Google
15,516 restaurants
2,723 parking spots

Yelp
9,100 restaurants

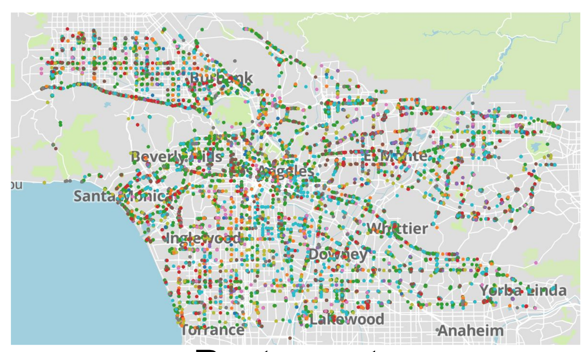
Data Driven Business Decisions for New Restaurants Using Heterogeneous Datasets

Barkha Bhojak, Dweep Kumarbhai Trivedi, Supicha Phadungsilp, Utkarsh Gera, Zeeshan Abid
University of Southern California: CSCI 599 - Data Science for Social Systems

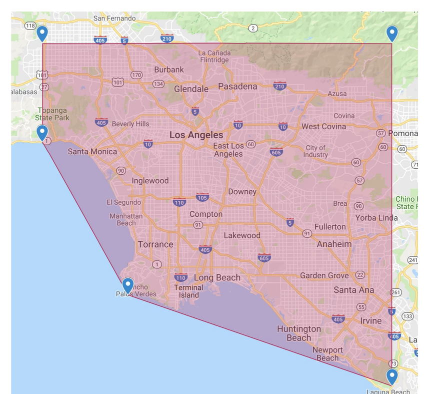
USC Viterbi School of Engineering

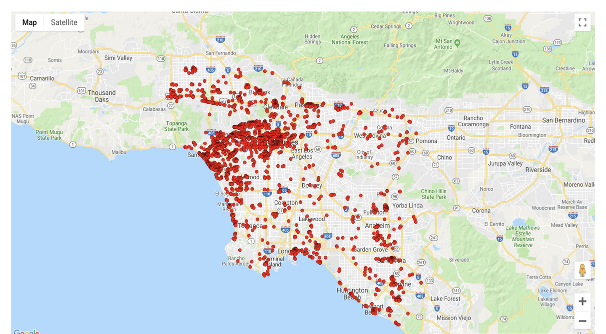
Abstract

New restaurant business owners have numerous factors to consider before making important decisions. With the rise in the number of social media platforms, large amounts of user-driven data are being generated on a daily business. We aim to help new entrepreneurs capitalize on this data before undertaking a venture. Utilizing data from various different sources our objective is to predict a list of cuisines that are likely to receive a high rating on Yelp at a given location in Los Angeles.



