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. 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, Mexican, 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. Do the frequent features described in the first question vary between 2 major cities?
3. Are there cities where restaurants are more likely or less likely to receive negative review sentiment?
4. Can seasonal changes improve or impair restaurant review sentiment? If so, which categories 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: Reviews for Italian restaurants with 4-5 stars focus on authenticity and service. In Phoenix, AZ Italian restaurants receive more negative review sentiment during the Fall and Winter than the Summer. While in Las Vegas, negative reviews are invariant to seasons. And review sentiment is focused on atmosphere and service

2 Yelp User Reviews:

These reviews are generated by Yelp users who post restaurant reviews infrequently ($n < 50$), frequently ($n < 1000$) and very frequently ($n < 6000$). In this set of reviews, we aim to identify common features among different types of users. We will focus on two questions.

1. Do infrequent, frequent, and very frequent reviewers cluster around certain attributes such as geographic locations, number of fans, average star review, or useful tags given?
2. How does review sentiment vary among infrequent, frequent, and very frequent reviewers?

In answering these questions, we hope to discover patterns among different types of Yelp users that may be beneficial to restaurant businesses. For example, if a certain area has clusters of Yelp users who review very frequently with negative sentiment, then it may be beneficial to avoid opening in that area. Conversely, if an area has a cluster of positive and frequent Yelpers. Then investments made towards

Yelp could prove to be worthwhile, i.e. focus on key customer preferences or Yelp partnerships.

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 have 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 similar to those that we will need to do, 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 from the words used in reviews.

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 tokenized.

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 cities 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 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 and review sentiment analysis.

DATA SET

The Yelp Open Dataset contains over 8 million user reviews on over 160 thousand businesses in 8 metropolitan areas. The raw data consists of 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. These are paired with dates, business identification and user identification.

5. yelp_academic_dataset_user.json

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

EVALUATION METHODS

We will use two methods when evaluating our results.

First, we will look at correlations between our findings and restaurants that have already closed. This will allow us to gauge the degree of negativity found in our restaurant profiles since we expect closed restaurants to, on average, have a high level of negative sentiment.

<Remember training vs testing subsets from zoom>

Second, we will split our reviews into 80/20 training and validation subsets. This split would allow us to compare if our frequent features found in the training set are held in the validation subset. Similarly, this evaluation approach will also be used for Users.

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.*

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 fiinal report complete.

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