Opinion Phrase Mining of Restaurant Reviews

Gazi Mahir Ahmed Naven, Azmayeen Fayeque Rhythm
University of Rochester

I. Background and Motivation

Customer reviews and ratings are a useful source of information to identify why some restaurants are more popular than others. Yelp has been one of the most popular sites for users to rate and review local businesses. Businesses organize their own listings while users rate the business from 1 − 5 stars and write text reviews. Users can also vote on other helpful or funny reviews written by other users. Using this enormous amount of data that Yelp has collected over the years, we have applied discriminating text mining to analyze the customer reviews and find text or phrases that appear frequently in customer reviews. Thereby, we identify what are the unique attributes that differentiates between similar restaurants that are doing good vs bad in their business. Our objective is to understand why similar restaurants in a high end or low end area receives various customer reviews. We expect that this project will help restaurants have a comprehensive feedback of customer reviews with insights that will help restaurants to work for improvement.

II. Related Work

With the advent of social media, blogs and peer to peer networks, opinion mining and sentiment analysis became very popular in the field of data driven research. A detailed overview of similar existing work was presented in (Pang and Lee, 2008). In their survey, the authors describe existing techniques and approaches for an opinion-oriented information retrieval.

Many predictive tasks previously performed on the Yelp Dataset worked in category prediction based on business attribute [1]. The research intended to classify the reviews as funny, useful or cool, the metric used by Yelp to evaluate user reviews. In addition to the general area of text mining and sentiment analysis, there are research on text mining for predictive tasks in review and rating systems. The impact of text derived information has been previously studied by using topic modelling on datasets[1].

Furthermore there has been research to utilize existing supervised learning algorithms to predict a review‘s rating on a given numerical scale based on text alone. Yelp dataset has been used to experiment different machine learning algorithms such as Naive Bayes, Perceptron, and Multiclass SVM and compare the predictions with the actual ratings. [2] In addition, microblogging websites have been used as source of data for opinion mining and sentiment analysis. This research [3] focuses on performing linguistic analysis of collected corpus and using the corpus builds a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document.

Other methods including regression [4] and bag of opinions method [5] have also been explored in the past. This paper [9] concludes that bag-of-opinion phrases outperform the bag-of-words topic model and using the grammar relationships outperforms the existing preprocessing techniques in extracting aspects from users. Hence, our approach has been to work with opinions to understand overall pattern in user reviews for restaurants.

III. Data

The data was downloaded from the Yelp Dataset Challenge website **https://www.yelp.com/dataset_challenge/dataset.** This database contains the Yelp business dataset and the review datasets, which has information on the businesses, reviews, users, address, reviews. We only focused on businesses which were Restaurants, Cafes, and Food places. The total number of such businesses were above 65,000. Hence we were only focusing on the user reviews on these businesses, which was over 2,569,000 reviews.

We also used the Data.gov Zipcode dataset from https://catalog.data.gov/dataset/zip-code-data. We were able to extract the Average Adjusted Gross Income for 27,791 Zip Codes. We only used the Zip Codes for the remaining businesses. Income was used a metric to define similarity of restaurants. The main idea was to distinguish the standard of living of the respective restaurant locations. We expect that this distinction and accumulation of restaurants, in close income ranges would provide us reviews for more accurate text mining for the respective types of restaurants. This relationship between level of income and types of restaurant people usually visit has also been explored in other researches. For instance, the papers in [6] and [7] found a positive correlation between Income and Eating behavior which further strengthens our idea of segmenting the

restaurants based on the income range of an area. Thus, for instance, in an area where the income level is high, the restaurants are likely to be more high end since they would target relatively more rich customers.

IV. Methodology

A. Pre-processing

We wrote a Python parser to read in the business.json and review.json, from Yelp Dataset, and Income.csv, from Data.Gov, data files. It reads each of the file in a pandas dataframe.

Then we created two new data frames called bus_df and review_df. We merged [Yelp dataset: bus_df, Income.csv *on* postal_code = ZIPCODE] and called it BI. Then we cleaned BI to keep only the relevant columns such as: name, business_id, review_count, starts, state, ZIPCODE, AGI and Avg AGI. Then we merged BI with review_df [Yelp dataset: BI, Yelp dataset: review_df *on* business_id]. This is Megaset dataset in the code.

Finally, according our restaurants similarity approach, we divided the dataset into four distinct datasets. We took Average Aggregate Gross Income (Avg. AGI) by binning on the following ranges [0-90,000], (90,000-200,000], (200,000- 300,000], [above 300,000). Each of the new datasets was then divided into three more datasets, according to the ratings of the restaurants. Hence, we defined a relatively "poor rated restaurant" for businesses rated between 0 - 2.5 stars, and then relatively "mid rated restaurants" for businesses rated between 2.6 - 3.9 stars and finally, "high rated restaurants" for businesses rated above 4 stars. This was the final dataframe division and, so we dropped all the other attributes, and kept the 'text' attribute, which contains the reviews from the users. Finally, we had a total of 12 datasets. These datasets were outputted into individual .txt files using the pandas library.

To prepare the data for review processing, the algorithm was given only was given only one user review at every iteration. We also processed each .txt file one at a time to keep the processed pattern from each of the 12 review datasets separate at the same time. The algorithm for processing each of the review sentences has been outlined below.

Figure 1, below, outlines the whole database pre-processing.

