ISM 6136: DATA MINING PROJECT

YELP DATA ANALYSIS AND RATING PREDICTION

Neti Sheth

Adarsh Patel

Ninad Mehta

Table of Contents

Introduction

Yelp is a regional directory platform and review site with social networking tools. The website provides crowd-sourced reviews for local businesses (spas, restaurants, department stores, pubs, home-local services, shops, cars). This helps users to give business ratings and reviews. Normally, the review is a brief text composed of a few lines of about 100 words describing various user experiences with respect to various dimensions. This offers business owners the opportunity to improve their products and customers to choose among the best available sector.

The goal is to design a system that will use existing yelp data to provide insightful analysis and to assist existing business owners, future business owners to make important decisions about a new business or business expansion.

Although there is just one score (one to five stars) received by a business review on Yelp, the comment text reveals a more complicated story. A customer may have loved the food, but the service might have irritated them; the design may have inspired them, but the long wait turned them off. The objective is to perform sentiment analysis to separate positive reviews from negative and topic modeling on the reviews in order to automatically reveal the underlying topics discussed by the reviewers and also predict the rating that the restaurant would have received based on each review.

Business Value

Not only in personal life, but also in business, emotions are important. In addition, how the customers feel about goods and your brand provides the requisite information to determine the advertising or interaction approach. The management constantly wants to know about customer expectations, experiences (good and bad), favorites, what they miss, likes and dislikes from feedback they have received in order to improve the restaurant services and qualities.

The management might not have enough time to go through each and every review. If valuable information and insights can be provided to them at a glance, it can be really helpful and time-saving. And not only for the management but also for the customers who are trying to know more about the restaurant and need some help in ordering or selecting the restaurants. After all, in today’s world, every person prefers to read reviews and feedbacks before they take a decision.

In our project, we have used Natural Language Processing and Machine Learning to achieve these business and customer goals. We have focussed on Sentiment Analysis, Topic Modelling, Data Analysis, and Classification for rating prediction

Data Source and Preparation

The Yelp dataset is a subset of companies, ratings, reviews and user data for personal, educational and academic purposes available as JSON files. The dataset is available at https://www.yelp.com/dataset. It contains 6,685,900 reviews for 192,609 businesses based in 10 metropolitan areas.

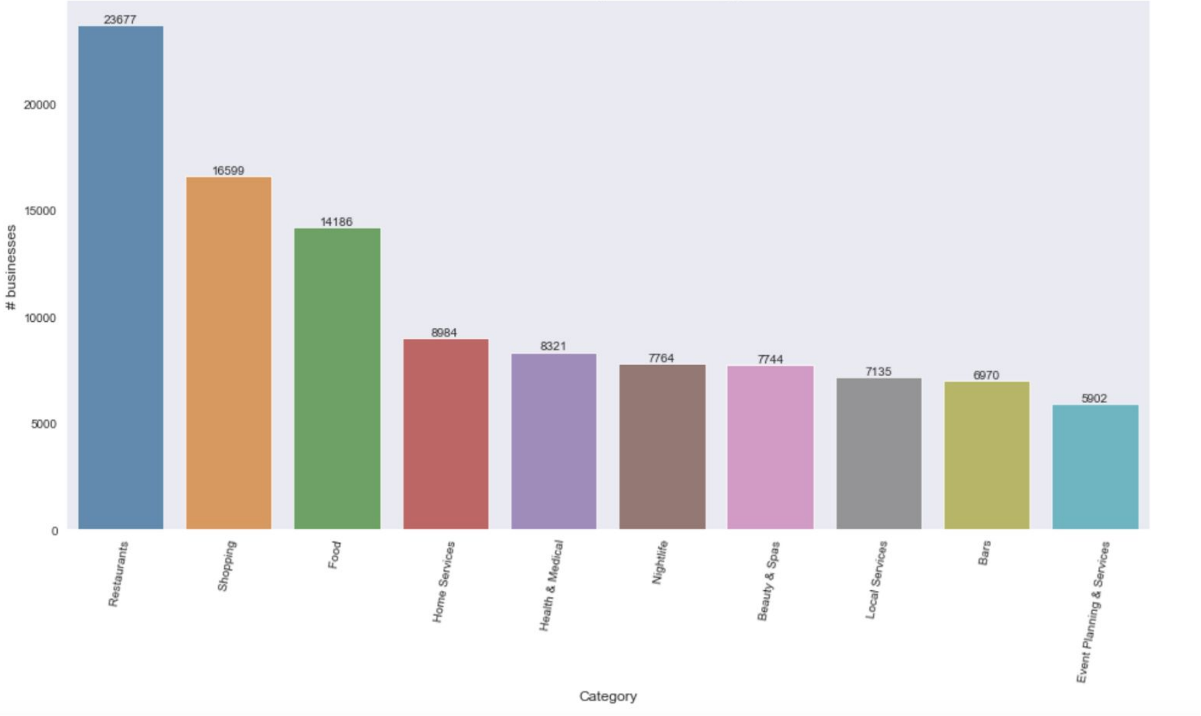
As a part of the data preparation and cleaning process, we corrected the datatypes of every column in the dataset and dropped all the unnecessary columns. We wanted to focus on only one city which had the maximum no of reviews for restaurants. In this case, it was Las Vegas. Hence, we filtered the dataset records to keep only reviews for restaurants in Las Vegas. The resultant dataset had 1,242,711 reviews for 6450 restaurant businesses.

