The Food Explorer

Atit Gaonkar 1217031322 | Haseeb Ur Rahman 1217181953 | Pauras Jadhav 1217154822 | Gayathri Alloju 1218539582 | Gouthami K 1218506822

***Abstract*— Ratings have been always considered by people, a measure to decide if trying out a particular restaurant will be worth the effort, time, and money. Ratings do not reflect the quality of the food served, rather convey an overview of the user's experience. For people interested solely in the quality of the food, the average rating does not serve their purpose. Foodies or people whose sole purpose of travel is to try out different food items, cannot base a decision to try out a restaurant just on average rating. Using techniques like Named Entity Recognition, Fuzzy Logic, and Sentiment Analysis, we aim to solve this concern by grasping the sentiment of the user's review and incorporating it in the context of the mentioned food. Based on the popularity, trend, and overall sentiment for a particular food, we suggest various restaurants for a given state, thereby providing one of the best food experiences.** [1]

I. INTRODUCTION

The user's rating of a restaurant is determined by various peripheral factors ranging from the ambiance, locality, customer service, and user experiences along with the quality of food served.

The following sections revolve around the below-listed research problem.

Research Problem:

(i). State-wise food popularity and food trend.
(ii). Most consumed food vs Most preferred food.
(iii). Recommend State wise top restaurant for its popular food items, based on the quality of the food and user's sentiment about the food.
(iv). Is the average rating the best measure to base restaurant recommendation to foodies?
(v). Rating trend of a state.

Using techniques like Named Entity Recognition, Fuzzy Logic, and Sentiment Analysis, we engineer a platform to provide users with a one-stop solution to satisfy their food craving by understanding user opinion about the food, mentioned in user reviews. Using our intelligent dashboard, users can comprehend the food trend, food popularity, and some of the finest food served by the restaurants of a given state. A list of the restaurant is recommended to the user based on the desired food and user's taste.

[1]**Keywords**: Data Visualization, Yelp, Food detection, Sentiment Analysis, Natural Language Processing

II. MOTIVATION AND RELATED WORK

Foodies are in constant search for their next destination to satisfy their insatiable desire for food. We intend to help these foodies find their next stop. Without a platform that recommends restaurants solely based on food quality and the user's sentiment about the food item, there is a need to develop an intelligent recommendation system that does the same. The decision to work in the food-industry was backed by the fact that one-fourth of the Yelp dataset was related to the food industry.

Average Rating of a restaurant has been a conventional and principal measure based on which a decision is made; Ratings provide an overview picture of the restaurant and hence doesn't serve the purpose of a foodie, as they are interested in the food. A foodie traveling to a different state is unaware of the food trend and famous local food items. We felt it would be convenient to know the most preferred food item in the state and try it from one of the best restaurants.

One of the notable approach taken by Boya Yu et. al [**?**]. where they makes use of sentiment analysis to highlight the key features of a restaurant. In our project, we have worked towards answering the research questions mentioned earlier. We have provided users to request details-on-demand as cited by Shneiderman, B [1].

III. VISUALIZATION

Our platform provides numerous visualizations for the user to interact with. Our Dashboard consists of several different interactive elements, a United States

map, a food bubble cloud, a restaurant recommendation cards, checkin-distribution, and the ability to locate a restaurant on google maps. In order to understand the flow of control, we display only the necessary visualizations at a given time.

A. Map

Users are initially provided with an interactive map of the United States of America where the user can select a map they wish to discover the food trend and food popularity for a given state. Using the principles of Borden D. Dent [3], we intend to keep the map in the focus of the screen.