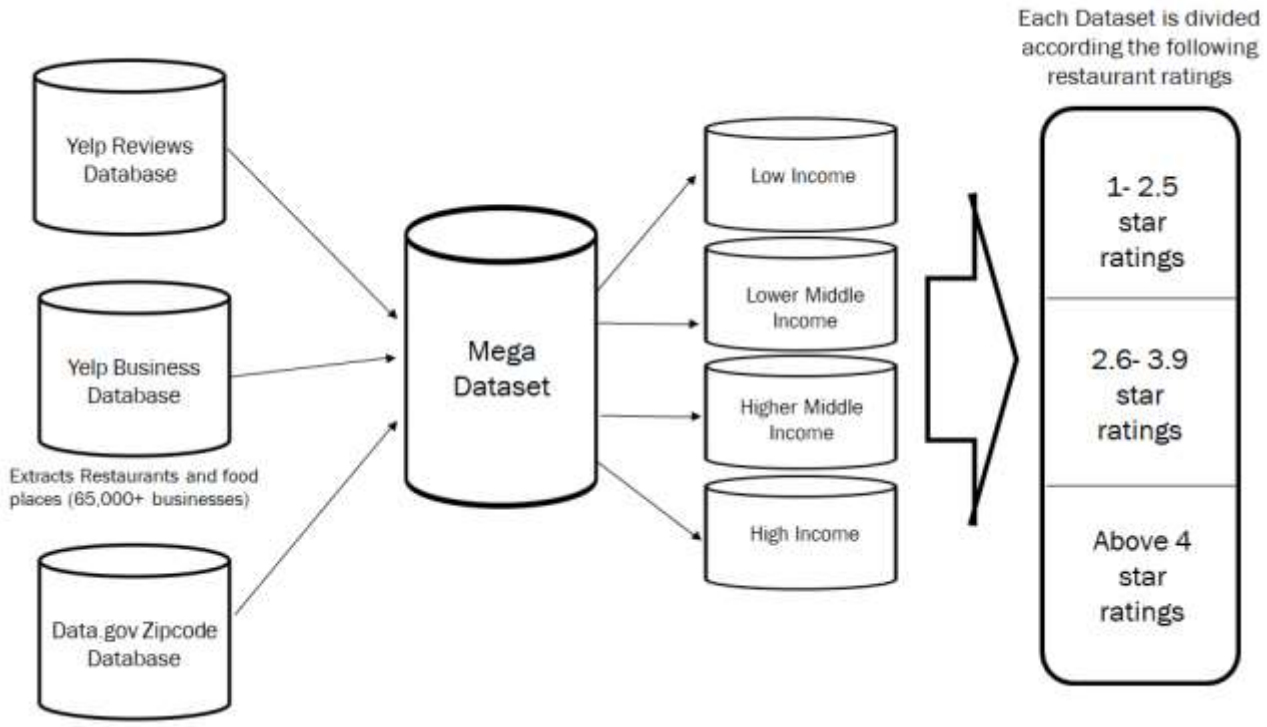


Fig :1

B. Review Processing

We implemented and developed our review processing algorithm from vlad sandulescu 's algorithm on opinion phrase mining. The idea is that when reading a review a simple way to extract people's opinion is to look for nouns and pick the nearest adjective around it. We were also able extract the compounded nouns. Thereby, it can show the syntactic dependencies between words and even predict the overall sentiment of a text.

The algorithm takes text input to begin the review processing. By tokenizing the sentences it takes each word apart and applies parts of speech tagging for that word. From these tags a basic pattern is formed. First, in the ReviewPhrases class it creates objects for phrases and reviews. In the Pattern class is creates an object for pattern and sets a relation for patterns as key, value pairs. When extracting the basic pattern, we use stopwords to clean the patterns so that we are left with only significant patterns. This is the pruning step.

[Used Stanford.nlp to allocate the parts of speech tag for each word]

The parts of speech tags used here are:

NN : Noun Singular

JJ : Adjective

VB : Verb

In order to end-up with opinion phrases, first, basic patterns are extracted [8]:

Here are examples on how the modifiers work and how they create the pattern.

1. Adjectival complement (acomp): *The camera looks nice*, parsed to *acomp (nice, looks).*

2. Adjectival modifier (amod): *This camera has great zoom* parsed to amod(zoom, great)*;

3. "And" conjunct (conj and): *This camera has great zoom and resolution* parsed to conj(zoom, resolution).

4. Copula (cop): *The screen is wide* parsed to cop(wide, is).

5. Direct object (dobj): *I love the quality* parsed to dobj(love, quality).

6. Negation modifier (neg): *The battery life is not long* parsed to neg(long, not).

7. Noun compound modifier (nn): *The battery life is not long* parsed to nn(life, battery).

8. Nominal subject (nsubj): *The screen is wide* parsed to nsubj (wide, screen).

The simple patterns are then combined in a tree-like manner to obtain more valuable opinion phrases. (N indicates a noun, A an adjective, V a verb, h a head term, m a modifier, and < h, m > an opinion phrase)[9]

1.amod(N, A) →< N, A > *This camera has great zoom and resolution → (zoom, great)*

2.acomp(V, A) + nsubj(V, N) →< N, A > *The camera case looks nice --> (case, nice)*

3. cop(A, V) + nsubj(A, N) →< N, A > *The screen is wide and clear --> (screen, wide)*

4. dobj(V, N) + nsubj(V, N0) →< N, V > *I love the picture quality --> (picture, love)*

5. < h1, m > +conj and(h1, h2) →< h2, m > *This camera has great zoom and resolution --> (zoom, great), (resolution, great)*

6. < h, m1 > +conj and(m1, m2) →< h, m2 > *The screen is wide and clear --> (screen, wide), (screen, clear)*

7. < h, m > +neg(m, not) →< h, not + m > *The battery life is not long --> (battery life, not long)*

8. < h, m > +nn(h, N) →< N + h, m > *The camera case looks nice --> (camera case, nice)*

9. < h, m > +nn(N, h) →< h + N, m > *I love the picture quality --> (picture quality, love)*

For example, one of the reviews in our dataset is as follows:

“Excellent food, large portions. As you drive into the lot your mouth begins to water because you have just taken in the heavenly smell coming from the smokers. You will find that the food is just as heavenly. "Ahhhh... BBQ. such a satisfying treat”

After tagging each word, it creates a basic pattern of <N,A> form. Some example patterns:

[Used Stanford.nlp to recognize the modifiers]

*Noun Compound Modifier (nn): Excellent food, large portions * parsed to nn * (food, portions)*

*Adjective complement (acomp): The food is just as heavenly * parsed to acomp * (just, heavenly)*

*Adjectival modifier(amod): You will find food is just as heavenly*parsed to amod*(food, heavenly)*

The pruning step gets rid of insignificant patterns, for example: (water begins), (smell smokers), (just heavenly)

In the final output we are left with opinion phrases which are in a <N,A> form:

[food Excellent, smell heavenly, food heavenly, treat satisfying].

