Yelp Dataset Analysis

Chetan Surana
crajende@asu.edu

Divya Kshatriya
dkshatri@asu.edu

Magesh Sridhar
msridh12@asu.edu

Sagar Parekh
sjparekh@asu.edu

ABSTRACT
The majority of the restaurant search websites have a lot of data that is not displayed in a manner that a user can apprehend. Our goal with this project is to provide the user as many insights about a restaurant as possible so that the user can accurately pick their go-to restaurants. All the analysis is visualized, so the user can quickly judge a restaurant.

KEYWORDS
Data Visualizations, Pie chart, Bubble chart, Spider graph, Line graph, Recommendation system, Sentiment Analysis, Web Scraping, Data Mining, Yelp Dataset Challenge

1 Introduction

Yelp is currently the most widely used restaurant and business information software across the United States. Yelp provides an overview about the restaurant like the overall ratings, few reviews out of thousands of reviews and other basic information. To improve yelp users' experience, we decided to do an in-depth analysis of the yelp dataset[7]. The data can be analyzed and visualized from the business owner's perspective, the user's perspective and the reviewer's perspective. We chose to explore the user perspective so we investigated what information a user needs to choose a restaurant and how to best present it visually. Our project gives users' insights about: restaurants in the neighbourhood, available cuisines, average rating of a restaurant, rating over a period of time, busy hours at a restaurant, sentiments analysis of all the reviews for a restaurant, popular opinion word list for a restaurant and a recommendation system for similar restaurants.

2 Motivation

According to [10], there are more than 1 million restaurants chains in the US right now. The restaurant industry was expected to add 4.9% new restaurants in 2019 and this rate is only going to increase. Some other interesting facts are: 90% of guests check out a restaurant online before going, 33% of people read other guests' reviews before selecting a place to eat and a one-star increase on Yelp can boost a restaurant's profits by up to 9% [11]. These statistics inspired us to build a visual recommendation system which could provide as many insights about a restaurant as possible so the user can make a well informed decision. User can pick a restaurant of his desired cuisine, distance he is willing to commute, time he is ready to wait for a table, and many such insights which the system can offer him in a visually appealing way.

3 Data Collection and Analysis

The dataset used is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge. Yelp Dataset (https://www.yelp.com/dataset) has over 1.2 million business attributes for each of 192,609 businesses. We limit our dataset to Tempe for easier and more precise analysis of the data. The dataset is cleaned to remove all the other businesses other than the restaurants in Tempe. Specifically, the dataset comprises of 5 files: one for each object type: business, review, user, check-in and tip.Each file is a json object. The business json file comprises of attributes like business id, name, location, stars, categories, etc. The categories attribute had different keywords that represented the restaurant. Based on the collective category keywords, we created a dictionary of valid cuisines. Each keyword was assigned to a particular cuisine. Later, the restaurants were categorized based on the cuisines present in the dictionary. The cuisines are used for creating a bubble chart and filtering the restaurants in the map. A text review is a json object in the 'review.json'file, which specifies the business ID, user ID, stars (integer values between and including 1 and 5), review text, date and votes. A total of 1044 restaurants were used for our analysis based on the location Tempe.

4 Visualization Design Implementation

4.1 Overall Design

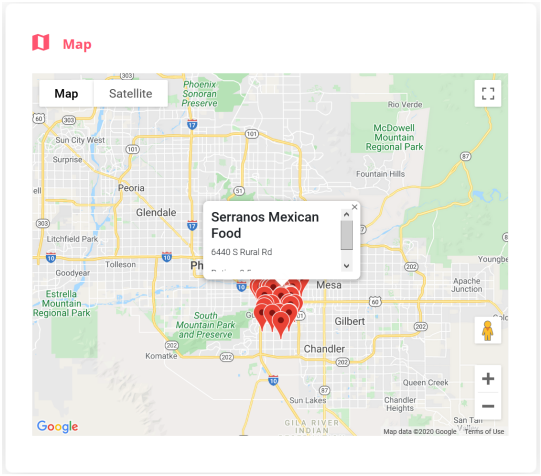
The system is designed with material design principles, which is inspired by material surfaces, lights and shadows. The background is a light gray, with each visualization presented on a white card, which gives the perception of shadows and depth. The overall effect is a clean, intuitive interaction. Each visualization has a clear title. There is a legend describing colors and symbols wherever required. There is a unified color scheme that ties in the components. We used Open Sans font, which is an open source font. We referred HappyHues[12], which provides curated color schemes, based on context. The titles are red, to excite and attract attention. Visualization lines are green for most part, as green is a happy, soothing, healthy hue. Bright colors on a white backdrop blend well.

