|  |
| --- |
| **Restaurant Menu Recommender: Yelp Dataset Review Sentiment Analysis** |

**Aditya Bhardwaj Karan Chaudhri Shubham Goyal**

[*abhardw2@ncsu.edu*](mailto:abhardw2@ncsu.edu) *krchaudh@ncsu.edu gsshubha@ncsu.edu*

**Abstract**

Thinking about which food item to order from the menu is a cumbersome task for any customer at a restaurant. In this project, we are performing sentiment analysis over the restaurant reviews given by users in the Yelp Dataset scoring every food item in the review with a value and creating list of food items in ascending order of score, top being the first food item. On basis of reviews and score given to that food item, we suggest 5 food items to try at restaurant.

**1 Introduction**

Restaurant reviews plays a vital role in selection of restaurant for any customer, the more important aspect comes when customer want to try some food at the restaurant. Such a task can be cumbersome and usually put a customer in frenzy on what to choose and what not to. To simplify such a task, we address the business problem where customer will be given food recommendation based on the restaurant reviews by other customers.

At current stage, recommendations are based on various factors and we leveraging the restaurant reviews which are given by users based on their experience and likings. We intend to use Sentiment Analysis over the data to produce a recommendation JSON file which can be utilized by any tool including Mobile Application, Web Application and APIs. We have created one such tool in Python CLI to show how the tool recommends food items after score value evaluation of food items in a review.

**1.1 Data set**

1. Business Data

Rows = 144K, Columns = 8

2. Review Data:

Rows = 4.1M, Columns = 9

3. Tips Data:

Rows = 947K, Columns = 6

4. Food Items List:

Food Items count = 4000+

**1.2 Tools used**

1.2.1 SQLite: For Cleaning and preprocessing of data.

1.2.2 Python: For running the algorithm and creating a wrapper app for Food recommendation.

**2 Business Questions**

2.1 What is the probability of a friend to be there at same restaurant as the user at same time?

Based on user reviews and check-in information for a particular restaurant, one can combine it with the social check-in of a user. Finding friends of users via social networks and compiling such information with the yelp profile of each user, we can find out the next time where both user and his friend could be together at same place.

2.2 At what time, Restaurants can be busy next time?

This is rather a simple problem and can be solved by using check-in information from tips and business dataset provided by Yelp. To add more specificity in context, we can use social network check-in along with yelp check-in data to analyze at what time restaurants can be busy. Checking this information over a period of time will result in forecasting of date/day/time when restaurant will be busy next time.

2.3 What is the probability of the next review to be useful, cool etc.?

In Yelp dataset, we have information about review rating by other users which gives information whether the review is useful, cool, hot etc. We intend to predict whether the review given by a user would be rated useful, cool etc. by other users or not. Such a model can be helpful for business owners to find their liabilities and the future impact of such a review on their business.

2.4 Which kind of food to suggest to the user on basis of restaurant reviews he is visiting?

The most viable business question in today’s date is to please a customer and decrease the bad publicity for a business. We intend to solve both the problems in a single problem statement intended to help customers decide what to order while visiting a restaurant. Such a model would result in food recommendations from menu based on user reviews for a particular restaurant.

**3 Targeted Business Question**

Which kind of food to suggest to the user on basis of restaurant reviews he is visiting?

Restaurant reviewing system can be a simple yet effective business model. In such a model, one of the problem which we are focusing on is recommending users with a menu which they would like to try while visiting restaurants to avoid any dissatisfaction at the user end. This model is a win-win situation for both user and restaurant owner to focus on things which needs to be improvised and which are praised by users.

The method which can be employed to solve such a question which is heavily based on the user review is Sentiment Analysis using NLP.

**4 Research Papers and Survey**

4.1 Ruchi: Rating Individual Food Items in Restaurant Reviews

Author has tried and rated each individual food item in a restaurant based on the reviews of the food item in restaurant. One of the biggest challenge in doing so was extracting the food items served in a restaurant from the reviews as this data was not made directly available. To extract the food names from reviews author has used trained NER model, using which corpus of online customer reviews was annotated with food items mentioned in the review. After the food item has been identified next step was sentiment analysis of the review to rate the food items. A simple sentiment analysis of sentences would not have worked here because a single review could contain reviews about multiple food items. The author performed entity level sentiment analysis to find the polarity of opinions. Finally, individual food items are rated based on the polarities of all the opinions received for each of the food items. We were faced with the same challenge of tagging the reviews with food items mentioned in the review and then performing sentiment analysis on the reviews. We chose a more simplistic approach for tagging the reviews. We were able to find an exhaustive list of food items online, which was not available to the authors of Ruchi. We matched each review with the list of food item to check which food items were available in that particular restaurant.