Overall, this gave us the ability to understand in general what the user is mainly focusing on about a restaurant. Hence, can generate a more simplified and accurate frequent pattern, as all the user’s reviews are parsed into the general idea and focus points. The following figure outlines the entire process.

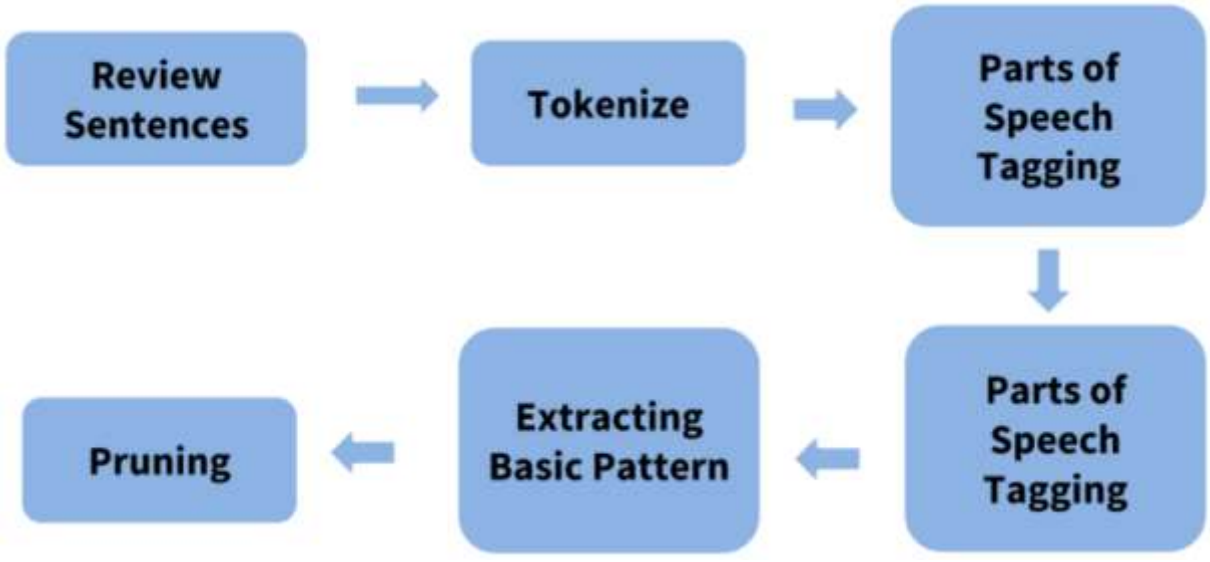


Fig: 2

V. Experiment

We used the pymining library in Python which contains Data Mining algorithms for frequent itemset and pattern mining. We implemented RElim (Recursive Elimination) algorithm. This algorithm is inspired from FP-Growth, however, it works without any prefix tree structure or complicated data structure. It also outperforms other Apriori and Elcat algorithms [10].

In our program we implemented this algorithm to generate the frequent itemsets in the opinion phrases that has been parsed from each of the reviews in each of the eight data frames we generate from the different combination of income level and restaurant ratings. We applied a 1% minimum support threshold for each of these datasets. Finally, we generated bar charts for each dataset to visualize the percentage of each opinion phrase relative to all the frequent opinion phrases generated.