The visualizations are interactive, to let the user have control. Tooltips appear where required to give additional information. The first page presents the user with a search bar, a map with restaurant markers, and a bubble chart of cuisines. The search bar is equipped with an autocomplete feature that suggests queries in a dropdown based on the partially typed query. This was enabled with the help Awesomeplete, a simple, customizable and lightweight autocomplete widget. The second page presents ratings spider chart, ratings time series, popular opinion word summary, review distribution donut chart, check-in heatmap, and similar restaurants recommendation.

4.2 Map

Using Google Maps Plugin API from Google Cloud Platform, we integrated a map plugin that displays markers for restaurants in the vicinity. The markers are placed by accessing the latitude and longitude property values for each business in the business.json file. The user can then select a restaurant interactively from the map. The markers can be filtered by cuisine by clicking on a cuisine bubble from the bubble chart alongside. On hovering over the map markers, a tooltip with the restaurant name, address and rating is presented. The map aids the user in making a decision based on the rating and distance preference.

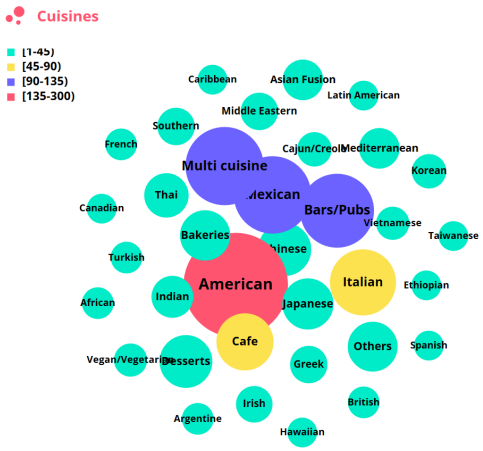
Figure 1: Google Maps



4.3 Bubble Chart

A bubble chart is used to visually represent the diverse cuisine options available in the user's location, with the size of the bubble encoding the number of such restaurants in the neighborhood. The bubble chart uses the categories attribute in the business.json file to get the cuisine. Our visual recommendation system incorporates this visual cuisine selection bubble chart as an intuitive filter into the decision making process of choosing a restaurant. Clicking on a bubble filters the results on the map to show restaurants belonging to the selected cuisine. Also, the selected bubble is highlighted and the other bubbles are faded.

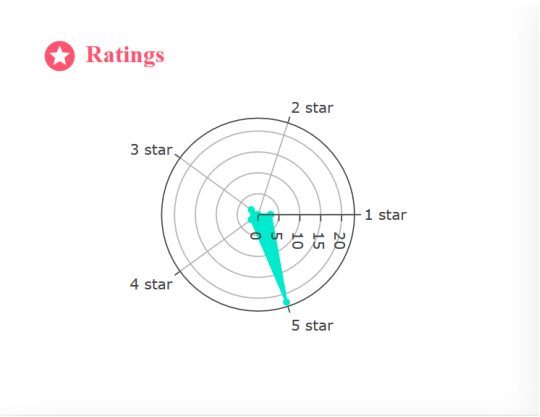
Figure 2: Bubble Chart



4.4 Ratings Spider Chart

A spider chart, or radar chart, of the ratings of a restaurant is created using the restaurant ratings in review.json file. The circle is divided into five spokes or radii, each representing one, two, three, four and five star rating. The spoke length is proportional to the number of reviews with the corresponding rating. The X-axis scales automatically, based on the total count for the restaurant. This chart gives the user an idea about the distribution of number of each star rating. It is a visually pleasing and less space consuming alternative to the simple bar chart or histogram.

