Yelp Business Insights: Visualization of the Restaurant Growth Trends using Yelp Dataset

Aravind Sundaresan
ASU ID - 1214862831
asunda23@asu.edu

Pallavi Balasaheb Kamble
ASU ID - 1215159595
pbkamble@asu.edu

Parth Rajendra Doshi
ASU ID - 1215200012
pdoshi4@asu.edu

Pravachan Thumati
ASU ID - 1215180161
pthumati@asu.edu

Kabini Salim Kumar
ASU ID - 1215098534
ksalimku@asu.edu

ABSTRACT

This project on Yelp Business Insights provides insights with regards to the growth of a restaurant, from the perspective of a business owner. The aim of this dashboard is to help owners estimate the performance of their business on the basis of 3 important attributes - neighborhood analysis, restaurant performance analysis and customer sentiment analysis. The dashboard consists of various interactive visualizations that depict key statistics pertaining to the 3 topics mentioned above. Sentiment analysis is performed on the customer reviews to help understand if customers are happy with the restaurant or not. By analyzing these metrics, the restaurant owner can figure out ways to enhance the performance of his business.

Keywords

Data Visualization, Yelp, Sentiment Analysis, Business Insights, Neighborhood Analysis

1. INTRODUCTION

The restaurant industry is really competitive with owners constantly trying to improve their businesses to stay ahead of the competition. The growth of a restaurant is determined by various factors ranging from key aspects like the quality of food to peripheral aspects such as restaurant location, ambience, etc. With the growth of business review websites such as Yelp and Zomato, the restaurant owners have large silos of invaluable information obtained directly from their customers. In this project, we have made use of the dataset provided by Yelp as part of their Dataset Challenge to analyze the performance of businesses specifically in the state of Arizona. The following research questions were answered in this project:

- What is going on in my neighborhood?
  - Number of similarly rated restaurants nearby
  - Trending cuisines
- How is my restaurant doing?
  - Number of check-ins per day of the week
  - Number of reviews per year
  - Star ratings over the years
- Are the customers happy with my restaurant?
  - Sentiment Analysis of customer reviews
  - Identification of top positive and negative aspects of the restaurant

Using our dashboard, restaurant owners can estimate not only how well their business is doing but also how well their competitors in the neighborhood are performing. The use of sentiment analysis also helps them understand which aspects of restaurants are well-received by customers and the ones which need to be improved upon. All these insights can in turn help businesses make changes that can accelerate their growth.

2. MOTIVATION

According to [1], there are more than 1 million restaurants in the U.S. right now. The industry adds about 10,000 units a year and this rate is only going to increase. People view restaurants as not just a place to eat but also a place where they can spend quality time with friends and family. To run a restaurant successfully, the owner must figure out ways to ensure that the customers have a good experience when they visit the restaurant while also sustaining profits. The owner must keep track of how the restaurant preferences of customers (such as cuisine) are changing in the neighborhood while also monitoring how their competitors are performing. Our dashboard can thus be used to view how the restaurant has grown over the years. The analysis of customer sentiment can be used to understand the aspects of the restaurant that please customers and the ones that don’t. With this knowledge, the owners can work towards enhancing the growth of their business by increasing profit margins while also ensuring customer satisfaction.

3. RELATED WORK

As part of the Yelp dataset challenge a lot of work has been using this dataset. The work done by Sindhu Hegde et. al. [4] have analyzed the Yelp academic dataset to determine the ideal way to run a business in order to improve the business. They have made use of attributes such as the location and crowd related data to determine how to setup an efficient business. The work done by Boya Yu et. al. [2] makes use of sentiment analysis to highlight the key features of a restaurant. We have used a similar approach to identify the aspects of a restaurant that pleases or displeases customers the most. In our project we have worked towards answering the 3 questions mentioned above. This in turn can help a business owner ensure that his restaurant has a positive growth trend.