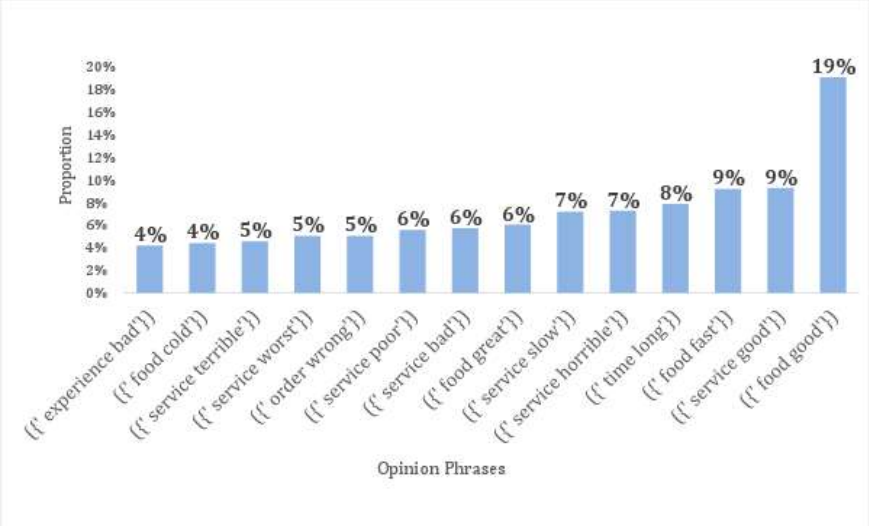
VI. Results

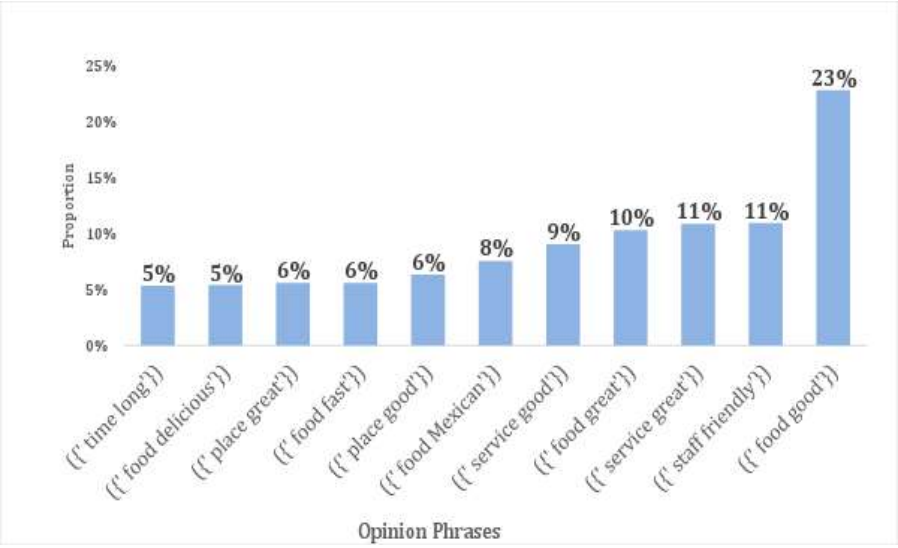
A. Low Income:

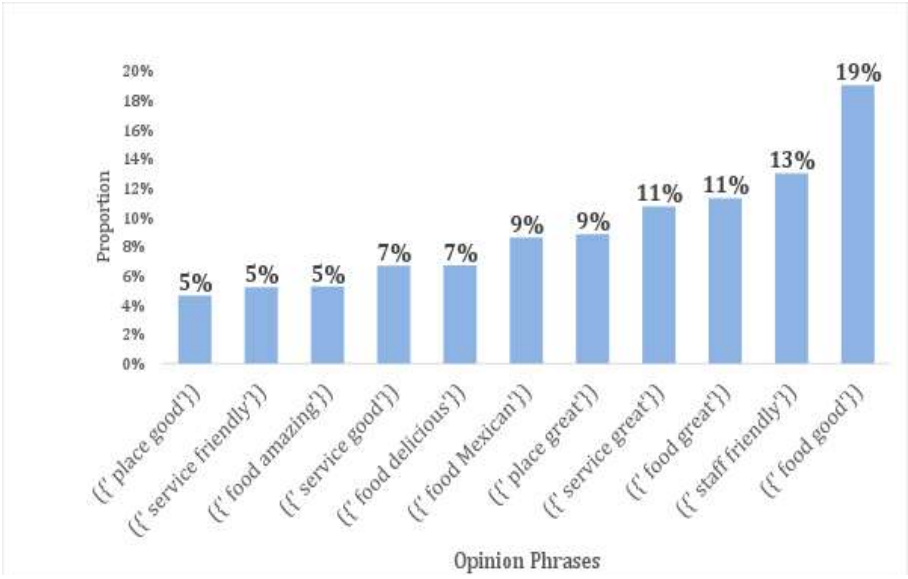
From the result we can observe the frequent review patterns in (1 - 2.5), (2.6 - 4) and above 4 star restaurants where we can see the type of food and service that appeared most in the result after going through the frequent mining algorithm. In the Income level 1 (low income) type restaurants with (above 4 stars) the most common opinion about the food is ‘food good’ at 19% and about the service is ‘service great’, “place great” and “staff friendly” comes at 11%, 9% and 13%

For the Income level 1 with (2.6 - 4 stars), the food quality is described with ‘food good’ at 23% which appeared most after frequent pattern mining. However, we see some contradiction in services for the low income restaurant reviews, since 5% of the reviews have ‘time long’ however ‘service great’ and ‘staff friendly’ has 11% each. Thus, it does not give a decisive answer for the level of service they provide.

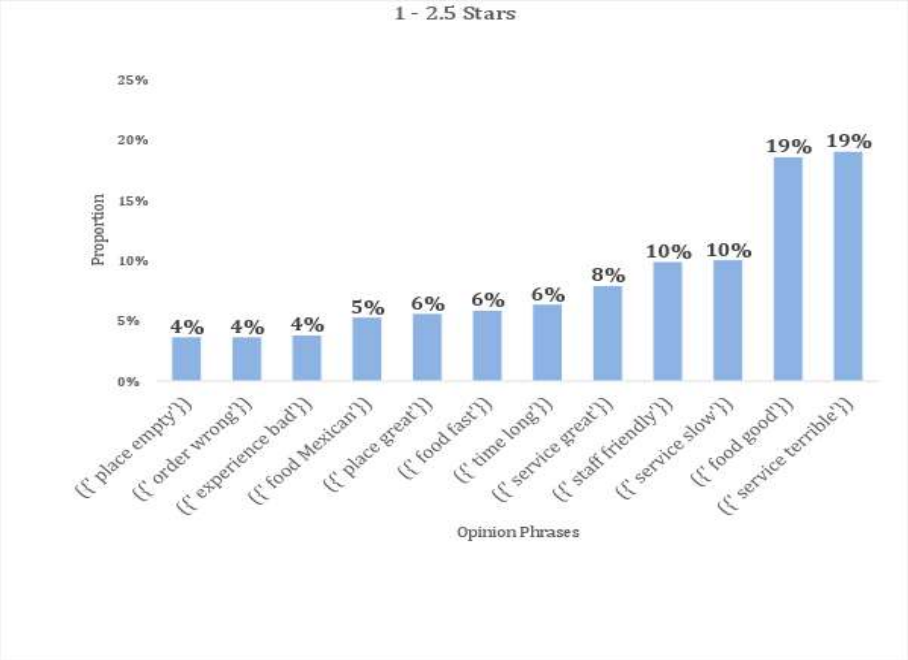
For (1-2.5 stars) it is clear that the service for these restaurants are not as good in service as that of 2.6-4 stars and above 4 star restaurants. Since “service horrible” and “service slow” comes at 7% and “experience bad” and “order wrong” comes at 4% and 5%. However, “food good” comes at 19%. Thus the low star ranked restaurants clearly has a bad customer service reputation even if the food is liked by just as many people.