Once the dataset was ready, the next important task was to clean the text of the reviews for analysis. The task, in particular, involved tokenizing text into words, removing stop words and words with less than three characters, and lemmatizing words.

Exploratory Data Analysis

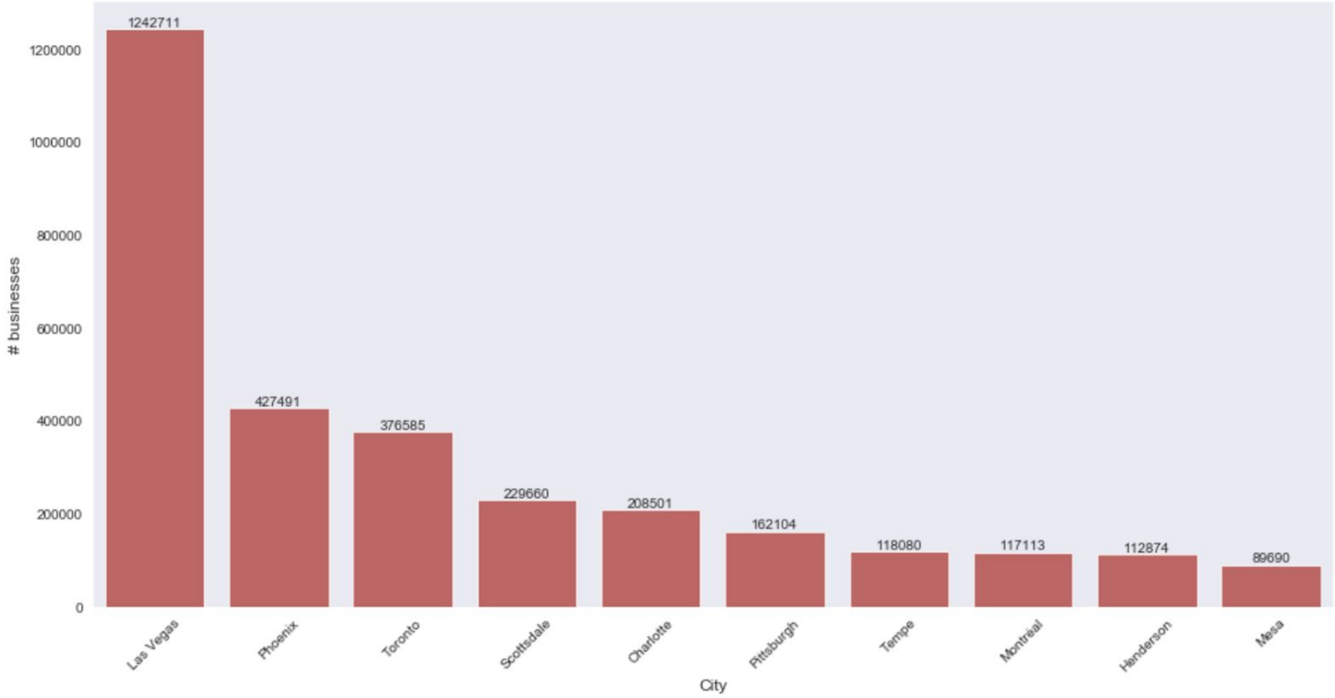
We used 6,685,900 reviews for 192,609 businesses to do overall data exploration. We addressed the following questions to quickly get a preliminary understanding of the dataset:

1. **Which are the top 10 categories with the maximum number of reviews?**



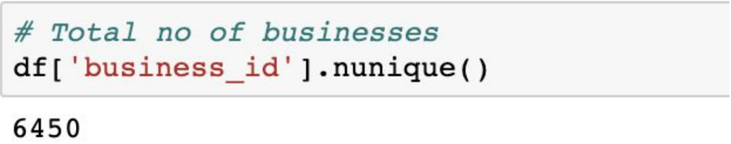
Restaurants are the most reviewed business category followed by shopping, food and home services. It is always better to focus on one category for analysis in order to get relevant insights from data. Hence, we filtered the reviews in the dataset to keep only restaurant category reviews.

2. **Which city has the maximum number of restaurant reviews?**



Las Vegas has the maximum number of reviews. We again filter the reviews to keep the Las Vegas restaurant reviews.