4. VISUALIZATION DESIGN

4.1 Dataset

The dataset for this project was obtained from Yelp. Altogether, this dataset consists of 6,685,900 reviews for 192,609 businesses. The data, which was available in JSON format was first converted to CSV format for this project. For the purpose of our project, this dataset was reduced to include only the data pertaining to restaurants present in the state of Arizona. The filtered dataset consists of 1,170,517 reviews for 11473 restaurants. For each of these restaurants, the average rating and count of reviews were computed for every year the restaurant was in operation. Sentiment Analysis was performed on these reviews using the Stanford CoreNLP library and each review was assigned a label ("1" indicating Negative, "2" indicating Neutral and "3" indicating Positive) to denote the sentiment associated with the review.

4.2 Technologies Used

The following technologies were used to implement this project:

- D3.js (Data Driven Documents) – An open source JavaScript library to create interactive data visualizations for web browsers.
- Leaflet – An open source JavaScript library for interactive maps.
- Stanford CoreNLP – A Java library that provides a set of human language technology tools to perform Natural Language Processing.
- User Interface – The UI has been designed using HTML 5, CSS and JavaScript.

4.3 Map

The map explores the Yelp business data to identify how the businesses and their competitors are performing in the selected neighborhood. The maps have the potential to show the complex data and makes it easier to understand the activity of relevant local businesses across the neighborhood. It gives clear and concise view of data and visualizing data spatially allows the business user to derive insights on the distribution of the local competitors. On selection of a restaurant on the map, the neighborhood of that restaurant is highlighted in a circle. The information of the selected neighborhood is further analyzed and displayed using two charts, star ratings of the neighborhood business and the count of restaurants in top ten cuisines, which gives insight on the performance of a business with respect to its competitors in a given area. Leaflet is one of the widely used JS library to build maps. It is interactive, flexible and lightweight and has many plugins available.

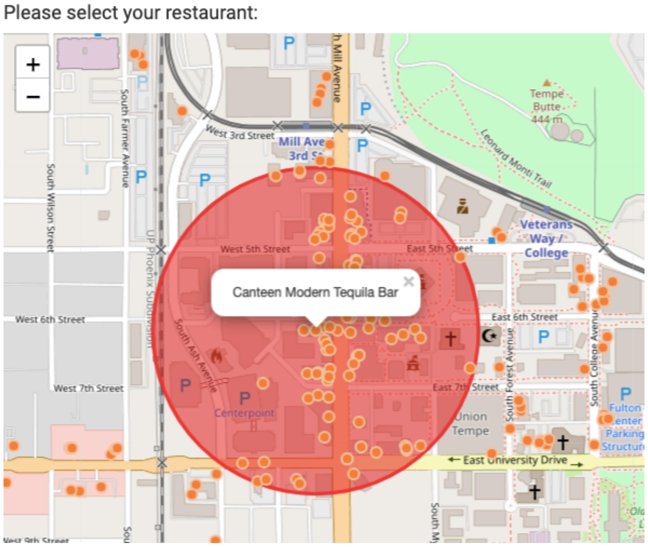
Please select your restaurant:

Figure 1. Map input to select a restaurant

4.4 Neighborhood Analysis

Neighborhood analysis involves the analysis of other restaurants that lie in the vicinity of the given restaurant within a 0.3 mile radius. The following charts display the information based on the selection of a restaurant on the map:

- Star rating distribution of neighborhood businesses
- Count of restaurants for each of the top ten cuisines

4.4.1 Bar Chart

The bar chart (Fig. 2) is used to show the star rating distribution of neighborhood businesses. The star rating consists of discrete values ranging from 0 to 5 and this chart depicts the number of restaurants in each category of star rating values.

4.4.2 Horizontal Bar Chart

The horizontal bar charts are best suited to display nominal variables like the top ten cuisines. The labels are easier to display for categorical variables on the y-axis. In the horizontal bar chart show in Fig. 3, the count of the restaurants in each of the top ten cuisines is represented by each bar.