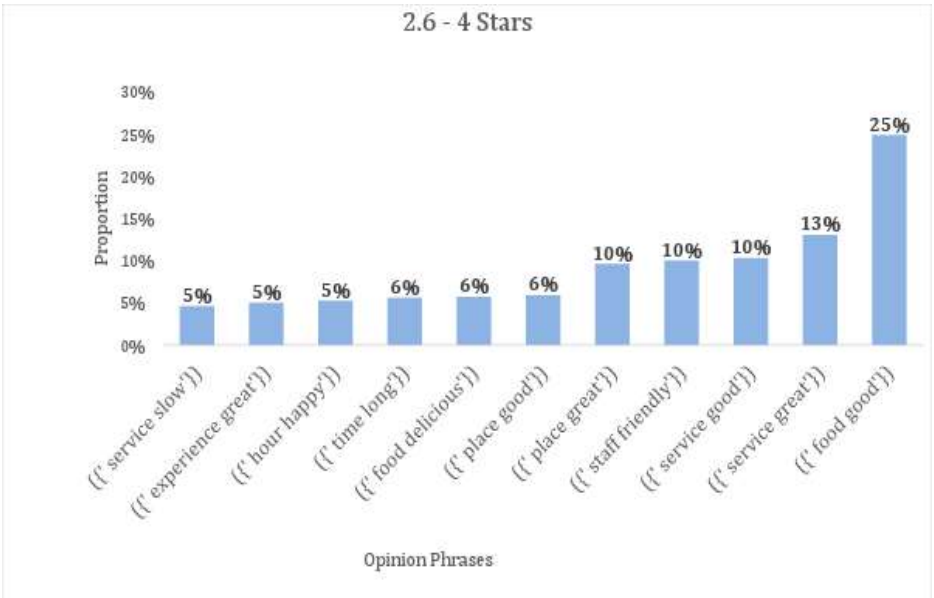

Proportion
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
4%
4%
5%
5%
5%
6%
6%
6%
7%
7%
8%
9%
9%
19%
{{' experience bad'}}
{{' food cold'}}
{{' service terrible'}}
{{' service worst'}}
{{' order wrong'}}
{{' service poor'}}
{{' service bad'}}
{{' food great'}}
{{' service slow'}}
{{' service horrible'}}
{{' time long'}}
{{' food fast'}}
{{' service good'}}
{{' food good'}}
Opinion Phrases

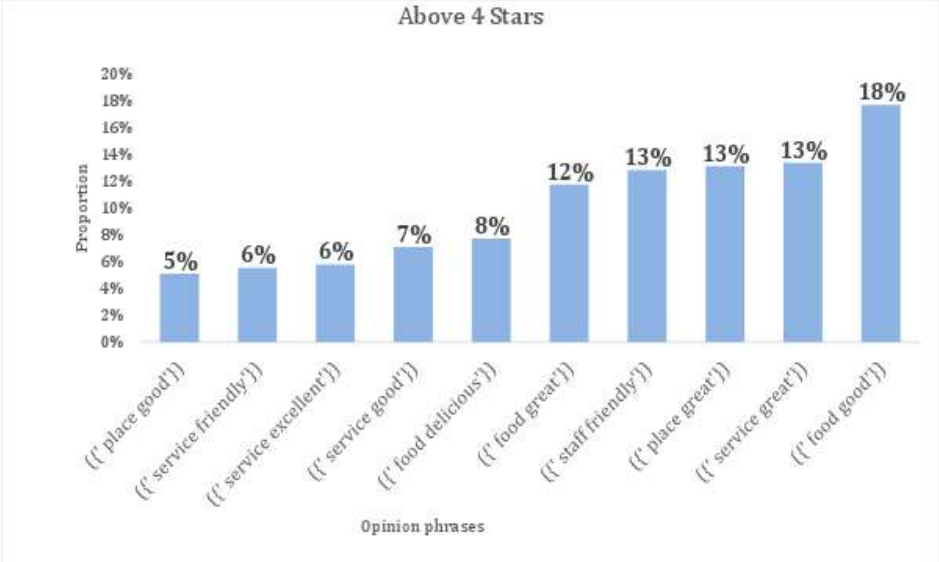

Proportion
25%
20%
15%
10%
5%
0%
5%
5%
6%
6%
6%
8%
9%
10%
11%
11%
23%
{{' time long'}}
{{' food delicious'}}
{{' place great'}}
{{' food fast'}}
{{' place good'}}
{{' food Mexican'}}
{{' service good'}}
{{' food great'}}
{{' service great'}}
{{' staff friendly'}}
{{' food good'}}
Opinion Phrases


Proportion
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
5%
5%
5%
7%
7%
9%
9%
11%
11%
13%
19%
{{' place good'}}
{{' service friendly'}}
{{' food amazing'}}
{{' service good'}}
{{' food delicious'}}
{{' food Mexican'}}
{{' place great'}}
{{' service great'}}
{{' food great'}}
{{' staff friendly'}}
{{' food good'}}
Opinion Phrases

