Analysis of ratings between Yelp and Zomato

Donovan Jeremiah, Shrey Shah and
Mubeen Ali Mohammed
National College of Ireland
Dublin , Ireland
x18181562@student.ncirl.ie,
x18192271@student.ncirl.ie,
x18180370@student.ncirl.ie

***Abstract*—21st Century has observed a fast-paced digital revolution which has led every industry to adopt digitalization. Many users with smartphones rely on online food review platforms like Yelp, Zomato to decide a place for any occasion like parties, client dinners, festival celebrations. Customers have started taking decisions based on reviews and learning from experiences shared by online community members.**

The first objective of our project is to identify if there are any difference in ratings between yelp and Zomato for identical restaurants and investigate what are the underlying factors and recognizing patterns associated with the rating difference. The second objective is to analyze reviews data (textual data). 391 common restaurants from Yelp and Zomato are being analyzed based on Business, Review, Users, tip data. Performed analysis to find trends on difference of ratings with respect to different cuisines across cities and state, Geospatial map of restaurants with high ratings differences, Text analytics using word cloud, Sentiment analysis of reviews to understand key aspects about restaurants. This project is implemented using Python language.

Keywords*— Yelp business, Zomato API, Rating Differences, Star ratings and reviews, Cuisines, Restaurants, Sentiment analysis, Google Colab, MongoDB, PostgreSQL, Jupyter)

I. Introduction

In this age of digitalization, data has become the most valuable resource in the world [1]. There is abundant data generated and consumed by netizens present all over the world by which every person takes their informed decisions based on reviews, experiences shared by users of community in many countries.

Food industry is one of the interesting industries in which customers irrespective of their race, ethnicity and age want to have the pleasure of trying out different types of cuisines and get value for their money satisfying their palate. Hence, successful online review platforms like Yelp and Zomato has been very reliable source for a user to decide best place to eat a particular type of cuisine based on online reviews.

Yelp is a San Francisco headquartered business directory service widely used in US and Canada. Zomato is Indian online review service now also available in US, Canada is specific to Restaurants. Many researchers have published various types of analysis using the publicly available Yelp and Zomato unstructured/semi structured data through API. We have carried out this analysis to understand the factors.

A. Motivation

Many papers had only studied Yelp and Zomato individually. We did not find papers which compared reviews for both sites. Hence, we have identified patterns based on their comparison for businesses to attract users by incorporating the expectations gap in either of review sites.

B. Research Questions

- Are there any differences in ratings between Yelp and Zomato for restaurants?
- Can word cloud explain yelp ratings?

II. Related Work

Many researchers have done extensive analysis of Yelp and Zomato datasets separately as the datasets have beenmade publicly available for developers and researchers to help connect customers and various business via online platfoms in the Food service Industry. Various type of analysis has been done in the past using Big Data Framework, Python, R in the form of Statistical analysis, Machine learning and Deep learning.

The authors in the research [2] carried out analysis on Yelp to identify if there is a strong or weak association between the star ratings and sentiment derived from the textual portion of consumer reviews. The implication of their research was that only some aspects significantly influenced the ratings. The limitations were that the relationship was not established, and ratings were not entirely describing the review sentiments as only 1000 random review comments were selected.

The journal paper [3] discusses about how online reviews affect demand of restaurants. The results suggested that 5-9 percent increase in revenue was led by one-star increase in Yelp rating and this effect was driven by independent restaurants. Chain restaurant were not affected by rating however, their market share declined as result of Yelp outreach.

III. Methodology

We are using KDD approach which stands for Knowledge discovery in databases [5].

A. Dataset Description

1) Yelp Datasets:

a) Business: 192609 Records
b) User: 1637138 Records
c) Review: 6685900 Records
d) Check-in: 161950 Records
e) Tip: 1223094 Records

2) Zomato Datasets:

a) Restaurant: 1200 Records

3) Technologoies Used:

- Python: It has a variety of packages that facilitate a thorough data analysis with interactive plots.
- Jupyter Notebook: Jupyter's ability to combine code, interactive plots, and markdown for

documentation, all in the same file, is the reason for choosing it for this data analysis project

- MongoDB: Due to MongoDB's ability for faster read/write operations, its lightweight implementation, and its simple querying language as Javascript made it the choice of storage for the unstructured data extracted from Zomato's API and Yelp's JSON files
- PostgreSQL: Since the transformed data was structured, PostgreSQL was used. It also worked well with python and the python package psycopg2 that was used to connect to it
- Google Collab: Due to limited memory and computational power on local machines, Google Collab was used for sentiment analysis of the large review dataset (6 GB)

4) Python Packages Used:

- Data Processing: Pandas, Numpy, Json
- Data Storage and Retrieval: Pymongo and Psycopg2 were used to connect to MongoDB and PostgreSQL
- Data Visualization: Folium, Seaborn, Matplotlib, NLTK, Wordcloud, and Textblob.