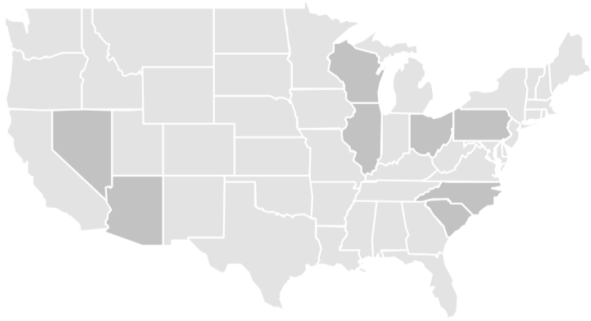

Fig. 1. US Map

Due to a huge discrepancy in the volume of the dataset available for a different state, we categorized the map into two categories, states with considerable data set to perform analysis and those with very low or no data. The ones with very few or no data tend to show unreliable output and as a result of this, we enable the user to select only those maps which contain enough data. States colored with dark gray implies that these states contain enough data and the rest aren't selectable.

B. Bubble Cloud

Bubble Cloud vs. Word Cloud: Word cloud comes with its caveats. The position, orientation, color, and the length of a word needs to be justified to use a word cloud. Whereas the bubble cloud inherently solves most of these issues. Bubble cloud takes in size and color of the bubble as prominent metrics for distinguishing each other.

Once a user clicks on a state, using techniques such as Named Entity Recognition(NER), Fuzzy Logic, and Food Detection techniques (*explained in later section*), we display a food bubble cloud. The bubble cloud contains a variety of information.

- **Food Popularity**: The size of bubble represents the popularity of the bubble and the percentage of people consuming it. For example, in Figure[2], among all the food items, chicken is more popular, followed by cheese, burger, and pizza.

- **Food Trend**: While comparing similar bubble clouds of different states, we can observe a popularity trend of a food item across the country.

- **Most Consumed food vs. Most Preferred Food**: The color of a bubble indicates the average sentiment score of that food in the state. Darker the shade of green, the more it's average sentiment score of the food. Average Sentiment Score conveys the degree to which people like the food. Wine being colored darker than chicken, although chicken is the highly consumed food in the state, conveys the fact that people prefer wine over chicken. Hence people tend to recommend wine over chicken in the state of South Carolina.

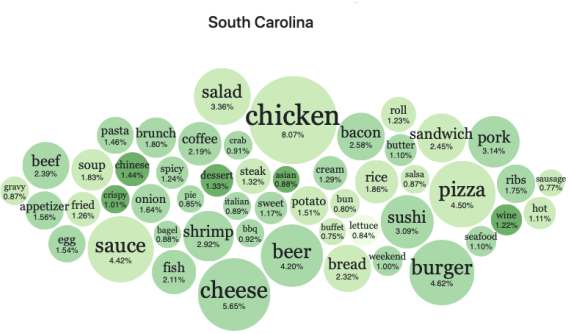

Fig. 2. Recommending Food Bubble Cloud

The bubbles needed to be colored in such a way that they convey the user's compassion, excitement. In order to represent the same, a green sequence color coding was employed, as green generally indicates compassion, excitement. Green is associated with nature and considered to represent freshness. Hence it was imperative to use shades of green to impart the degree of freshness (quality) and compassion.

C. Restaurant Recommendation

Once a user selects a food item to explore the restaurants, we recommend a list of the restaurants based on the restaurant's Hungry Score (*explained in further section*). The user is provided with Restaurant's name, address, it's average rating a