3. **Las-Vegas has how many restaurant businesses?**

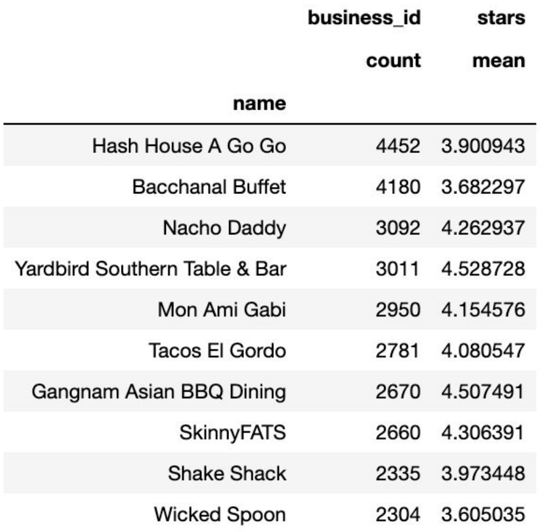


```
# Total no of businesses
df['business_id'].nunique()

6450
```

Las Vegas has 6450 restaurant businesses registered on Yelp.

4. **Which are the 10 top-rated and popular restaurants in Las Vegas since 2016?**

Based on the total no of reviews and average rating, the 10 top-rated and popular restaurants in Las Vegas since 2016 are:

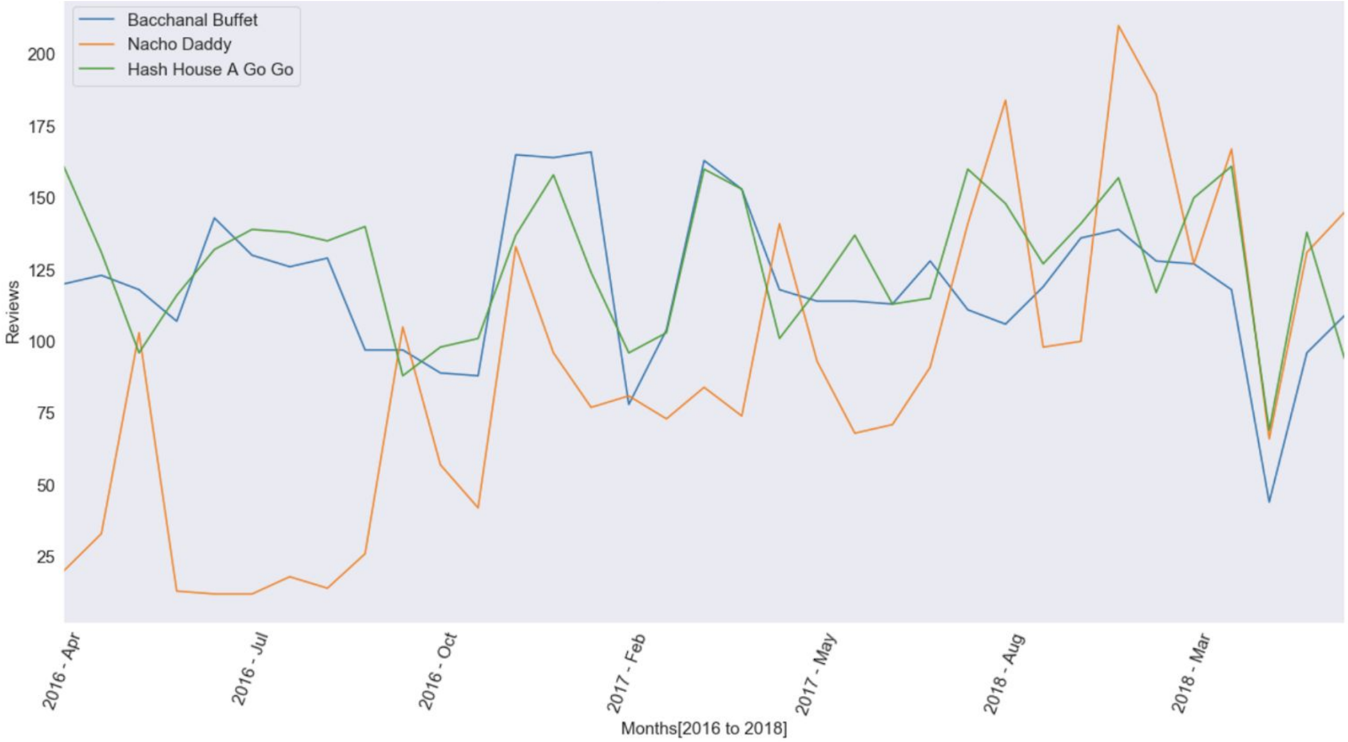


	business_id	stars
	count	mean
name		
Hash House A Go Go	4452	3.900943
Bacchanal Buffet	4180	3.682297
Nacho Daddy	3092	4.262937
Yardbird Southern Table & Bar	3011	4.528728
Mon Ami Gabi	2950	4.154576
Tacos El Gordo	2781	4.080547
Gangnam Asian BBQ Dining	2670	4.507491
SkinnyFATS	2660	4.306391
Shake Shack	2335	3.973448
Wicked Spoon	2304	3.605035

Analysis: 3 best-rated restaurants in Las Vegas

The 3 top-rated restaurants in Las Vegas since 2016 are Hash House A Go Go, Bacchanal Buffet and Nacho Daddy. To derive meaningful insights from the positive and negative reviews for all three restaurants, we performed sentiment analysis and topic modeling on the reviews.