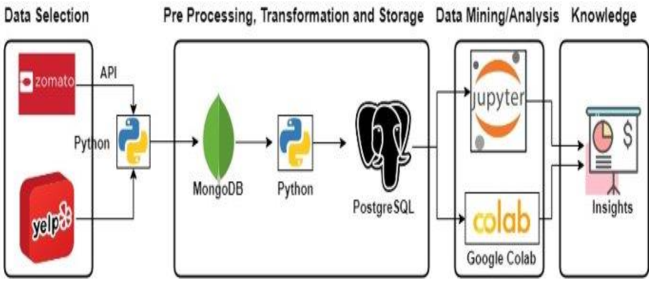


Fig. 1. KDD Methodology, Fayyad et al 1996

B. Dataset Extraction

1) Yelp Datasets:

- The 5 JSON files (i.e. business, user, tip, checkin, and review) were downloaded from the Yelp dataset challenge website and placed on the local machine. These JSON files were then stored in MongoDB using python's pymongo package which ran on a Jupyter notebook. They were then stored as 5 collections in MongoDB under their respective names.

2) Zomato Datasets:

- The top 12 cities from US and Canada with the most restaurants from the Yelp's business dataset were identified, and a limit of 100 restaurants were programmatically extracted for each city from Zomato's API. These 1200 records were then stored as the 'restaurant' collection in MongoDB. The cities are Las Vegas, Toronto, Phoenix, Charlotte, Scottsdale, Calgary, Pittsburg, Montreal, Mesa, Tempe, Chandler, Cleveland.

C. Data Preprocessing

The data was extracted from the 6 MongoDB collections using python's pymongo package and the below operations were performed:

1) Discarding Attributes:

- Attributes that would not contribute to the data analysis performed later were removed at this stage. For example, in the Yelp dataset, the information related to photos, thumbnails, and URLs of the

2) Normalization:

- The 6 collections read from MongoDB were normalized. Challenges faced normalizing nested JSON objects were resolved by only extracting the necessary attributes and discarding the rest. This made for an easy transition from nested JSON objects to the normalized relational tables in PostgreSQL which is mentioned below
- Stopwords and punctuations were removed using NLTK for text analytics on the Yelp review dataset, the text column of the review table was used

D. Data Preprocessing

1) Postgres Table Creation :

- Tables with appropriate structures and datatypes were created in PostgreSQL using the python package psycopg2. These tables contained columns only of the information which was required for the analysis.

2) Data Insertion to PostgreSQL :

- The data was extracted from the 6 collections from MongoDB using python. The data was then inserted into mongo Single, double quotes, and unicode characters had to be escaped before inserting them into PostgreSQL. The insertion was done using SQL statements executed by python's psycopg2 package

E. Data Analysis

- Jupyter Notebooks were used to analyze the ratings between Zomato and Yelp
- Google Colab was used to perform text analytics on the review dataset

IV. RESULTS

The objective was to compare ratings of same restaurants on Zomato and Yelp. Text matching has been used on restaurant names, and the latitudes and longitudes of the matched restaurants were compared for similarity. This is because the same restaurant can have multiple branches with different ratings but with the same name. The main purpose was to compare the exact same restaurant between both the sites. From the below chart it can be observed that for most of the instances, Zomato has a higher rating for the same restaurant as compared to Yelp. However, it must be noted that Yelp rates its restaurants in increments of 0.5, while Zomato rates in increments of 0.1.