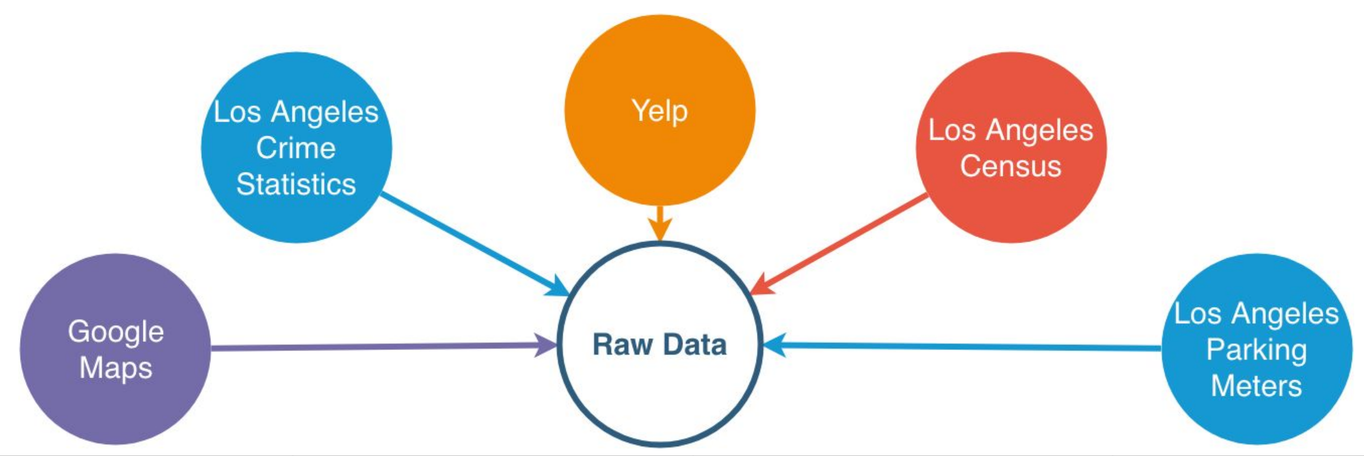
Restaurants



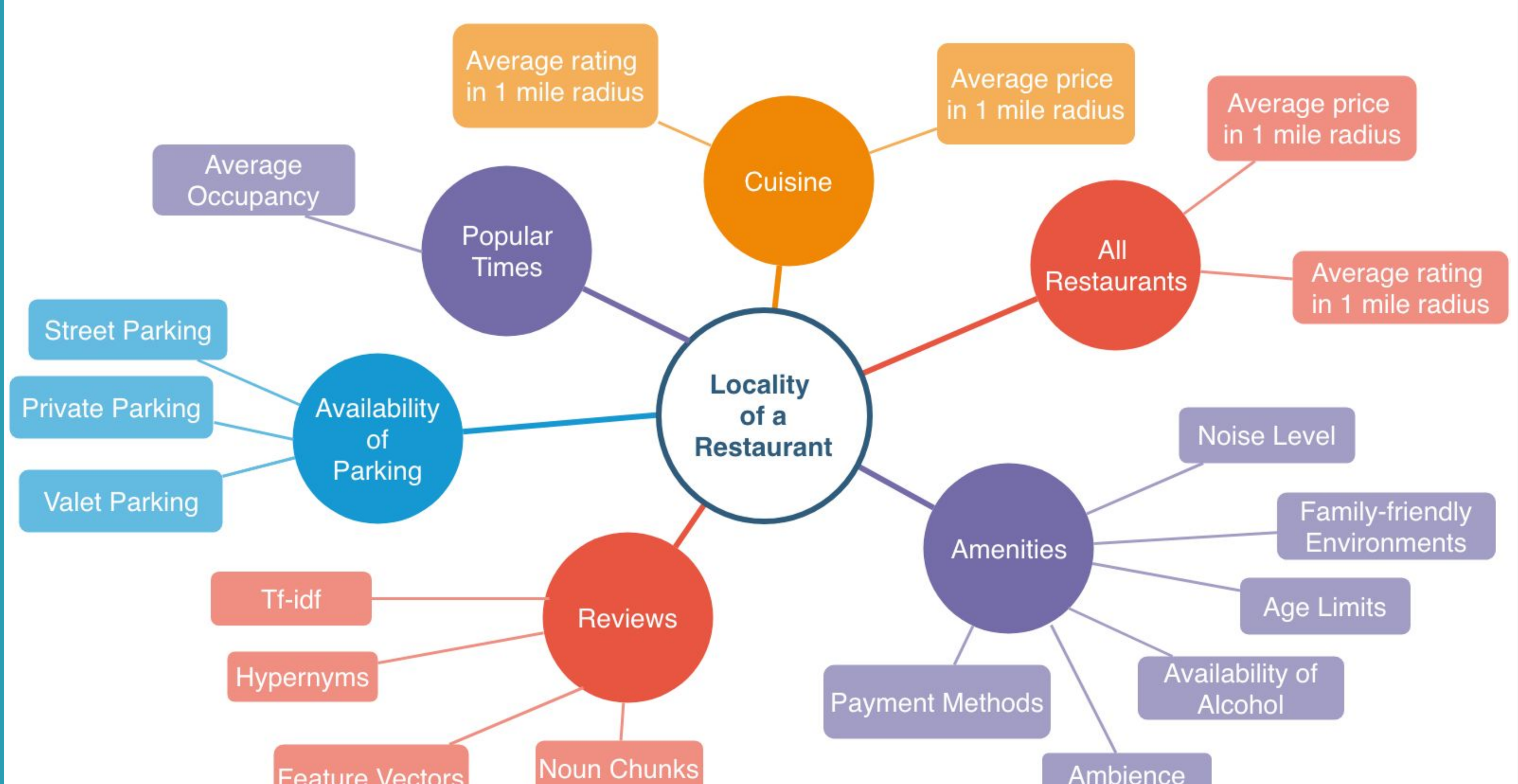


Parking

Data Collection and Feature Extraction

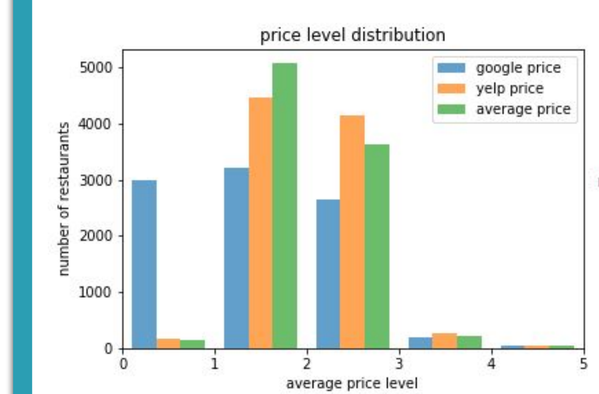


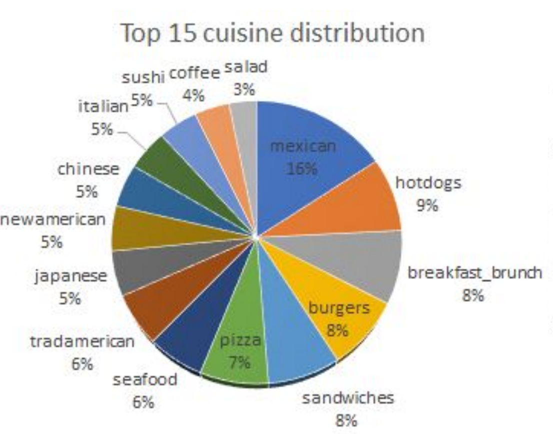
Google Maps	Yelp	UI Crawler for Google	UI Crawler for Yelp
Rating	Rating	URL	Yelp ID
Price Level	Price	Popular Times	Review Text
Place ID	Location	Place ID	Amenities
Address	Name	Reviews	Posted Date
Phone Number	Phone Number	Amenities	
Name	Review Count		