- **Reviews posted over time**

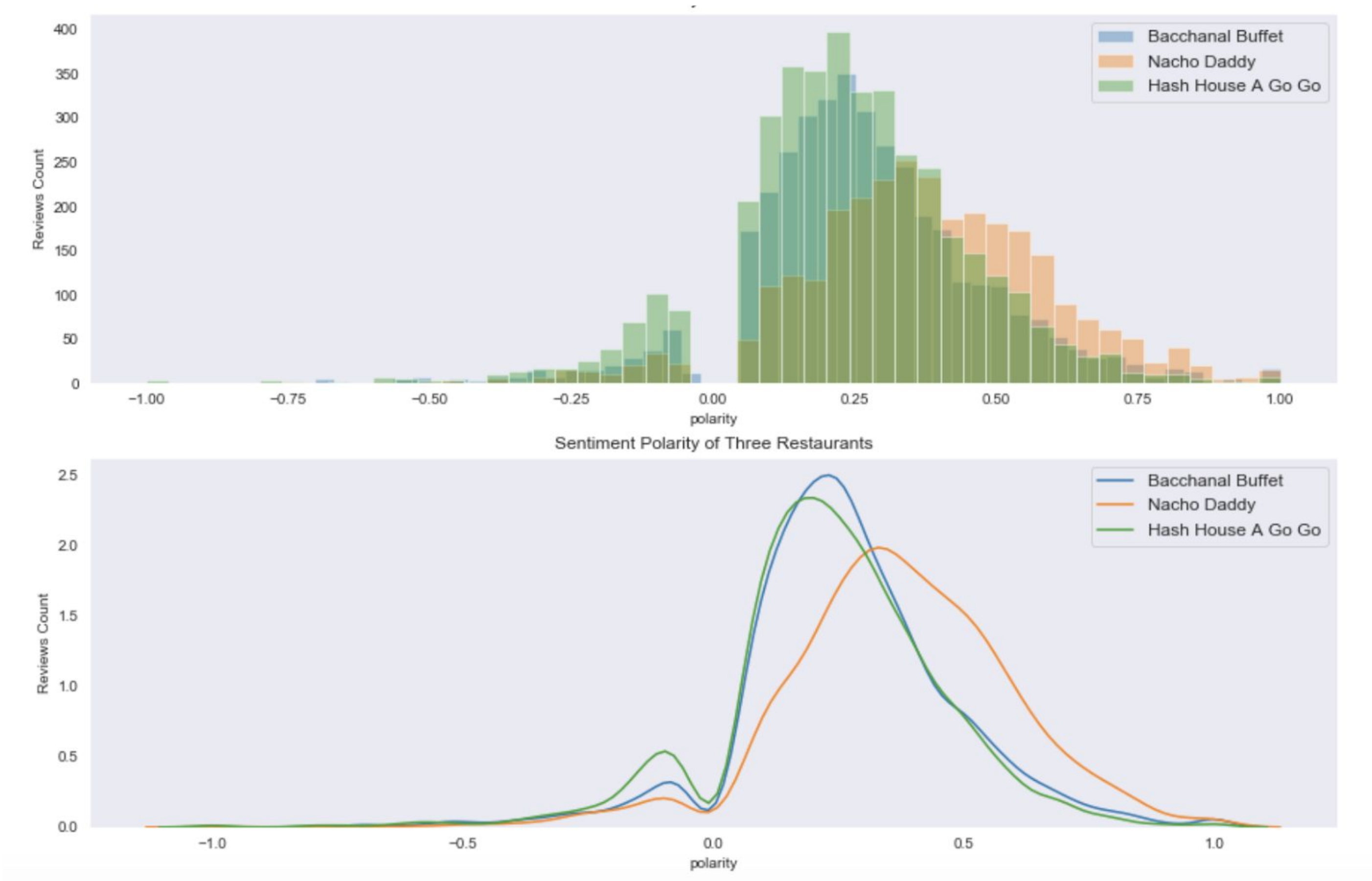


From April 2016 to December 2018, we can observe how the number of reviews for each restaurant varies. For example, for Nacho Daddy the reviews dropped from 100 in May 2016 to 20 in June 2016. It stayed the same for quite some time until Sept 2016. This information can be useful for the management to evaluate the possible reasons behind the drop and gain in number of reviews and when it happened.