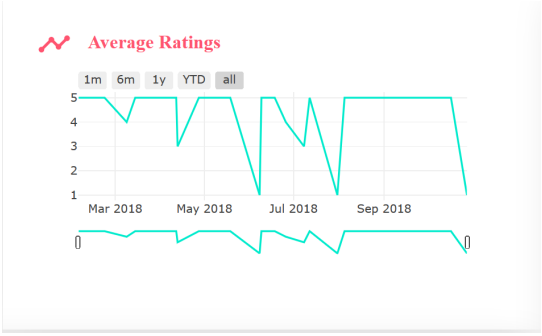
Figure 3: Ratings Spider Chart



4.5 Time Series

The reviews.json file has a rating and date property for each review posted by users. Using PlotLy and D3, A time series plot of the restaurants ratings is displayed over a time window. The ratings are obtained from With buttons, the user can select the time frame scale as one month, six months, one year or entire history of existence. The Y-axis is scaled from 1 to 5 to represent the restaurant rating and the X-axis is scaled based on the window frame scale chosen. An interactive slider allows the user to scroll over time. The user can then visualize the restaurant's rating trend during the selected time frame. A restaurant may have been around for a long time and a simple average rating may not adequately represent how well a particular restaurant is doing. This visualization helps the user to dive in and see the trend in a restaurant's rating.

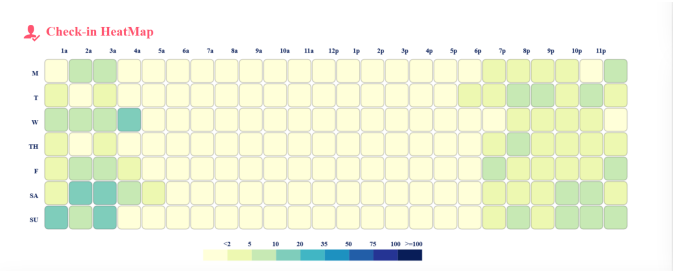
Figure 4: Time Series



4.6 Heat Map

The heat map visualization presents 2D grid of boxes for each of the seven days of the week and 24 hours of each day. Each box is filled with a color that quantifies the number of user check-ins for that time of the day. A python script processes the user checkins and gives a count for each hour for each day for every business. The checkin.json data is first processed to transform the time checkin values to day-hour frequency counts for visualization. The processed data is then read in and bound to svg rect elements using D3. The count is mapped to a color on the Yellow-Green-Blue color scheme. This color scheme was chosen so that the eye can quickly distinguish between few, moderate and large number of checkins. A tooltip with the exact count on hover over the grid is also presented. This helps achieve our goal, which is to give the user an idea about the restaurant's busy hours and probable waiting time in the shortest space, with the least ink and minimal clutter.

Figure 5: Heat Map



4.7 Popular Opinion Word List

This section presents the top five negative and positive words that occur in the reviews for the selected restaurant. This is done by sentiment analysis on the reviews.json file. Firstly, text preprocessing steps are applied on each review before calculating the sentiment. Stop words like 'a', 'an', 'the', 'of' and others are removed. Words are lemmatized so that all words in the simplest form and tense. Afinn Sentiment Lexicon is then used to match and extract positive and negative words from each review. Afinn Sentiment Lexicon is a dictionary of English words that are manually rated with a value between -5 (negative) and +5 (positive). The five most frequently occurring positive and negative words are listed. This gives the user an idea about the popular opinion.

Figure 6: Frequent Words

Sentiment Analysis

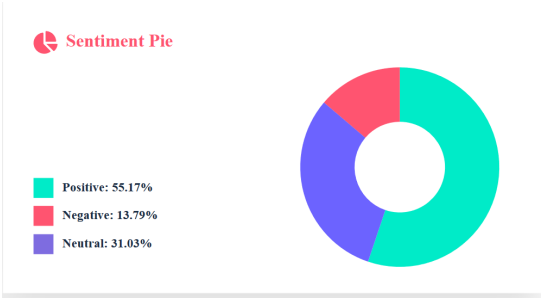
Negative	Positive
stop	great
disappoint	good
miss	best
worst	love
bomb	super

4.8 Donut chart

The total sentiment score for each review for a restaurant in review.json file is calculated using the Afinn Sentiment Lexicon. If the total score is greater than zero, the review is considered positive. It the score is less than zero, the review is classified as negative. The review is considered neutral if the total sentiment score is zero. In this manner, all reviews for a business are categorized as positive, negative and neutral

reviews. A donut chart is created to visually represent the distribution of positive, negative and neutral reviews. For this, the D3 pie layout is used. The pie area is proportional to the count. With this informed summary, the user has an idea about the visitors' opinion about the restaurant. The colors chosen are red, blue and green for negative, neutral and positive reviews, respectively. The colors themselves implicitly convey the sentiment, as red is generally perceived as danger or negative, blue as neutral or calm, and green as positive or healthy.