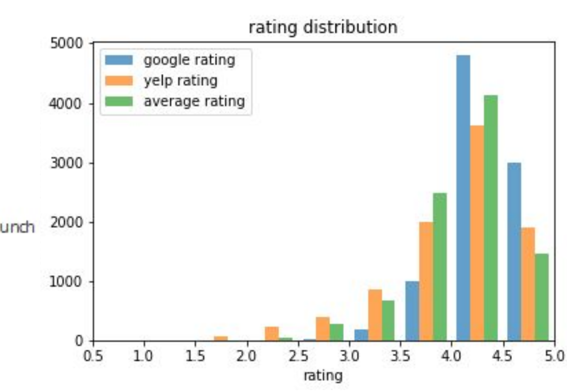


Interesting Facts

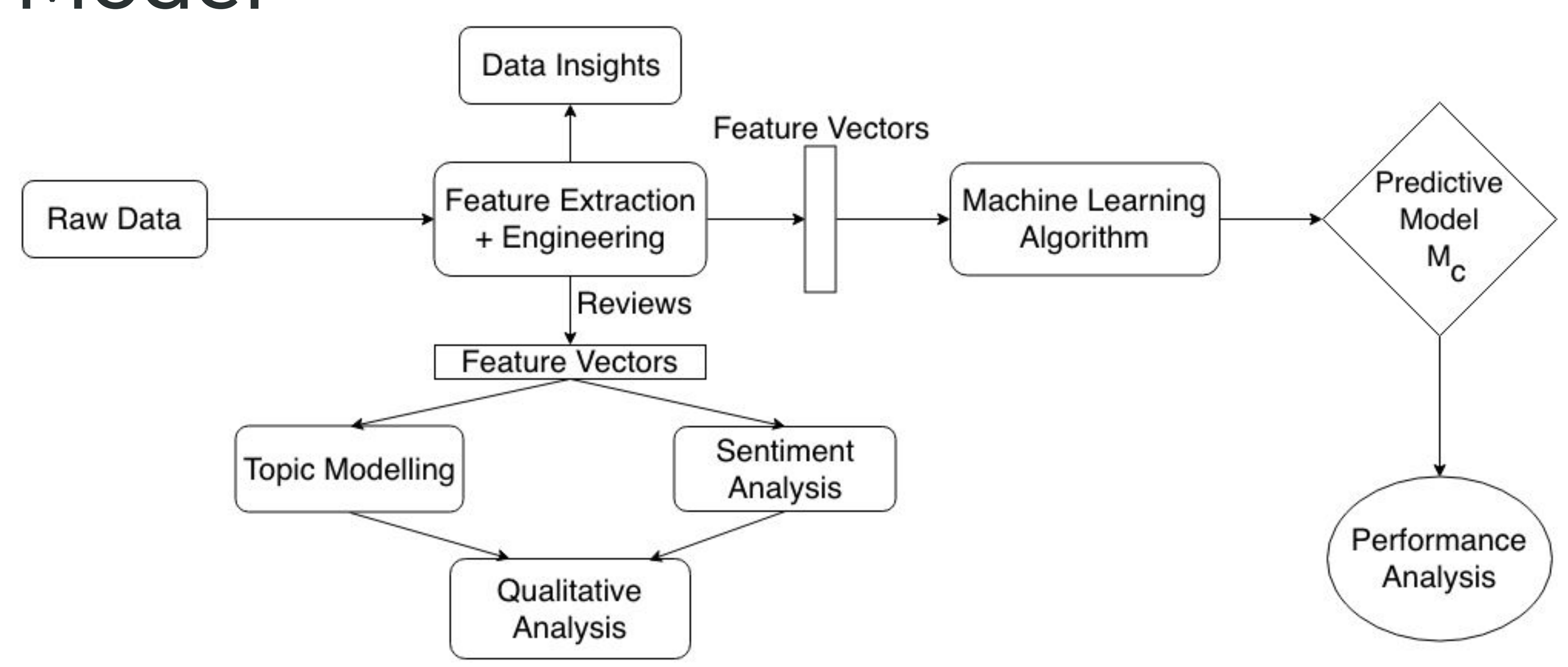
- Most Common Cuisine: Mexican
- Most Correlated Amenity: Street Parking
- Most Correlated Cuisine: Hot dogs
- Most Correlated Feature: Average rating of nearby restaurants
- Highest Average Rated Cuisine: Vegan, Tacos, Mediterranean
- Lowest Average Rated Cuisine: Hot Dogs, Chicken Wings, Burgers
- Cuisine with the Highest Average Price Level: Steak, Bars, Sushi
- Cuisine with the Lowest Average Price Level: Hot dogs, Tacos, Mexican







Model



Predictive Model

Input: Coordinates - latitude, longitude

Output: Top 3 cuisines with the highest rating

Algorithm:

1. *Generate geolocalized features*
2. *For each **cuisine**:*
   a. *For each **price level**:*
      i. *Predict rating using* M_c
3. *Return **top 3** cuisines and price level for a location based on ratings*