B. Lower Middle Income:







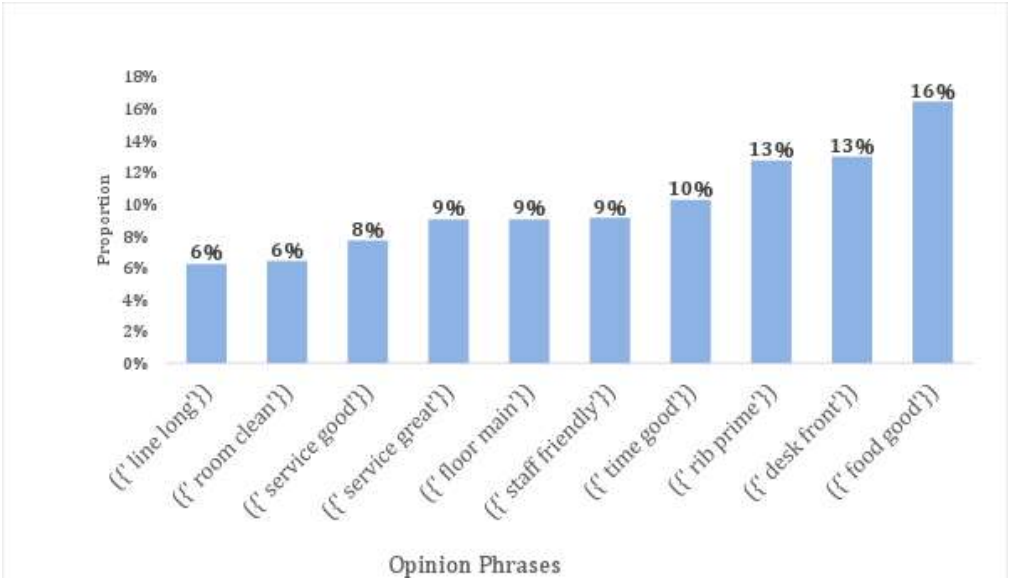
From the result we can observe the frequent review patterns in (1 - 2.5), (2.6 - 4) and above 4 star restaurants where we can see the type of food and service that appeared most in the result after going through the frequent mining algorithm. In the Income level 2 (lower middle level) type restaurants with (above 4 stars) the most common opinion about the food is 'food good' at 18% and about the service is 'service great' at 13%

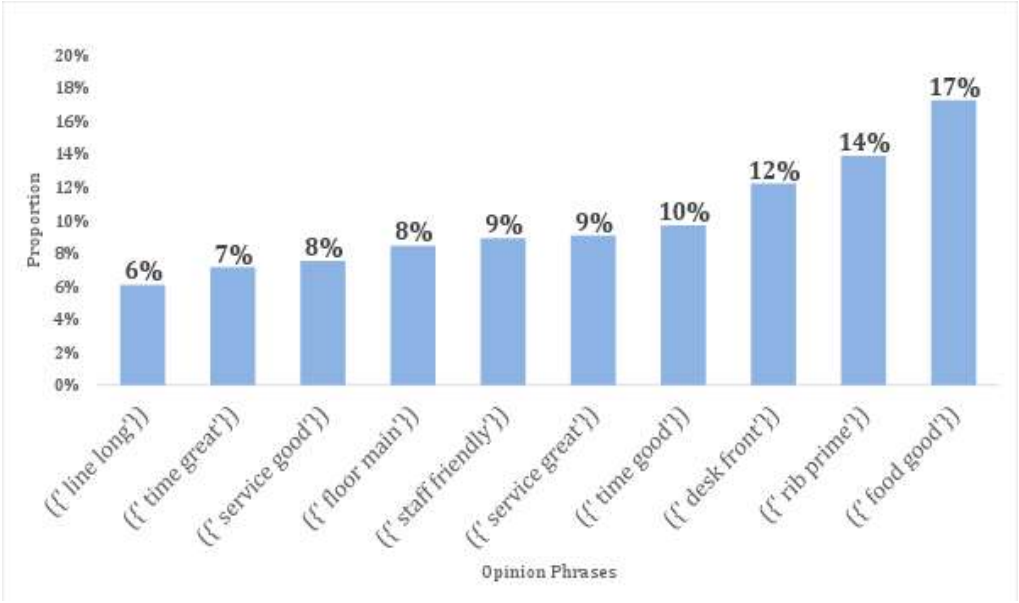
For the Income level 2 with (2.6-4 stars), the food quality is described with 'food good' at 25% which appeared most after frequent pattern mining. However, we see some contradiction in services for the lower middle income restaurant reviews, since 5% of the reviews have 'time long' and 5% has 'service slow', however 'service good' and 'staff friendly' has 10% each. Thus, it does not give a decisive answer for the level of service they provide.

For (1-2.5 stars) it is clear that the service for these restaurants are not as good in service as that of 2.6-4 stars and above 4 star restaurants. Since "service terrible" comes at 19% and "experience bad" and "order wrong" comes at 4%. However, "food good" comes at 19%. Thus the low star ranked restaurants clearly has a bad customer service reputation even if the food is liked by just as many people.

C. Higher Middle Income:

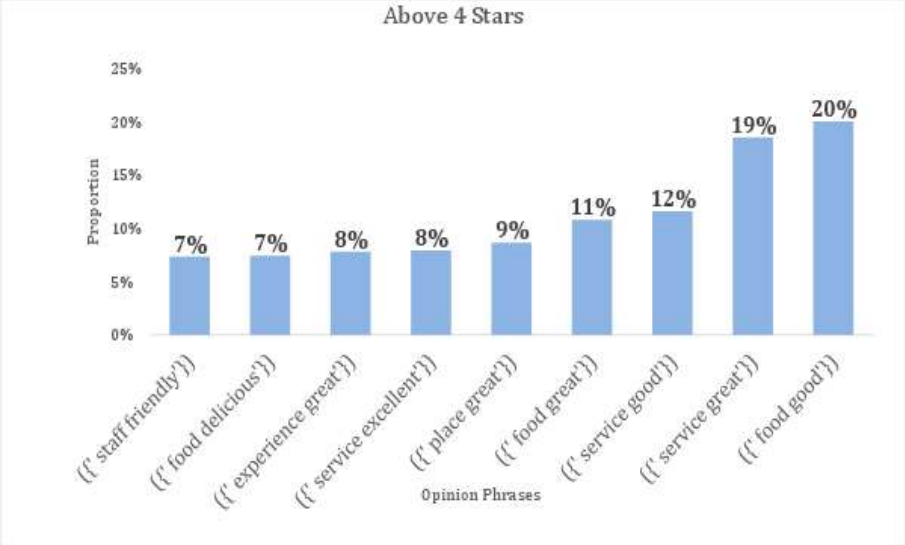
From the result we can observe the frequent review patterns in (1 - 2.5), (2.6 - 4) and above 4 star restaurants where we can see the type of food and service that appeared most in the result after going through the frequent mining algorithm. In the Income level 3 (Higher Middle Income) type restaurants with (above 4 stars) the most common opinion about the food is 'food good' at 20% and about the service is 'service great' at 19%