- **Sentiment Analysis**

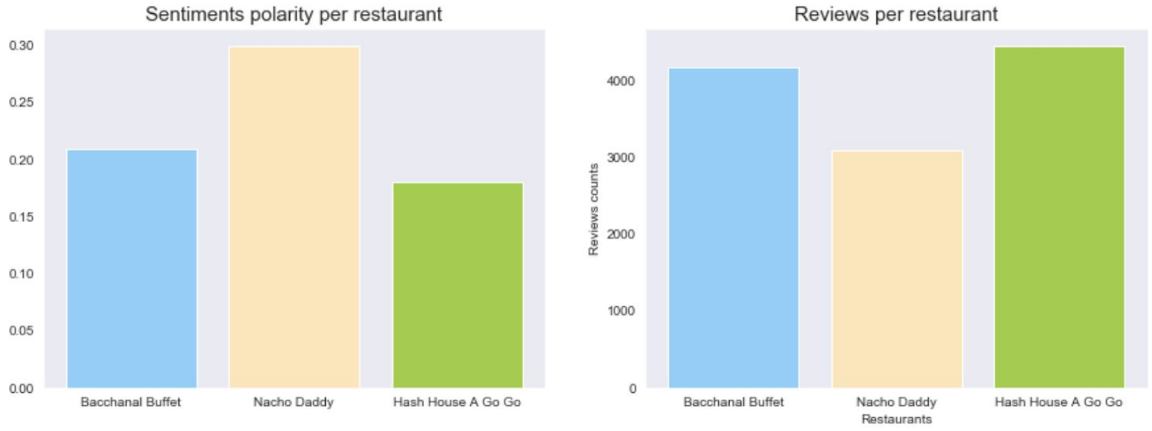
After cleaning the data, we conducted sentiment analysis to separate the positive reviews from the complaints. We used **TextBlob** for this purpose. TextBlob is a python library that provides a simple API for accessing its methods and carrying out basic NLP tasks. Textblob's sentiment function returns two properties, polarity, and subjectivity. Polarity is float, which is within the range of[-1,1] where 1 means positive and -1 means negative. Based on the polarity value, we classified the review as positive or negative.

- **Sentiment polarity distribution for the restaurants**



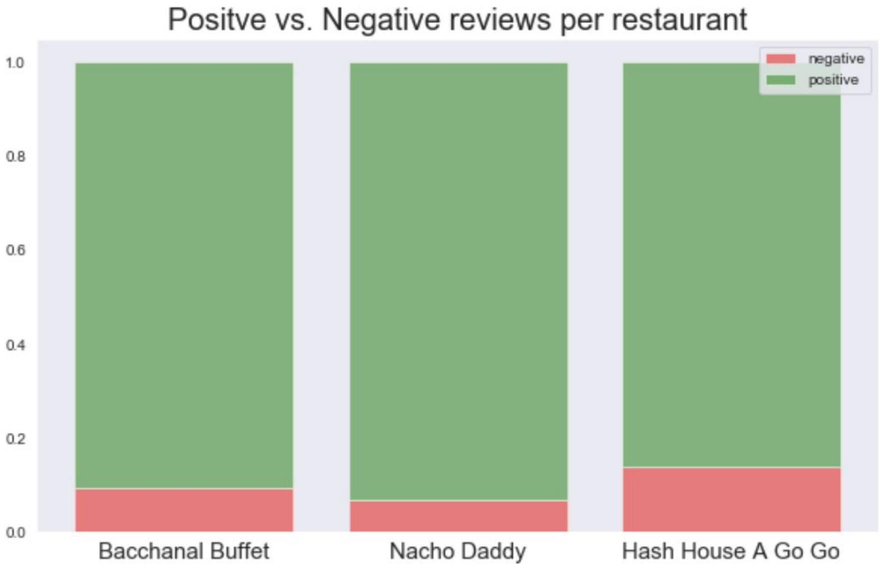
We can observe how the sentiment varies for each review for different restaurants. It helps us to understand how many reviews were positive and negative. For example, Hash House A Go Go had the maximum number of reviews with -0.1 polarity (negative). Most of the reviews for each restaurant were positive and were between 0.1 and 0.6.

- **Overall sentiment polarity of each restaurant**



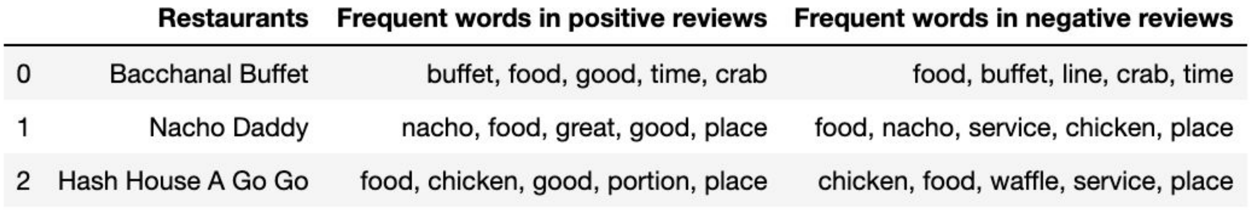
Overall sentiment score for all the restaurants is positive as expected but its more on the neutral side. Hash House A Go Go gets the lowest sentiment score and Nacho Daddy gets the highest. An interesting observation is that the restaurant with more reviews has less sentiment polarity score.

- **Positive vs Negative reviews for each restaurant**



We can see that for all the three carriers, the number of positive reviews is more than negative reviews. But as the number of reviews for a restaurant increases, the proportion of negative reviews also increases. Hash House A Go Go has maximum reviews and more negative reviews compared to others.

- **Representative frequent words for positive and negative reviews**



	Restaurants	Frequent words in positive reviews	Frequent words in negative reviews
0	Bacchanal Buffet	buffet, food, good, time, crab	food, buffet, line, crab, time
1	Nacho Daddy	nacho, food, great, good, place	food, nacho, service, chicken, place
2	Hash House A Go Go	food, chicken, good, portion, place	chicken, food, waffle, service, place

The frequent words in positive and negative reviews are quite similar. It’s good to have this as the starting point. We can see that reviewers talked a lot about the portion size in positive and service in negative reviews for Hash House A Go Go.