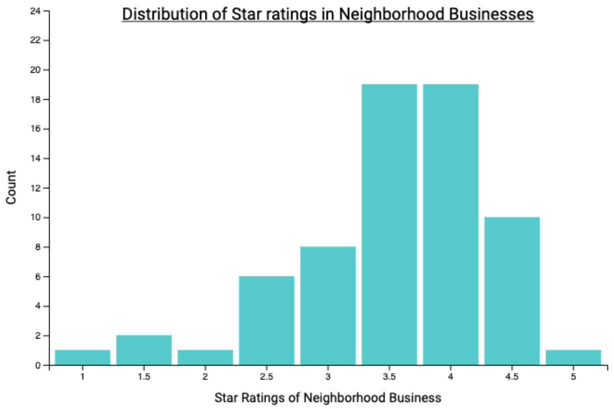

Figure 2. Bar chart showing the distribution of ratings in neighborhood restaurants

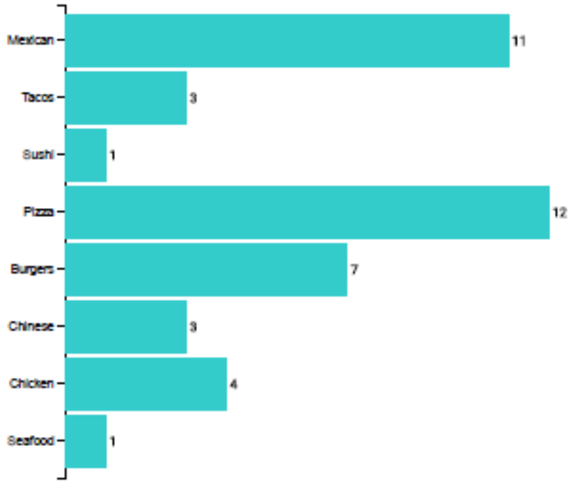

Figure 3. Horizontal Bar chart showing the distribution of ratings in neighborhood restaurants

4.5 Restaurant Performance Analysis

Restaurant Performance Analysis involves the analysis of the growth of a restaurant over the years using key business metrics such as the star rating, number of reviews and the number of check-ins. The following charts display the information based on the selection of a restaurant on the map.

- Annual Number of Reviews.
- Weekly Check-In Distribution of the restaurant.
- Annual Average Rating of the restaurant.

4.5.1 Bar Chart:

The bar chart (Fig. 4) is used to show the distribution of the annual number of reviews for a business. This chart shows the distribution of reviews over the years, helping the restaurant owner understand the popularity of his restaurant while observing the trend.

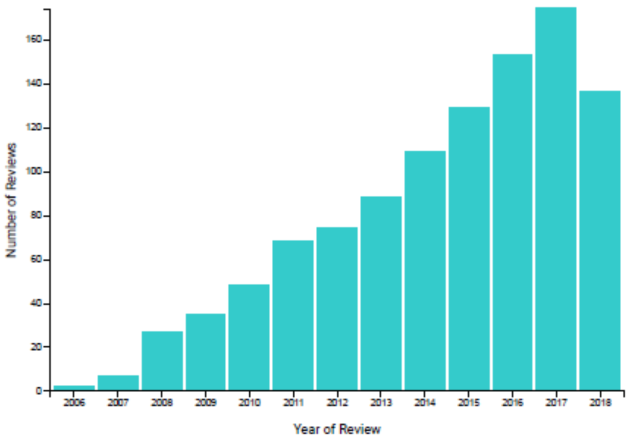

Figure 4. Bar chart showing the number of reviews over the past years

4.5.2 Horizontal Bar Chart:

Horizontal bar charts are best suited to display nominal variables like weekdays. The labels are easier to display for categorical variables on the y-axis. The count of check-ins for each day of the week is displayed using this chart (Fig. 5).

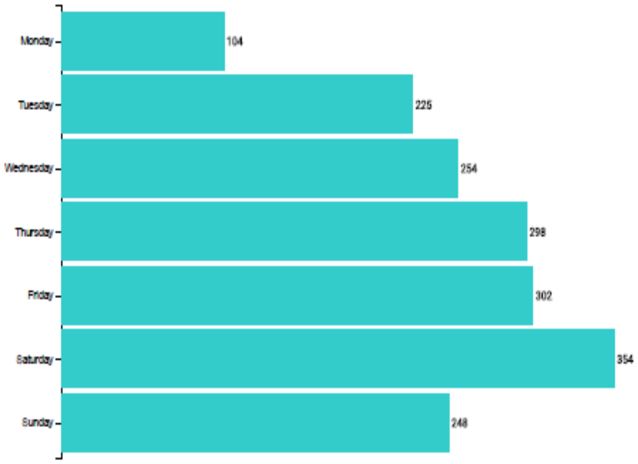

Figure 5. Horizontal bar chart depicting the number of check-ins for each day of the week

4.5.3 Line Chart:

The line chart is primarily used for visualizing the distribution of a continuous variable like average rating over the years. This chart (Fig. 6) can be used to see how the restaurant rating on Yelp has varied over the years, thereby helping the business owner understand trends in the growth of their restaurant.