Figure 7: Donut Chart



4.9 Recommendation System

We developed an algorithm to recommend restaurants similar to the selected restaurant so that the user can explore multiple similar options and choose the restaurant that is most to his liking. We recommend three restaurants in the same category, with same or higher rating and distance within three miles. Since the business.json data has latitude and longitude values for each restaurant, we calculated distances between the restaurants using the latitude and longitude values by applying the Haversine Distance Formula:

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + \cos\phi_1 cos\phi_2 \sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

Figure 8: Recommendations

4.10 Technologies Used

The following technologies were used in this project:

- D3.js (Data Driven Documents) – An open source JavaScript library to create interactive data visualizations for web browsers.
- Python – A high level general processing language for scripting, pre-processing, etc.
- Prototyping - Adobe XD (Experience Design).
- User Interface – The UI has been designed using HTML 5, CSS and JavaScript.

5 Methodology

The data of restaurants which fall in a neighborhood is computed by checking the geocoordinates of each restaurant if they are within the radius of the circle. This data is used to further analyze the top cuisines and the ratings in that neighborhood.

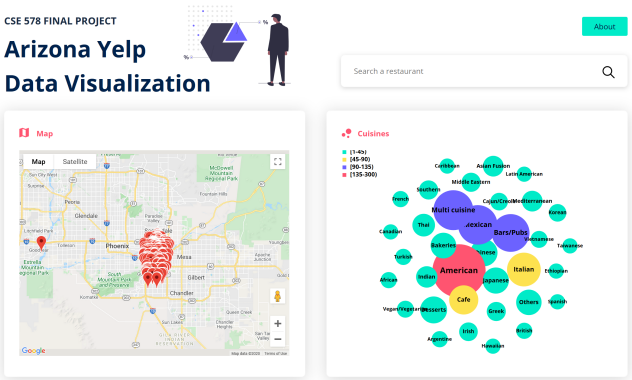
5.1 Neighbourhood Analysis

5.1.1 Search Bar If some users aren't comfortable with using interactive visual navigation the search bar is paramount for them. The search bar aids the user to directly look up any restaurant he wishes. It also saves time by auto-complete suggestions. It can be seen on the top right side of Fig. 9.

5.1.2 Bubble Chart Cuisine is the driving factor in the decision making process. So our system analyzes the top cuisines in the neighbourhood and displays diverse cuisine options in the form of a bubble chart as seen on the right side in Fig. 9. Looking at the size of the bubbles, the user gets an idea about the number of such restaurants in his neighbourhood. Once a cuisine's bubble is selected, the restaurants in the map overview get filtered which in turn makes the decision making process easier.

5.1.3 Map The Map shows all the restaurants in the neighbourhood [Fig. 9.]. This gives an insight to the user about the restaurants close to him and their ratings in a tooltip on hover over the marker. After filtering by any cuisine the map only shows restaurants of that cuisine. Decision can now be made considerably easily. Once the user selects a restaurant he is directed to the restaurant's specific page.

Figure 9: Page 1



5.2 Restaurant Performance Analysis

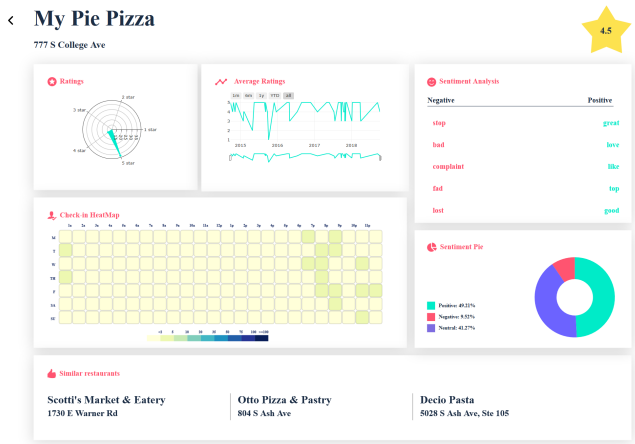
The restaurant specific page tries to convey all the relevant information about the restaurant in a visually appealing way to the user. The restaurant's name and address is the first thing that user sees. On the top right side the overall rating of the restaurant is presented.

