset_user.json**

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

EVALUATION METHODS

We will use three methods when evaluating our results.

First, we will look at correlations between our findings and closed restaurants. This will allow us to evaluate negative sentiment in our restaurant profiles since we expect closed restaurants to be highly correlated with negative sentiment.

Second, we will split our reviews into 80/20 training and validation subsets. This split allows us to determine whether frequent features found in the training set are in the validation subset.

Third, we will search the literature and look for studies with similar methodology and see if similar conclusions were made.

TOOLS

We will be using Python as our programming language and github for version control. Our project code will be organized in jupyter notebooks within our project repository. We will be using a virtual environment for dependency management and a variety of python packages, including: *pandas, scipy, numpy, matplotlib, nltk, textblob, sqlite, geopandas.*

Update: We used the apriori algorithm implementation provided by the mlxtend library (rasbt.github.io/mlxtend)

MILESTONES

We are expecting to complete portions of the project by these dates:

March 29, 2021: Preprocessing, data cleaning and transformation completed.

April 5, 2021: Data integration complete and sentiment analysis has been tested on smaller subsets of review text.

April 16, 2021: Jupyter notebooks complete for restaurant review questions. Progress report complete.

April 23, 2021: Jupyter notebooks complete for user review questions.

April 29, 2021: Presentation slides and final report complete.

MILESTONES COMPLETED

We have completed the preprocessing and data cleaning milestones for our dataset. Each of the initial five json files have been cleaned in different ways which has resulted in a new set of five cleaned json files.

The initial json file containing business data was filtered to only contain businesses in Las Vegas, then filtered again to only contain restaurant businesses and food related vendors.

The initial reviews json file containing review data was filtered to only contain text reviews for our filtered restaurants in Las Vegas. Furthermore, 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.

After we cleaned and pre-processed our five initial json files, we integrated the data into an Sqlite relational database. This allows us to access all the cleaned data from a single source during mining.

MILESTONES TODO

April 20, 2021: Jupyter notebooks complete for user review questions. This includes clustering of reviewers and geopandas visualizations. Sentiment analysis tested on smaller subsets of review text.

April 26, 2021: Jupyter notebooks complete for restaurant review questions. This includes review sentiment related questions, such as seasonal changes in review sentiment. Work on a draft of the final project report.

April 28, 2021: Draft of final project report complete. Team meeting to discuss our findings and further refine the final report.

April 30, 2021: Presentation slides and final report complete. Review report and slides for corrections.

May 1, 2021: Submit the presentation slides and final report. Ensure all source code exists on the project github page and the README follows the guidelines listed in the project guidelines.

RESULTS SO FAR

1. Apriori pruning business categories:

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 be found to 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:

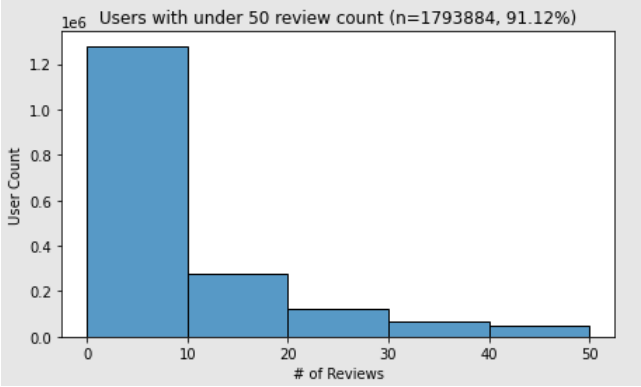
Association rules (High to Low Lift)

	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 that passed, we expect cafes to be associated with breakfast, and sports bars to be associated with American food.

2. Review count frequency of users:

Initially, we planned on separating Yelp users by their review counts into categories, such as: low, infrequent, or frequent review count. The number of reviews across our dataset was graphed. This revealed that the vast majority of all users (~90%) only submit between 0 and 10 total reviews:



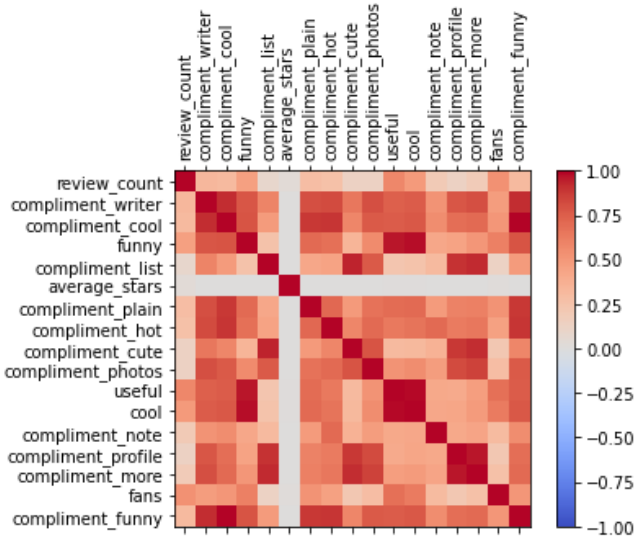
Since review counts are so heavily skewed, we need to change our approach to categorizing users based on their review counts.

As an alternative, we may perform a logarithmic transformation and then categorize users based on the number of standard deviations from the mean. Using these user categories, we can then move on to cluster analysis in order to determine if low (<25%), medium (25% to 75%), or high (>75%) review counts cluster by street, star review, or restaurant category attributes.

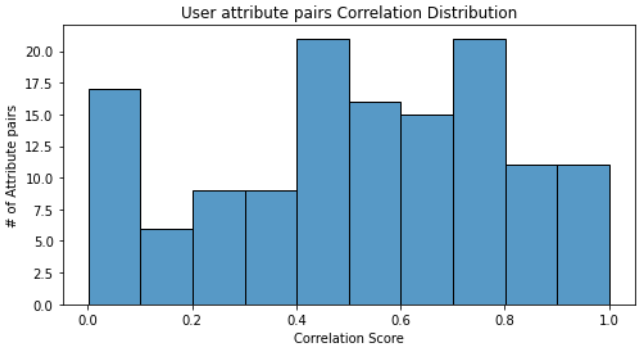
3. Reducing redundant user attributes

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”.

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/

[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