Qualitative Analysis

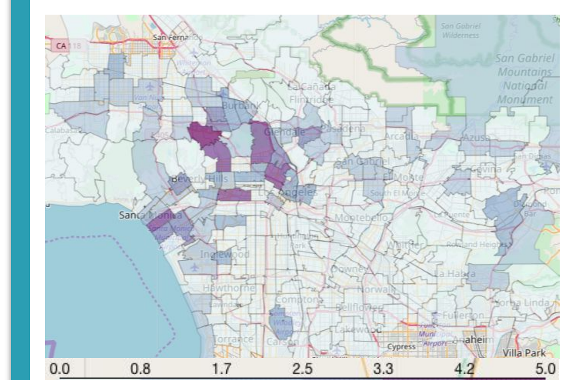
To understand the quality of food, service, and ambiance of each restaurant, the following methods were used to extract the above topics from reviews:

- Method 1: Topic modeling on each review to extract topics using LDA algorithm.
- Method 2: Topic modeling on each sentence of each review to extract topics using LDA Multicore algorithm.
- Method 3: Generate hypernyms of each word to get the topic of each sentence using NLTK.
- Method 4: Generate feature vectors for noun chunks and the context of the chunks to train a LSTM network where the labels are topics which were generated using NLTK.

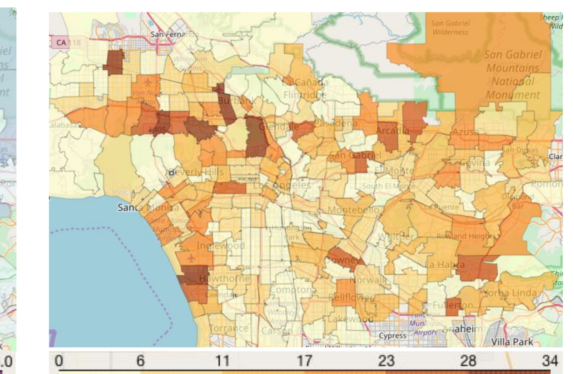
Once the sentence/review had been classified, VADER scores were calculated to get the compound sentiment for each class.

Analysis

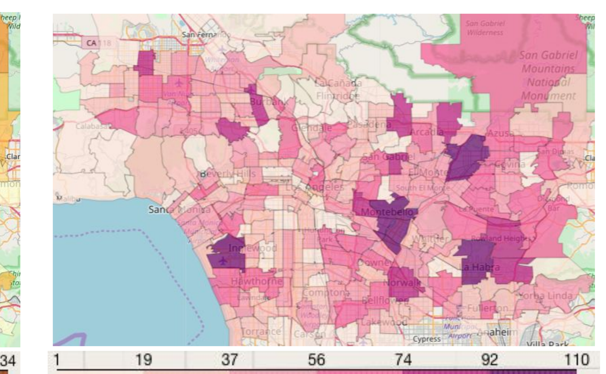
Distribution of vegan restaurants



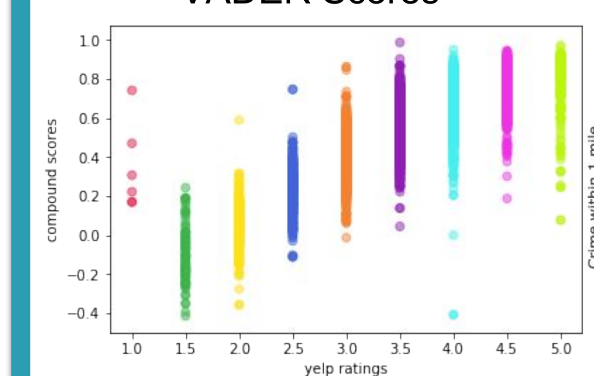
Distribution of family-friendly restaurants