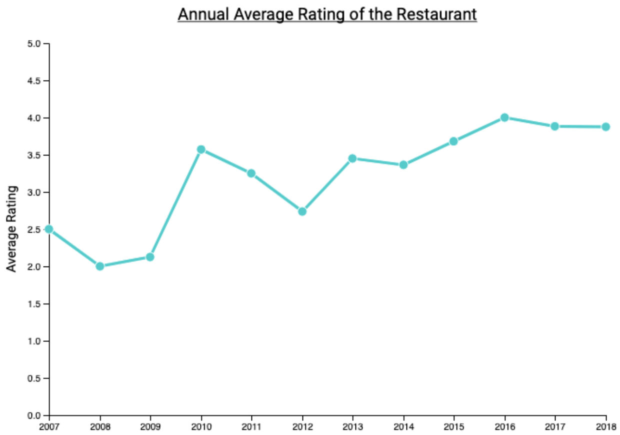

Figure 6. Line chart showing the average rating each year

4.6 Customer Sentiment Analysis

Text Analytics is performed on the reviews to understand the customer sentiment for each restaurant. The following charts display the customer sentiment based on the selection of a restaurant on the map.

- Word Bubble showing the most frequent words in reviews for each star rating.
- Liquid Gauge chart showing the percentage split of sentiment across the reviews.
- Tornado chart to show the top positive and negative words/phrases.
-

4.6.1 Word Bubble

A bubble plot is a scatter plot with a third numeric variable mapped to the circle size. In Fig. 7, the bubbles represent the words with frequency counts greater than 50 while its size corresponds to the frequency of the words. The word bubble is generated based on the value of the star rating input by the user.

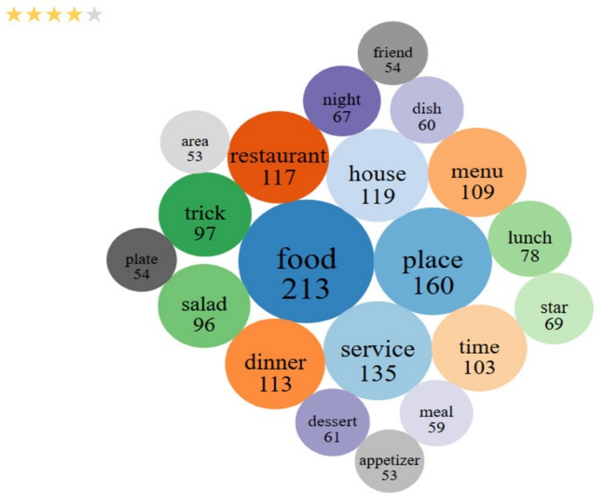

Figure 7. Word bubble chart depicting the counts of the most frequent words for a given star rating

4.6.2 Liquid Gauge Chart

A liquid gauge is a chart which shows the split of a measurable quantity across multiple classes. The split is shown in terms of amount of liquid in the containers. This chart (Fig. 8) is used to show the percentage splits of positive, negative and neutral sentiment of customers in the reviews pertaining to a restaurant. The blue color is used to represent positive sentiment, grey is used to represent neutral sentiment and red is used for negative sentiment.