4.2 Buon Appetito - Recommending Personalized Menus

In this paper author has built a menu recommender system and is very similar what we are trying to do in our project. Particularly, our approach for performing sentiment analysis on the reviews is inspired from the approach used by author in this paper. The results from this paper indicate that sentiment analysis at the sentence level helps to improve the quality of recommendations. In order to reduce noise and sparsity, author split the reviews in sentences and remove stop words from each sentence. To understand the polarity of text i.e. the amount in which it is positive or negative author has used the LIWC 2007 dictionary of sentimentally-annotated words. We have also used similar list of sentimentally-annotated words made available to us during the coursework. Given a text author has defined the sentiment score S = (p - n) / (p + n). We have used similar formulae where each sentence is tagged positive if the score is >= 0.5 and negative if score is <= -0.5.

4.3 Generating Recommendation Dialogs by Extracting Information from User Reviews

In this paper authors tries to generate more relevant and specific questions to be asked from user to streamline the search of restaurants. For e.g. Yelp provides each business with category labels for top level cuisine types like Japanese, Vegetarian, Coffee and Tea. Many of these top-level categories have natural subcategories. By identifying these subcategories, author enabled questions which probe one step deeper than the top-level category label. The sentiment propagation technique used by author is of our particular interest. The author begins with seed set from Opinion Finder Lexicon. We begin with the seed set provided to us during the course work. Author has assigned negative seeds value of 0, positive seeds value of 1 and all other adjective with value of 0.5. For the sake of simplicity, we assigned the negative seeds value of -1 and positive seeds value of 1. This allowed us to eliminate reviews that are neutral in sentiment which would have been difficult with value of 0 and 1. Author has eliminated candidates which do not link at least two positive or negative seeds. This was to avoid spurious propagation caused by one-off parsing errors. We avoided this approach because we were working on a smaller dataset and doing that would have a lead to not taking into account lot of one sentence reviews like “Sushi here was good”. The author then goes on to discuss techniques used for detecting question selection from dialog which are not relevant to our project work.

**5 Methods and Implementation**

**5.1 Data Cleaning**

The original data set contained many business, their information and review from various customers which was an overhead in our project which is focusing on the Restaurants. During data cleaning, we needed to remove the business overhead by removing all the business which does not fall under the category of restaurants.

As the data was preprocessed, so we didn’t have to worry about null values in the data but we ran the query to eliminate any null values from the data. After removing all other business categories, the data size was still significant. Extra fields in the tables were removed from each dataset in order to perform accurate analysis over the data.

After removal of unnecessary data from the tables, the tables contain following information in them:

Business Dataset: Business ID, Stars, City, State

Review Dataset: Business ID, Stars, Text

Tips Dataset: Business ID, Tips

As we are focusing only on Business, the dataset only contains information pertaining to evaluate reviews and tips for a particular business. The next steps before analysis includes sampling of data to make it manageable.

**5.2 Sampling**

As the data provided by the Yelp dataset was huge even after cleaning. Therefore, there was need to sample the data to make it more manageable for sentiment analysis. The reviews and business tables were sampled on state values and we extracted various tables for states including Illinois, South Carolina, North Carolina etc. to perform Sentiment Analysis more robustly over a smaller dataset.

Following queries were implemented in SQLite Browser to clean and sample the data:

|  |
| --- |
| CREATE TABLE businessClean AS  SELECT b.\*, yelp.city, yelp.state, yelp.latitude, yelp.longitude  FROM business AS b INNER JOIN yelp\_academic\_dataset\_business AS yelp ON b.business\_id = yelp.business\_id; |
| CREATE TABLE tipsCleaned AS  SELECT tip.business\_id, tip.user\_id, tip.text, tip.date  FROM tip INNER JOIN business  ON tip.business\_id = business.business\_id; |
| CREATE TABLE business\_SC AS  SELECT \* FROM  businessClean WHERE businessClean.state = 'SC'; |
| CREATE TABLE review\_SC AS  SELECT \* FROM  reviewCleaned INNER JOIN business\_SC ON reviewCleaned.business\_id = business\_SC.business\_id; |
| CREATE TABLE tip\_SC AS  SELECT \* FROM  tipsCleaned INNER JOIN business\_SC ON tipsCleaned.business\_id = business\_SC.business\_id |
| CREATE TABLE cleaned\_SC  SELECT f2.field1, r.text FROM foodlist AS f2, review\_SC AS r  WHERE r.text LIKE '%'||f2.field1||'%'; |