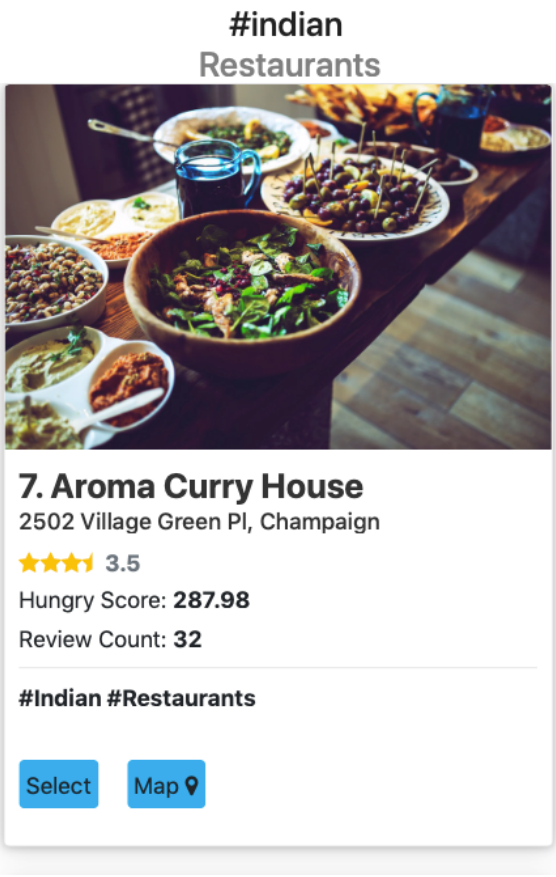

Fig. 3. One of the recommended restaurants

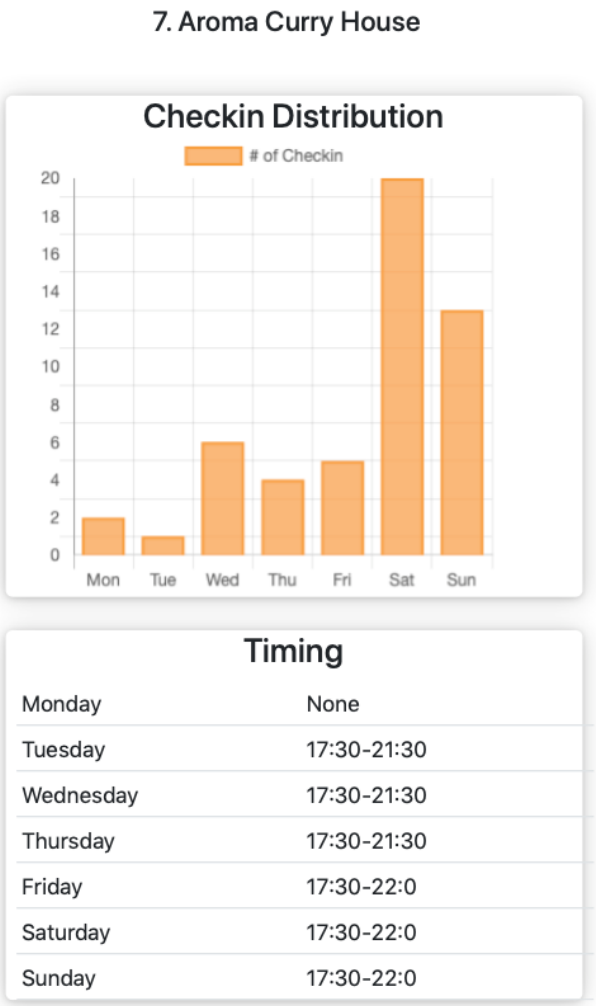

Fig. 4. Restaurant's Timing and Weekly Check-in Distribution

restaurant's hungry score along with their custom tags. To know more about the restaurant, the user can look at the restaurant's timing, weekly check-in distribution to evaluate a convenient time to visit or order from the restaurant. Figure[4] represents weekly check-in distribution of a restaurant.

Furthermore, we have also enabled users to view the location of the selected restaurant on Google Maps. Incorporated Google API calls for the user to interact live with google maps. Figure[5] represents the output of one such API calls.

IV. Challenges

Implementing this project provided us with numerous challenges, among which two challenges namely Food Detection and Food list inculcated us with a wide range of techniques to work-around such an issue.