We cant make any conclusions from these results. We need to dive deeper into the reviews and understand what they talk about - the topics.

- **Topic Modelling**

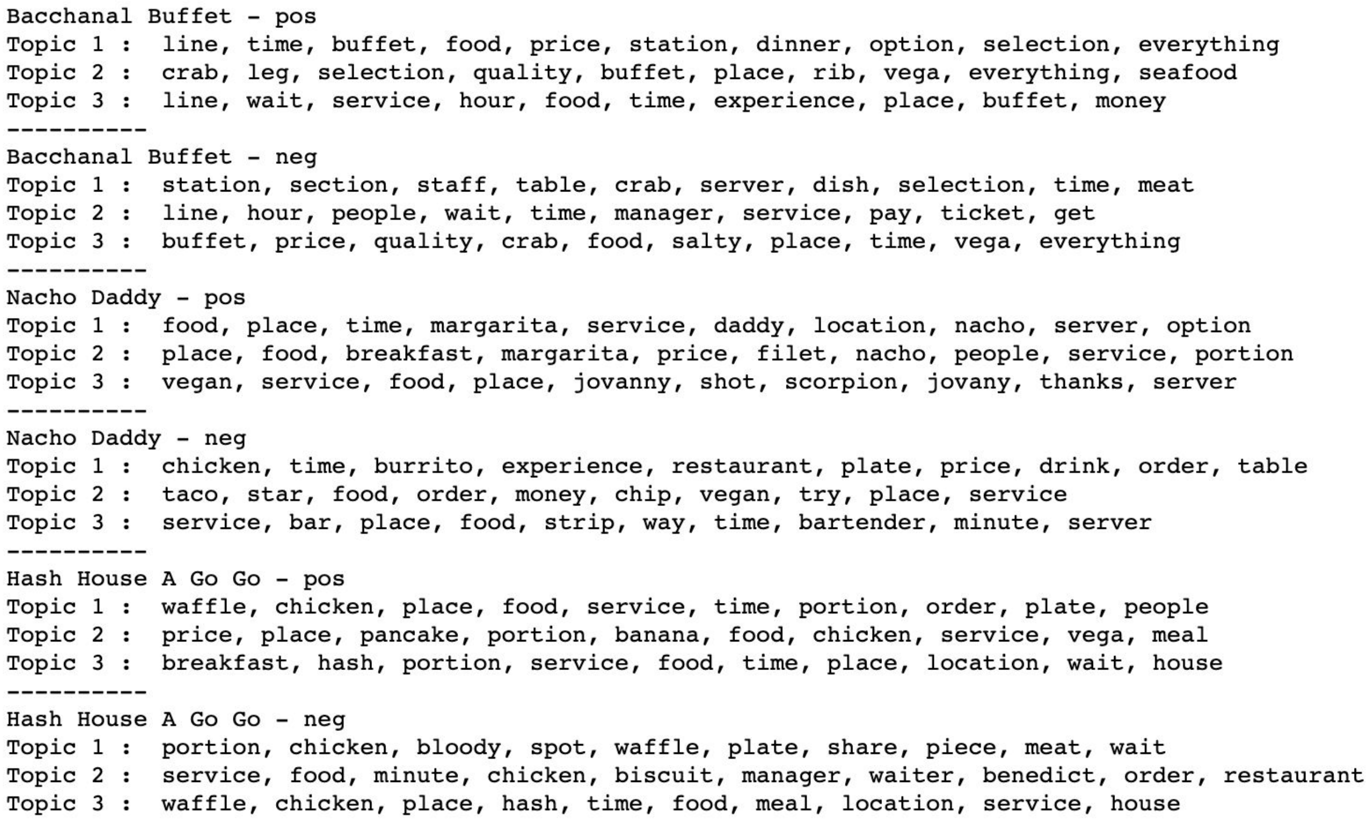
After separating the negative from the positive reviews, we built topic models using **TF-IDF(Term Frequency - Inverse Document Frequency)** and **LDA(Latent Dirichlet Allocation)** from the gensim package to understand what reviewers are talking about in the reviews.

Words in a sentence can be represented in terms of their counts (frequency of occurrence), presence vs. absence, weighted counts, or a similarity vector. If the ordering of words is not preserved, it is a bag of words approach. The distinct words in all the reviews compose our Bag of Words and then we use TF-IDF to re-weight the count features. This will help to reduce the importance of more frequent(which are generally less valuable) words and enhance the importance of rare yet more valuable words.

LDA model is used to fetch three topics for each restaurant's positive and negative reviews and the ten terms that accompany them. LDA is designed to uncover some of the latent subjects in a document corpus. It applies Bayesian probabilities to a bag of terms (in our case, TF-IDF).

It uses co-occurring keywords to arrange and interpret large collections of textual data by identifying obscure topics and annotating documents with them. It is assumed that the topic distribution has a sparse Dirichlet prior, which means that only a small collection of topics are covered by documents and that subjects use only a small collection of frequent words.