For the Income level 3 with (2.6-4 stars), the food quality is described with 'food good' at 17% which appeared most after frequent pattern mining. However, we see some contradiction in services for the higher middle income restaurant reviews, since 6% of the reviews have 'line long' whereas it also has "time good" 10%. Thus, it does not give a decisive answer for the level of service they provide.

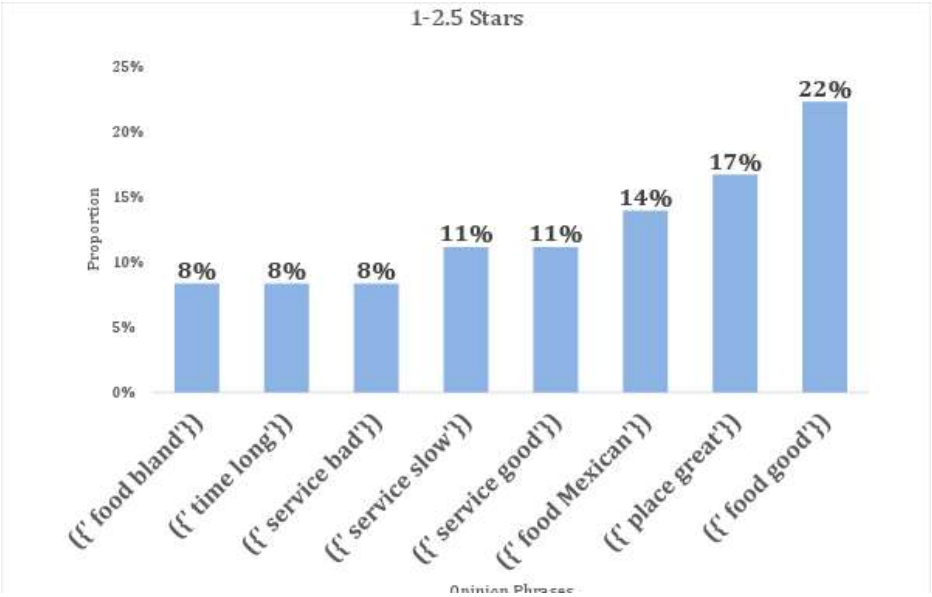
For (1-2.5 stars) it is clear that the service for these restaurants are not as good in service as that of 2.6-4 stars and above 4 star restaurants. Since "line long" comes at 6%. Even though, "food good" comes at 16%, the low star ranked restaurants clearly has a bad customer service reputation even if the food is liked by almost as many people.