Unable to find a suitable and existing mechanism to detect food accurately, we had to come up with our Food detection algorithm. This process involved understanding of sentence semantics and the context in which the sentence is being used. In order to proceed further, we decided to come up with a huge list of food items. New York Open Library provided a readily available list of 6000 food items. In need of more food items, we used Named Entity recognition, to detect nouns from the Yelp Image Captions. These nouns were then processed with Fuzzy Logic to provide a list of correlated food list.

For example, the food noun 'Almond' was mapped to several other nouns as mentioned below.

[Almond: (Almond Cake, 92), (Salted Almond Ice Cream, 93), (Salmon,90)....]. Similarly, every other captured noun is matched to find a list of corresponding nouns. As the yelp captions were primarily based on food items, we observed a dense list of items corresponding to a food item, whereas non-food nouns,

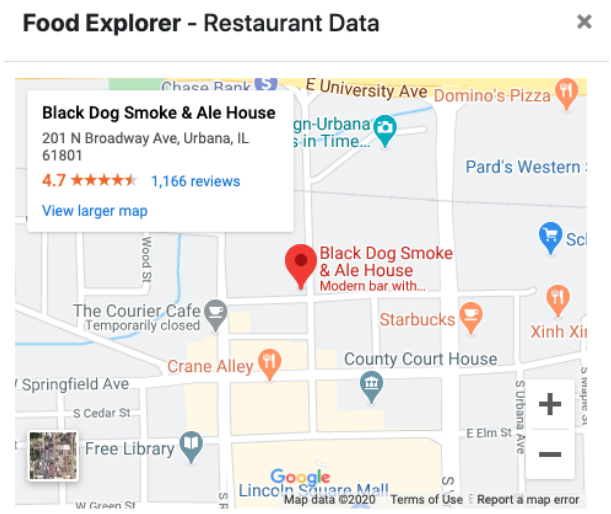

Fig. 5. Locate restaurants location on google map.

turned out to have a spare list. Hence we were able to capture a total of 16000 food items. As the list consisted of complex and high redundant usage of food items, along with the limitation of computing power we then consolidated this food list. For example 'Lemon Cake' and 'Almond Cake' in simpler terms represent 'cake'. As a result, the original food list was consolidated to 1000 food items. This list was further used as a comparative match used with Fuzzy Logic to detect food items in user reviews. Hence we were able to develop a food detection algorithm that could detect food items with an accuracy of 82%.

The second most challenging task was the computational power needed to detect these food items from 2 million reviews. As each review took on an average of 1 sec to process, a single standalone machine would need 28 days of continuous computing power. Using Google's Colab, we were able to split the computational work into 65 sessions among ourselves to produce the result within 24 hours.

V. Methodology

In order to obtain key metrics to recommend a restaurant for a given food, we devised an algorithm to detect food items from reviews while keeping a tab of their count. Furthermore, we have also kept track of their average ratings using the sentiment score of reviews.

A. Dataset Description/ Cleaning

The dataset was obtained from Yelp, consisting of about 6.68 million reviews for 192,610 businesses. For data cleaning, we first converted the data from JSON format to CSV. Thereafter, the data was filtered to only include restaurants from all over the United States. The new filtered dataset contained about 2.56 million reviews from 174,568 restaurants. This dataset was further split based on the state in which the restaurant is located.

A large portion of the reviews was from Arizona and Nevada with nearly one Million reviews each. Ohio, North Carolina, Philadelphia states had more than 100,000 reviews. Other states had few to zero reviews and hence were considered as unreliable data to recommend food and restaurants.

We have also used the Yelp photo dataset to extract the captions and labels from photo descriptions to later use this in the food detection task.

B. Sentiment Analysis

User reviews were analyzed to obtain a polarity score.

We have used two different algorithms namely, TextBlob and VADER (Valence Aware Dictionary and Sentiment Reasoner) in combination to offset their individual drawbacks.

A polarity score in the range of (-1,1) was calculated where -1 indicates negative, 0 indicates neutral and +1 indicates positive.