The topics for positive and negative reviews for the three top restaurants in Las Vegas are:



```
Bacchanal Buffet - pos
Topic 1 :  line, time, buffet, food, price, station, dinner, option, selection, everything
Topic 2 :  crab, leg, selection, quality, buffet, place, rib, vega, everything, seafood
Topic 3 :  line, wait, service, hour, food, time, experience, place, buffet, money
----------
Bacchanal Buffet - neg
Topic 1 :  station, section, staff, table, crab, server, dish, selection, time, meat
Topic 2 :  line, hour, people, wait, time, manager, service, pay, ticket, get
Topic 3 :  buffet, price, quality, crab, food, salty, place, time, vega, everything
----------
Nacho Daddy - pos
Topic 1 :  food, place, time, margarita, service, daddy, location, nacho, server, option
Topic 2 :  place, food, breakfast, margarita, price, filet, nacho, people, service, portion
Topic 3 :  vegan, service, food, place, jovanny, shot, scorpion, jovany, thanks, server
----------
Nacho Daddy - neg
Topic 1 :  chicken, time, burrito, experience, restaurant, plate, price, drink, order, table
Topic 2 :  taco, star, food, order, money, chip, vegan, try, place, service
Topic 3 :  service, bar, place, food, strip, way, time, bartender, minute, server
----------
Hash House A Go Go - pos
Topic 1 :  waffle, chicken, place, food, service, time, portion, order, plate, people
Topic 2 :  price, place, pancake, portion, banana, food, chicken, service, vega, meal
Topic 3 :  breakfast, hash, portion, service, food, time, place, location, wait, house
----------
Hash House A Go Go - neg
Topic 1 :  portion, chicken, bloody, spot, waffle, plate, share, piece, meat, wait
Topic 2 :  service, food, minute, chicken, biscuit, manager, waiter, benedict, order, restaurant
Topic 3 :  waffle, chicken, place, hash, time, food, meal, location, service, house
```

From the results, we can observe the following for each restaurant:

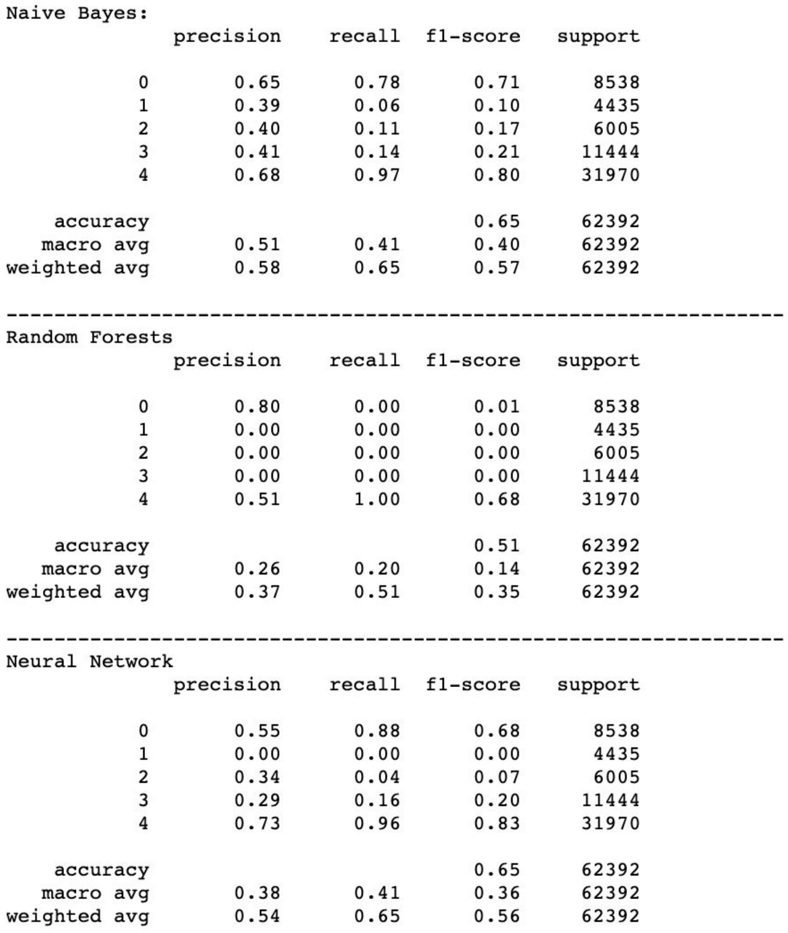
- Bacchanal Buffet: Reviewers probably liked the quality of seafood and buffet and were dissatisfied with the manager, service and wait time.
- Nacho daddy: Reviewers praised margarita pizza, location, portion size, and breakfast but left negative reviews for bartenders, tacos, burritos, and servers.
- Hash House A Go Go: Reviewers left positive feedback for pancakes, vega, and portion size but were not happy with biscuits, benedict, manager, waiter, and service.

These insights are really helpful for management and future customers. They can easily identify the strengths and weaknesses of each restaurant. Management can use this to improve customer service and loyalty. Future customers can make decisions based on what dishes are mentioned the most and what to expect from a restaurant.