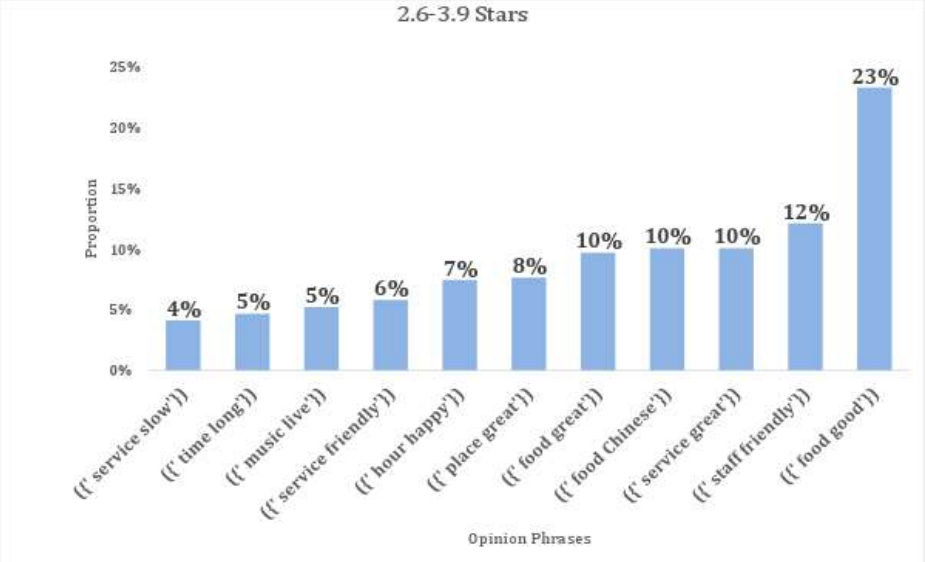


D. High Income:

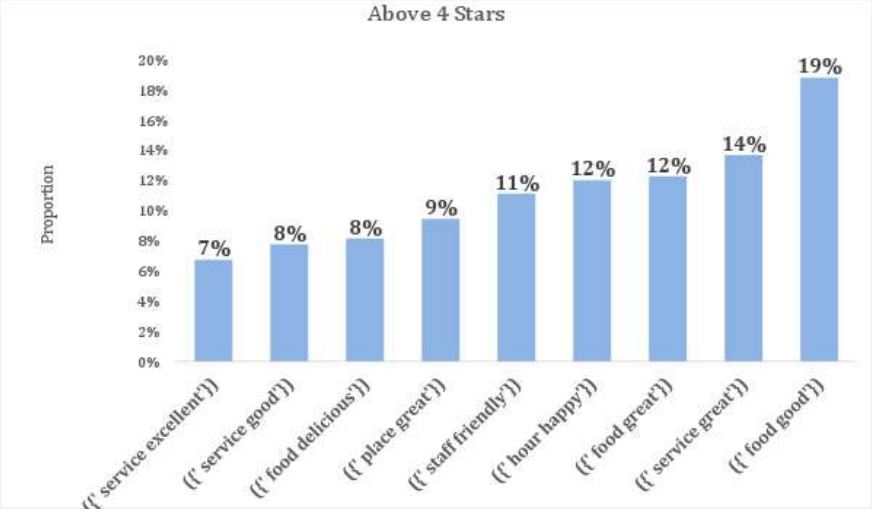
From the result we can observe the frequent review patterns in (1 - 2.5), (2.6 - 4) and above 4 star restaurants where we can see the type of food and service that appeared most in the result after going through the frequent mining algorithm. In the Income level 4 (High Income) type restaurants with (above 4 stars) the most common opinion about the food is ‘food good’ at 19% and about the service is ‘service great’ at 14%

For the Income level 4 with (2.6-4 stars), the food quality is described with ‘food good’ at 23% which appeared most after frequent pattern mining. However, we see some contradiction in services for the high income restaurant reviews, since 5% of the reviews have ‘time long’ and 4% has ‘service slow’, however ‘service great’ and ‘staff friendly’ has 10% and 12% each. Thus, it does not give a decisive answer for the level of service they provide.





For (1-2.5 stars) it is clear that the service for these restaurants are not as good in service as that of 2.6-4 stars and above 4 star restaurants. Since “service bad” and “service slow” comes at 8% and 11% and “food bland” comes at 8%. Even though, “food good” comes at 22%, the low star ranked restaurants clearly has a bad customer service reputation even if the food is liked by almost as many people.



VII. Conclusion

Among all the income levels we were able to identify four common themes that users generally wrote about. It was about how good the food is, the surroundings or place of the restaurant, and the staff or service of the restaurant. This finding can satisfy our initial motivation behind this project.

We are able to provide the key themes customers generally look for. However, there is not enough evidence to draw a concrete comparative analysis between the “low rated” (1-2.5 stars) and : high rated” (above 4 stars) restaurants. There was an obvious correlation between the positive phrases among themes. However, in the case of the restaurants with ratings between 1 - 2.5, stars we found a mixture of both positive and negative opinion phrases in these themes. It may also be surprising that the opinion phrase (food good) had the highest proportion in all the datasets focusing on restaurants below 2.5 stars, with an exception of the respective result in Lower Middle Income where (service terrible) had the highest proportion. There was enough similarity among the “low rated” restaurants to find enough direct frequent opinion phrases. In the Results of this paper, there are opinion phrases such as (service slow) or (time long). However, those opinion phrases lie in the lower proportion range in comparison to other positive opinion phrases among the results from the “low rated” restaurants.

In comparing the similar ratings across the income levels, it is quite interesting to see that we firstly found a similar distribution of the phrases. For instance, focusing on the restaurants with ratings between 1 and 2.5 stars, review patterns showed that food was both good and bad, service was both good and bad, and it took too long. In all four of the income levels we found the positive patterns to be in the general higher proportion region whereas the negative patterns to be in lower proportion region. We also found that restaurants above 2.6 stars did not really have any significant negative patterns of opinion phrases.

One of our primary deduction has been the case that certain restaurants which are relatively moderate may have reviews where people talk about the food being good and these restaurants may just have a lot more user reviews than the restaurants where users provide negative comments on certain aspects about the restaurants. Hence, it may just be the disproportionate number of user reviews driving the proportion of certain opinion phrases. However, as our findings were quite similar especially in the case of “low rated” restaurants across all four pairs of income level, we question how much can we depend on user reviews as a whole. User reviews are certainly subjective, and so it by mining patterns of opinion phrases, we may objectify these opinions. Although there is not much evidence in this research to make such a conclusion it could a good area to explore

VIII. Future Work

With this research we have been able to identify a couple of places and methods that can be implemented in the future to gain more accurate and better findings.

- The Review Processing takes on average $O(n^3)$ runtime. This is certainly expensive as we are mining through large sets of data. More heuristic based approaches could be taken to generate opinion phrases and training classification models.
- Sentiment scoring could also be used to apply weights to opinion phrases. Hence, this would give us the ability to understand how much importance each of the opinion phrases carry. Therefore, a more accurate proportion of each of the opinion phrases itemset can be generated to draw better comparisons.

Reference:

Bo Pang, Lillian Lee. (Pang and Lee, 2008) “Opinion mining and sentiment analysis”

[1] Sasank Channapragada, Ruchika Shivaswamy. “Prediction of rating based on review text of Yelp reviews.”
[2] Yun Xu, Xinhui Wu, Qinxia Wang. “Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews”
[3] Alexander Pak, Patrick Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”
[4] Ganu, Gayatree, Noemie Elhadad, and Amelie Marian. "Beyond the Stars: Improving Rating Predictions using Review Text Content." WebDB. Vol. 9. 2009.

[5] Qu, Lizhen, Georgiana Ifrim, and Gerhard Weikum. "The bag-of-opinions method for review rating prediction from sparse text patterns." Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.

[6] Susan E. Chen, Jing Liu, and James K. Binkley. “An exploration of the relationship between income and Eating Behavior”

[7] Saeideh Bakhshi, Partha Kanuparthy, Eric Gilbert. “A Review of Food Service Selection Factors Important to the Consumer”
[8] Vlad Sandulescu “Opinion Phrases Mining Algorithm”
https://github.com/vladsandulescu/phrases

[9] Abbasi Moghaddam, Samaneh. “Aspect-based opinion mining in online reviews”

[10] “Find Frequent Itemsets with the RElim Algorithm (Recursive Elimination)”
http://www.borgelt.net/doc/relim/relim.html