This range was further scaled to a range of 1 to 5 to maintain consistency with the original Yelp rating scale. Hence the scores were modified as follows: ('1': Highly Negative, '2': Moderately Negative, '3': Neutral, '4': Moderately Positive, '5': Highly Positive).

We have observed many cases where the ratings are low due to non-food related factors such as ambiance, locality, customer service, and user experiences, but the reviews for the food are good. The motivation behind performing sentiment analysis was to minimize the discrepancies between the review for the restaurant and the rating given.

C. Food Detection

In-order to perform food detection, we compiled a list of dishes from a variety of sources, including

the New York public library and yelp photo captions. Using the label attribute, a total of 44,000 relevant captions were extracted. Using these captions and the list of dishes compiled earlier, we calculated the top thousand food dishes based on their frequency. The food detection task involved performing NLP (Natural Language Processing) tasks such as sentence cleaning, sentence parsing, semantic analysis, and named entity recognition (NER) on user reviews. Using pre-trained NLP libraries such as SpaCy, NLTK, and fuzzy-wuzzy, along with the compiled food list with a precision of 70% for similarity matching, we were able to detect food items mentioned in a review with 82% accuracy.

D. Recommendations

We perform recommendations for two different instances, firstly for recommending top food items for a given state and secondly to provide a list of a top restaurant for a given food. To obtain the recommended food bubble cloud for a given state, we employ NLTK libraries such as FuzzyWuzzy and Spacy for food detection techniques. We keep a tab of the occurrences of food items (Count) in reviews and the sentiment of those reviews in context with the food (Average Sentiment Score). To analyze the sentiment of user's reviews, we take advantage of a hybrid combination of TextBlob and VADER algorithm.

Once we obtain a food bubble cloud, we recommend restaurants based on their Hungry Score (H-Score). Hungry Score is a product of a restaurant's 'Average Sentiment Score' and 'Count' in the context of a given food item. Larger the Hungry score, it is highly likely to be recommended. Top recommended restaurants for food have the highest Hungry Score.

E. Technologies Used

The following technologies were used to implement this project:

- NLTK, FuzzyWuzzy, and Spacy: In order to extract nouns, further is refined to contain food items, we make use of various Natural Language Processing libraries, such as NLTK, FuzzyWuzzy, and Spacy. We also get rid of unnecessary details, such as stopwords, non-noun words to limit the volume of the dataset. These libraries, make use of Regular Expression, tokenization, semantic analysis to extract food items from user reviews. Using fuzzy logic we try to consolidate a huge food list into simpler food items. Hence giving us an advantage by reducing unnecessary computations

- D3 and Leaflet: D3 and extension provided by Leaflet, we use it to our advantage to render the United States map along with state borders.

- TextBlob and VADER: Sentiment analysis has played a huge role in our recommendation system. Sentiment analysis algorithms, in general, have performance bias. Some perform well on negative reviews and some on positive ones. In order to counter their individual drawbacks, we make use of a combination of TextBlob and VADER (Valence Aware Dictionary and Sentiment Reasoner)

F. Food Popularity Analysis

Highly rated food items are analyzed from the customer reviews from the Yelp Dataset and categorized concerning states. Based on the chosen state the popular food items are presented to the customer in a Bubble Chart. A list of popular restaurants is displayed when the user selects their desired food item from the multiple food items in the Bubble chart.

Figure:2 Bubble chart with Food items.

VI. Evaluation Plan

- Food Detection: Bubble cloud is a validating measure that our novel approach to detect food from reviews and score the review based on the sentiment of the review in the context of given food work. In-fact we are able to detect the food with an accuracy of 82%.

- Validating restaurant recommendation: In order to validate the recommended restaurants for a given food item, we can cross-verify it using the restaurant categories provided by the Yelp dataset. For example, selecting pasta as a food item may result in a restaurant with tags, italian to be recommended. This correlation between the restaurant tags and selected food item implies some degree of validation.