Rating Prediction

There can be times when a user leaves a review and no rating for the restaurant. To predict the rating that the restaurant would have received based on each review we have used three different models - Naive Bayes, Random Forest, and Neural Network. We partitioned the dataset into training(70%) and testing(30%) dataset. Using the TF-IDF vectorizer, we transformed the clean preprocessed reviews text and used it as an input for all the three models.

Comparing the three models to pick the best



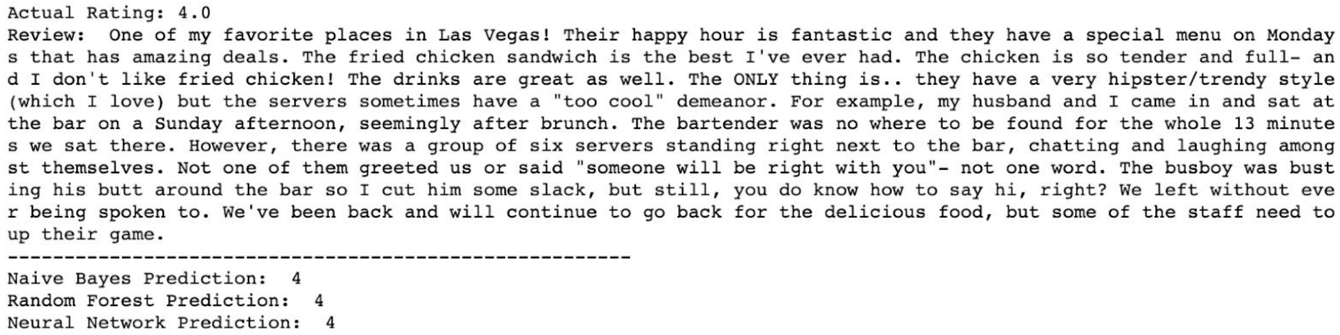
```
Naive Bayes:
              precision    recall  f1-score   support

           0       0.65      0.78      0.71      8538
           1       0.39      0.06      0.10      4435
           2       0.40      0.11      0.17      6005
           3       0.41      0.14      0.21     11444
           4       0.68      0.97      0.80     31970

    accuracy                           0.65     62392
   macro avg       0.51      0.41      0.40     62392
weighted avg       0.58      0.65      0.57     62392

-----------------------------------------------------------------
Random Forests
              precision    recall  f1-score   support

           0       0.80      0.00      0.01      8538
           1       0.00      0.00      0.00      4435
           2       0.00      0.00      0.00      6005
           3       0.00      0.00      0.00     11444
           4       0.51      1.00      0.68     31970

    accuracy                           0.51     62392
   macro avg       0.26      0.20      0.14     62392
weighted avg       0.37      0.51      0.35     62392

-----------------------------------------------------------------
Neural Network
              precision    recall  f1-score   support

           0       0.55      0.88      0.68      8538
           1       0.00      0.00      0.00      4435
           2       0.34      0.04      0.07      6005
           3       0.29      0.16      0.20     11444
           4       0.73      0.96      0.83     31970

    accuracy                           0.65     62392
   macro avg       0.38      0.41      0.36     62392
weighted avg       0.54      0.65      0.56     62392
```

From the classification report, we can see how the three models performed on the same dataset and classified the reviews in ratings on a scale of 1 to 5. Out of all the three models, Naive Bayes and Neural Network performed well with 65% accuracy. Random Forest accuracy was poor - 51%.

The training time for Neural Network was 3 minutes whereas, for Naive Bayes, it was just a few seconds. The model performance can still be improved with hyperparameter tuning and we are working on it.

Prediction Example

We used one of the review text from the dataset to test how the three models predicted.



```
Actual Rating: 4.0
Review:  One of my favorite places in Las Vegas! Their happy hour is fantastic and they have a special menu on Monday
s that has amazing deals. The fried chicken sandwich is the best I've ever had. The chicken is so tender and full- an
d I don't like fried chicken! The drinks are great as well. The ONLY thing is.. they have a very hipster/trendy style
(which I love) but the servers sometimes have a "too cool" demeanor. For example, my husband and I came in and sat at
the bar on a Sunday afternoon, seemingly after brunch. The bartender was no where to be found for the whole 13 minute
s we sat there. However, there was a group of six servers standing right next to the bar, chatting and laughing among
st themselves. Not one of them greeted us or said "someone will be right with you"- not one word. The busboy was bust
ing his butt around the bar so I cut him some slack, but still, you do know how to say hi, right? We left without eve
r being spoken to. We've been back and will continue to go back for the delicious food, but some of the staff need to
up their game.
-------------------------------------------------------
Naive Bayes Prediction:  4
Random Forest Prediction:  4
Neural Network Prediction:  4
```

The actual rating for the review was 4 and all the classifier predicted it as 4 correctly.

References

- https://www.yelp.com/dataset/documentation/main
- https://textblob.readthedocs.io/en/dev/quickstart.html
- https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
- https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/
- https://www.kaggle.com/ambarish/a-very-extensive-data-analysis-of-yelp