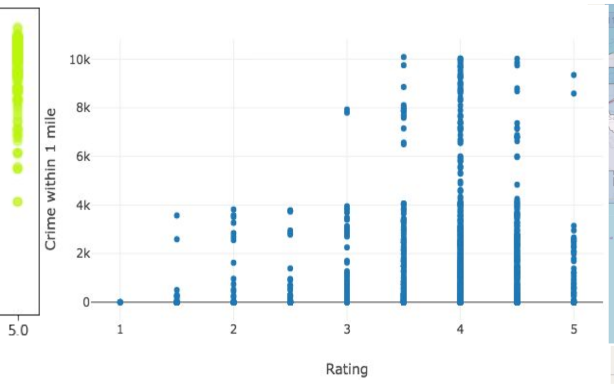
Distribution of all restaurants



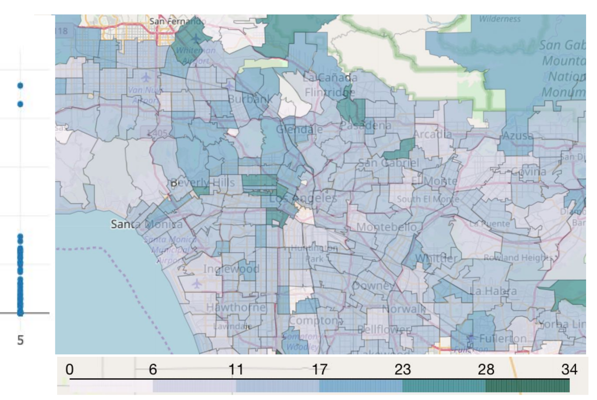
VADER Scores



Crime vs Rating



Average Popular Times



Performance Analysis

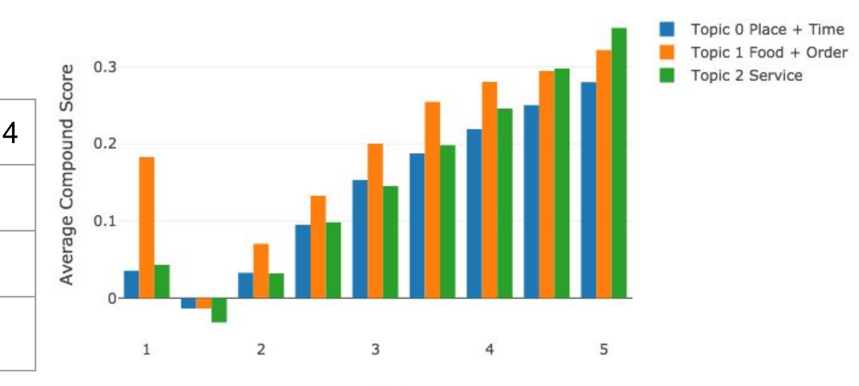
The size of the test dataset is 910. The number of restaurants that:

- are currently performing at their best: 10
- can increase their rating by at least 0.5 by changing the cuisine: 77
- can increase their rating by at least 0.5 by reducing the price: 31
- can increase their rating by at least 0.5 by introducing valet parking: 28

Qualitative Analysis

Topic	Method 1	Method 2	Method 3 & 4
Food	Order, Food	Food, Order	Food
Service	-	Time, Service	Service
Ambience	Place, Food	Place	Ambience

Topic Modeling on Reviews



VADER score of each topic vs rating

Results

Method	MAE	R^2 Score	MAE for restaurants with rating less than 3
Ridge Regression	0.2771	0.4086	0.6177
Support Vector Regression	0.2690	0.4345	0.5231
Gradient Boosting	0.2895	0.3685	0.8005
Adaptive Boosting	0.2836	0.3749	0.6717
Random Forests	0.2681	0.4393	0.53
Artificial Neural Networks	0.30	0.3059	0.3943

Regression

Method	F1 Score
Support Vector Machines	0.9705
Decision Tree	0.9647
AdaBoost	0.9683
Random Forests	0.9686
Artificial Neural Networks	0.9683

Classification