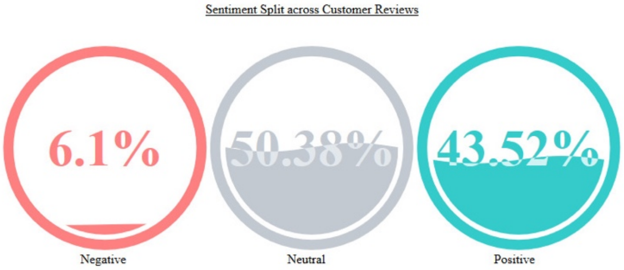

Figure 8. Liquid Gauge chart showing the percentage split of sentiments across customer reviews

4.6.3 Tornado Chart

One of the business insights given by the dashboard is the positive and negative aspects of the restaurant. A tornado chart is best suited for visualizing it. A tornado chart is a vertical bar chart used to visualize the change in a quantity with respect to independent variables. Here, the quantity measured is the sentiment strength and the independent variables are the key words/phrases found in the reviews. The x-axis represents the sentiment score. The keywords are plotted on the y-axis based on the sentiment strength and the polarity. Positive keywords are represented in blue on the positive side of x-axis whereas the negative words are represented with red color and on the negative side of x-axis. The length of the bar represents the absolute sentiment strength. This helps in visualizing restaurant characteristics with opposing polarities in a lucid manner.

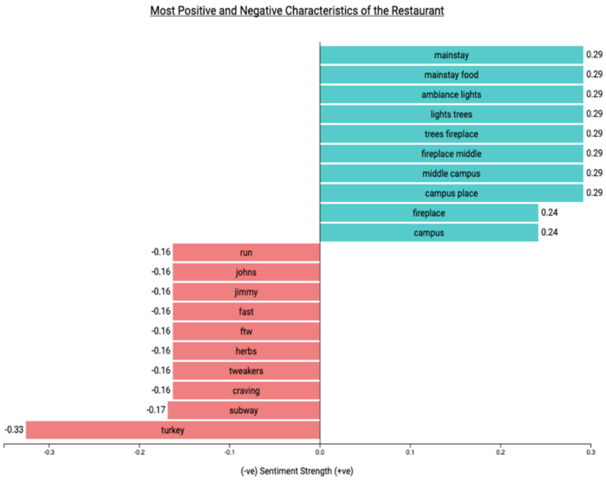

Figure 9. Tornado chart depicting the top positive and negative words/phrases with sentiment strength

5. METHODOLOGY

5.1 Neighborhood Analysis

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

5.1.1 Bar Chart

This chart shows the distribution of the ratings of how the businesses are performing a given neighborhood. The aggregated count of the restaurants based on the star rating in a neighborhood is depicted using each bar in this chart. Bar charts can be used to show the distribution of values in various discrete buckets. In this case, the star rating has discrete values based on the which restaurants are grouped.

5.1.2 Horizontal Bar Chart

This chart shows the top ten cuisines and the count of the restaurant for each cuisine. This gives an insight to the business user of which cuisines people love the most in that given neighborhood and how many of the restaurants are doing well in that category. For each of the restaurant in a neighborhood, the number of restaurants which fall under the given ten cuisines are counted. The chosen cuisines are – Seafood, Chicken, Chinese, Burgers, Pizza, Sushi, Tacos and Mexican.

5.2 Restaurant Performance Analysis

The data found in the Yelp Academic Dataset must be preprocessed for our analysis purpose. Hence, we use Python, mainly Pandas library, to perform data preprocessing and store it in the csv files. Using this preprocessed data, we use the data of the selected restaurant to populate the following charts:

5.2.1 Annual Number of Reviews chart

This chart shows the distribution of the reviews of business over the years. The business owner can understand how popular and trending his restaurant has been over the years. A bar chart is used to visualize a metric over time. This way, the user can spot trends in values over time. Here, the increase in the number of reviews over time indicates that the restaurant is gaining more traction with more customers visiting and rating the restaurant.

5.2.2 Weekly Check-Ins distribution chart

This chart shows the total number of check-ins made on each day of the week. This gives an insight to the business owner that how his restaurant is doing on different days and it can be potentially used to have an efficient allocation and effective utilization of resources. For example, the restaurant owner can come up with deals such as happy hours for the days in which the customer count is low. This way he can drive up sales for the days that are usually less active.