5.2.1 Rating Distribution The radar chart shows the cumulative distribution of all the ratings received by the restaurant over the years. The user can understand how popular and trending the restaurant has been over the years. Here the filled quadrant between two ratings indicates how the restaurant is doing.

5.2.2 Time Series Analysis This line chart shows how the restaurant's rating has varied over the years based on customer's rating. It can be used to understand the highs and lows of the restaurant. The user can select to view the ratings of past 1 month, 6 months, 1 year, Year-to-date or all the time from the selector button available. The user also has the option to use the range slider to select the time range of which he wishes to see the rating trend of the restaurant.

5.2.3 Check-In Analysis The Heatmap visualization shows the total number of check-ins made on each day of the week. This gives the user insight about the restaurant's busy hours which could help the user to understand how much wait time would be there at the restaurant. With this insight the user can plan ahead and manage his time more efficiently. A color scheme of Yellow-Green-Blue is used for easy interpretation of number of check-ins where yellow is low, green is medium and blue is high number of check-ins.

Figure 10: Page 2



5.3 Customer Sentiment Analysis

Sentiment analysis is performed on all the reviews received by the restaurant.

5.3.1 Overall Sentiment Analysis A donut chart is presented to visually represent the distribution of positive, neutral and negative reviews. This gives the user insight about the generalized feeling of all the reviewers towards the restaurant. Also the five most frequently occurring positive and negative words are listed to give the user an insight about what most people are talking about in their reviews without actually going through hundreds of reviews.

5.4 Recommendation

The user is recommended restaurants based on the cuisine, distance and rating. This helps the user to look for similar options easily without actually going through the entire process all over again.

6 Evaluation Plan

There are multiple objectives that need to be validated. First, the system needs to be tested for errors. Different queries and interactions must be input to test robustness and correctness of the visualizations. The recommendations generated also need to be evaluated for relevance. To evaluate the overall effectiveness of the system as a visual recommender system for selection of restaurants, we can survey users. The users will be asked to interact with the system and rate it on several criteria on a scale from one to ten. The criteria include ease of use, effectiveness, intuitiveness, completeness, and visual appeal.

7 Future Work

The analysis and related visualizations are done on Tempe dataset only. This could be extended all over the United States. The recommendation system created is a very minimal system to help the user search similar restaurants based on cuisine and higher ratings. This recommendation system can be more powerful and complex for more precise recommendation if done based on user profiles. The following use cases can be created in future :

Use Case 1

The analysis of the dataset can be extended to business owner's perspective. The visualizations can help the business owners to understand how their businesses are performing. Added to that, a recommendation system that suggests what can be done to improve the business could be incorporated to this.

Use Case 2

The user login and sign up could be added to get user profiles. Using the user profile details like the restaurants searched often, the cuisines liked, etc , a stronger recommendation

system can be created to precisely provide the similar restaurants.

Use Case 3

The sentiment analysis used here is only to provide the insights about the restaurant. Based on the analysis of the reviews of the restaurant, the visualization incorporates the pie chart. Added to this, based on user activity, the current mood of the user can be predicted which could be used to precisely select his/her next restaurant. For example : a user in the state of sadness would be prefer a cold desert.

8 References

[1] Sentiment Analysis:
https://itnext.io/functional-sentiment-analysis-in-javascript-754f58628746

[2] https://d3js.org/

[3] Donut Chart :
https://www.d3-graph-gallery.com/graph/donut_basic.html

[4] https://jquery.com/

[5] https://getbootstrap.com/

[7] Dataset:
https://www.yelp.com/dataset

[8] Map:
https://developers.google.com/maps/documentation/javascript/tutorial.

[9] Search Auto-completion:
https://leaverou.github.io/awesomplete/

[10] https://www.ibisworld.com/industry-statistics/number-of-businesses/chain-restaurants-united-states/

[11]https://www.smallbizgenius.net/by-the-numbers/restaurant-industry-statistics/gref

[12]https://www.happyhues.co