**5.3 Sentiment Analysis**

The sentiment analysis discussed under the Ruchi paper was based on the Stanford NLP toolkit which we are not using in the project. Our approach is based on the Sentiment Analysis as taught in the class using the NLP approach over Doc2Vec. We don’t have the set of positive and negative review for Doc2Vec to be implemented efficiently, so we adapted NLP approach based on the list of positive and negative words as provided in the Sentiment Analysis project.

After Sentiment Analysis, we are organizing positive and negative reviews for a particular food item in the list in dictionary based format, more the positive review for that particular food item, more would be its positive count. The organized data is saved in a JSON format for the convenience to be implemented in the wrapper application for faster recommendation to a user.

**5.4 Data Obtained after Sentiment Analysis**

Below is the structure of the data which is generated after performing Sentiment Analysis on the data. This structure would be utilized in a wrapper application to support our deductions on suggesting food items to the user.

|  |
| --- |
| {‘business\_id1’: {  “business\_id”: ‘business\_id1 value’,  “food\_items”: {  “food1”: {  pos\_review: [review text],  pos\_count: positive count value,  neg\_review: [review text],  name: ‘food name’,  neg\_count: negative count value  },  “food2”: …  }  },  ‘business\_id2’: …  } |

Figure 1 Structure of the data created after sentiment analysis

**6 Team Roles**

6.1 Aditya Bhardwaj

Data cleaning and sampling of Data into a manageable data.

Writing wrapper application on Python for a User perspective.

Writing and compiling project report

6.2 Karan Chaudhri

Running Sentiment analysis on the review csv file for the state: South Carolina.

Running Sentiment analysis on the top csv file for the state: South Carolina

6.3 Shubham Goyal

Understanding the dataset and come up viable problem set.

Reading and critique research papers for the targeted business questions.

Cleaning review data file out of extra attributes for the project.

**7 Conclusion**

Users are generally confused on which food item to order from the menu in a restaurant. The normal tendency for a user is to check the reviews for a particular restaurant and find what food item in the menu is good or not. Going through reviews can be a tiring task, and users might not be able to find relevant information about menu in them. We aim to eliminate the need for user to read lot of reviews before deciding which food item is good to try at a particular restaurant by automating the whole task and giving a list of food items which are reviewed good. This task has been performed by using Sentiment Analysis NLP as a technique and giving a positive and negative score to the food items. With an output in JSON format, this analysis can be used a tool for any application to leverage its information for suggesting user food items. This have been implemented using wrapper.py in the code.

The future scope of the idea would be creating a graph of user and restaurant reviews to connect user nodes with restaurant nodes based on their reviews. This way one will be able to suggest type of restaurant user should visit next, type of food user should try next and predict the pattern of user ordering food items from restaurants which would prove to be a nice business model and a tool for restaurants to improve and analyze their services.

Our code can be found here:

<https://github.com/intellectape/Yelp-Dataset-Challenge>

**References**

[1] Buon Appetito - Recommending Personalized Menus: Michele Trevisiol, Luca Chiarandini, Ricardo Baeza-Yates, Yahoo Labs, Universitat Pompeu Fabra

[2] Ruchi: Rating Individual Food Items in Restaurant Reviews: Burusothman Ahiladas, Paraneetharan Saravanaperumal, Sanjith Balachandran, Thamayanthy Sripalan and Surangika Ranathunga

[3] Dataset: <https://www.yelp.com/dataset_challenge>

[4] Converting Yelp Data from JSON to CSV: <https://github.com/Yelp/dataset-examples>

[5] Food list: <http://www.foodtimeline.org/>

[6] Food list: <http://www.oxfordreference.com/view/10.1093/acref/9780192803511.001.0001/acref-9780192803511>