- Restaurant Recommendation Trend: Not all top recommended restaurants outperform other restaurants in terms of an average rating, but they certainly do outperform in terms of their Hungry Score. As the hungry score is proportional to the average sentiment rating of the restaurant, this

indicates that the restaurants being recommended corroborate the validity of our program.

VII. Results and Findings

After successfully integrating the individual module to provide one of a kind recommendation system, we discovered the following trend:

- Understanding food popularity and trend: Based on the demographics of the bubble cloud, we observe the following information regarding the popularity and trend of various food items:
  1) The size of the individual bubbles in the bubble cloud conveys how popular a food item is for a particular selected state.
  2) Realizing the popularity of food items in various states indicates the trend of a food item across borders. For instance, customers in coastal states prefer seafood more than those in inland states. Likewise, southern states show a higher preference for Mexican food as those compared to northern states.
- Comparing the most consumed food vs. most preferred food: Although the food may be heavily consumed, it is not necessarily the most preferred in the state. People who consume wine in South Carolina are more likely to rate its positively, but this cannot be said about chicken, even though chicken is highly consumed in South Carolina.
- Average Rating alone doesn't justify the quality of food: Restaurants are recommended by calculating a restaurant's Hungry Score which reflects the user's sentiment of the restaurant for a given food item. We observe that not all the top-recommended restaurants have a higher average rating than others.

Also, the current generation tends to order food online or take out rather than dine-in. With the introduction of UberEats, GrubHub, Postmates as a platform-based delivery system, there has been an exponential growth in people using the platform to order food online. Hence, their decision to order food from a restaurant will highly depend upon the quality of the food, and hence our platform provides recommendations with the importance of the quality of the food.

VIII. Discussion and Future Work

- **Travel guide**: Users aiming to visit multiple destinations to try out different restaurants would need an efficient navigation system to travel and enjoy the journey. Hence we could suggest some of the worthy restaurants and food items try out which fall in the user's route. For example, a user traveling from Arizona to Nevada to try some of the best sushi restaurants, we could navigate as well as recommend restaurants to the user from Arizona to Nevada in such a way that if the user chooses to explore various other restaurants along the way would be able to do so.

- **Usage of Back-end**: To enable further dynamic analysis of reviews, recommend food items and restaurants, the introduction of back-end would be highly beneficial. Ability to apply multiple custom filters, such as locality, hungry-score requires, the support of back-end.

- **Customised User Recommendation System**: Introduction of back-end would then enable the ability to perform user-specific and personalized recommendation system based on user's food taste and previous selection. Users can set their preferences, based on which all further recommendation will be made to a particular user, providing a customized tailored experience that fits the user's needs.

- Dynamic Photo: While recommending a list of restaurants, we could dynamically match the restaurant with its corresponding pictures so that the user's if needed to base a decision to visit a restaurant on the look and feel of it, can do so.

- **Improving Named Entity Recognition**: Our current customized Named Entity Recognition algorithm detects food with an accuracy of 82% (testes on a sample data-set of 100 food items), which can be further made robust by integrating with prominent algorithms in the field of Named Entity Recognition and Food detection.

- **Detection of complex food dishes**: Our novel approach to detect food items currently detects simple food items. For example, Peanut Butter Sandwich would be further decomposed into three different food, Peanut, butter, and sandwich. But further improvisation on context analysis and food detection could result in the extraction of a complex food item.

As of now, we could only plot data available from

certain states, but we would like to include all the states upon the availability of the dataset.

REFERENCES

[1] Shneiderman, B, The eyes have it: a task by data type taxonomy for information visualizations, 1996.

[2] Andrew Gelman, Jennifer Hill, and Aki Vehtari, Regression and Other Stories, Chapter25 Missing data Imputation.

[3] Borden D. Dent: Cartography – Thematic Map Design, 5^{th} ed. Chapter 13

[4] Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao, Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews, 2017.