5.2.3 Annual Average Rating chart

This chart shows how the annual rating of the restaurant has varied over the years based on customer feedback. It can be used to understand the highs and lows of a business over the years and urge the business to explore more about such years in particular to unearth more insightful stories. The line chart can be used to study trends over time. In this case, an increase in the average time over time indicates that the business is performing well and that customers are happy with the same.

5.3 Customer Sentiment Analysis

5.3.1 Word Bubble

This chart shows the most prominent words occurring in the customer reviews. We have used NLTK libraries to remove stop words and only display nouns. The frequency has been set to 10 so that the visualization is prominent and understandable. The coloring is standard as we cannot understand whether a word is positive or negative without understanding the context.

5.3.2 Liquid Gauge Chart

The purpose of the chart is to show the split of customer sentiment. For this, each review of the selected restaurant has been classified into positive, negative and neutral sentiment using Stanford CoreNLP library. The sentiment analysis done by the library makes use of a recursive deep neural network model for semantic compositionality, based on the work of Richard Socher et. al. [5]. After the classification, percentage of positive, negative and neutral reviews of the restaurant are calculated and the same are displayed in the chart.

5.3.3 Tornado Chart

The goal of the chart to show the positive and negative aspects of a restaurant. This is a feature extraction problem, where important words or phrases which decide the polarity of a review, have to be extracted. First, the restaurants reviews are classified as positive or negative, using the Stanford CoreNLP library. For each restaurant, the reviews are segregated depending on their polarity resulting in two classes of reviews for each restaurant. Using Python's NLTK toolkit, punctuation and stopwords were removed from the reviews. NLTK parts of speech tagging is used to extract nouns found in each review to form a document. All documents extracted from positive and negative reviews were grouped together to form a corpus. This resulted in two corpora, a positive and negative corpus for each restaurant. Adjectives were not included as they don't provide any additional information. The polarity of the review is already known, so a keyword or phrase extracted from a positive review has positive sentiment and negative sentiment if otherwise. So, adjectives are no longer required to be included in a keyword or phrase. Now, the goal is to find important keywords in the positive and negative corpus of the restaurant. This is done by calculating the tf-idf weight using sklearn's TfidfVectorizer. N-grams of length two are also considered to give a more precise characteristic. This will give phrases like 'chicken parmesan' instead of only the word 'chicken'. Based on the tf-idf weight the words and phrases are sorted and top 10 are displayed in the chart.

5.4 Evaluation Plan

The evaluation of Yelp Business Insights is based on the user query and how they get benefitted from the tool. This system enables the user to explore and analyse competitors, restaurant performance and customer sentiment. Following are the findings:

- We see that Yelp does not offer neighborhood analysis which would be an invaluable insight for anyone who is planning to establish a business in a particular neighborhood or has an established business looking to stay ahead of its competitors.
- Yelp offers no charts for reviews or check-ins. After analyzing the data we decided to add charts for these as they can influence important business decisions.
- Yelp currently lets business owners see the reviews left by the users in the yelp portal. When we look at reviews alone, the emphasis is on good food. But by performing sentiment analysis, we observed that things like ambience (e.g fireplace) actually contribute to the customer sentiment and it's not always about the food. Thereby, helping business owners improve their services in all facets.

For further evaluation, we can ask the user how much they find the system useful.

6. DISCUSSION AND FUTURE WORK

The dashboard can be extended to include all the restaurants available in the yelp dataset across the world to make it usable for any business owner. More interactions can be added to the charts to facilitate the analysis across more dimensions. At present, the sentiment analysis is done on the entire list of reviews for each restaurant. Instead, this can be performed on the basis of star ratings or year for each restaurant as it enables the user to analyze from various perspectives. Metrics such as the star rating and review count can also be modified to be analyzed on a more granular level such as month or day of the week rather than just annually.

7. REFERENCES

[1] **https://www.restaurantbusinessonline.com/operations/heres-how-competitive-restaurant-industry-really**

[2] Yu, Boya, et al. "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews." *arXiv preprint arXiv:1709.08698*(2017).

[3] https://leafletjs.com/

[4] Hegde, Sindhu, Supriya Satyappanavar, and Shankar Setty. "Restaurant setup business analysis using yelp dataset." *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017.

[5] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013.