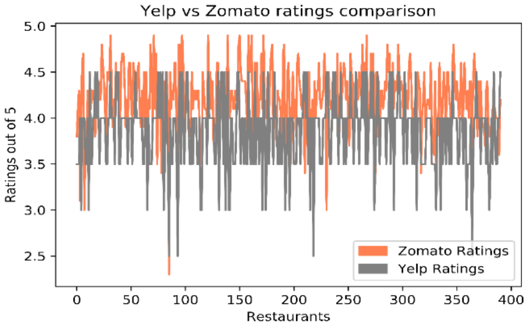


Fig. 2. Yelp vs Zomato ratings

Zomato ratings were converted to intervals of 0.5 to achieve one standard rating scale across both sites. The ratings on Zomato were converted as follows:

TABLE I. Rating Standardisation

Zomato Original Ratings	Zomato Standardized ratings
4.75+	5
4.25 - 4.74	4.5
3.75 - 4.24	4
3.25 - 3.74	3.5
2.75 - 3.24	3
2.25 - 2.74	2.5
1.75 - 2.24	2

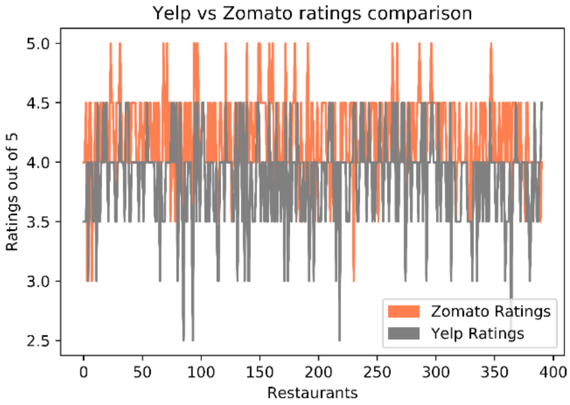


Fig. 3. Yelp vs Zomato ratings

For a better understanding, scatterplot of the same data was plotted and as observed earlier, Zomato mostly has higher ratings for every restaurant as compared to Yelp. Also, the average rating lines were plotted in order to understand the actual differences between ratings on both the platforms. The dashed lines show the average ratings for individual platforms. Zomato has an average rating of approximately 4.2 whereas Yelp has an average rating of around 3.8. This means that, on average, restaurants are likely to have a rating of more than 0.5 on Zomato.

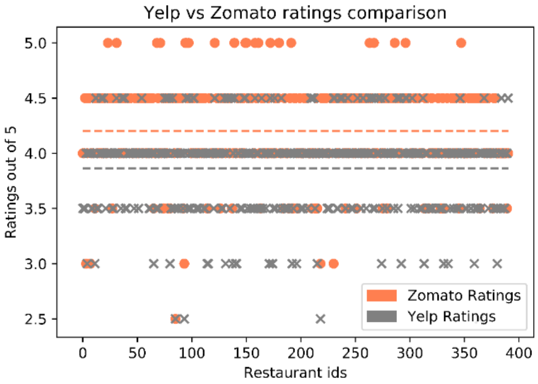


Fig. 4. Ratings by Restaurant Id

To further investigate the difference in ratings, data has been plotted to see how the number of votes between these restaurants between both the sites. From the below graph it can be clearly observed that the number of users voting for a restaurant on Yelp is much higher when compared to Zomato.

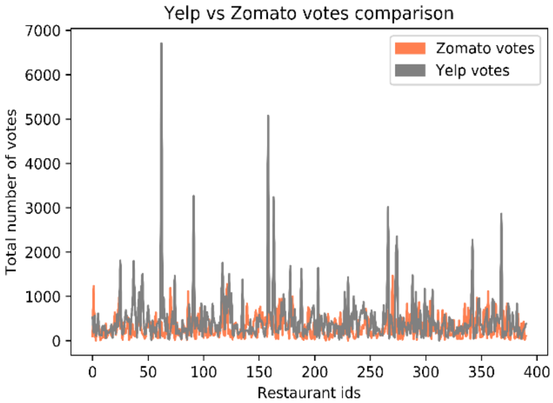


Fig. 5. Number of votes by Restaurant Id

Hence, from the above four visualization plots, we can see the difference between ratings of same restaurant on Zomato and Yelp. In most of the cases, the Zomato ratings were observed to be higher when compared to Yelp. A mere difference of just 0.5 on Zomato can change the restaurant's rating text from 'Good' to 'Excellent'.

Extent of Rating Difference between Yelp and Zomato:

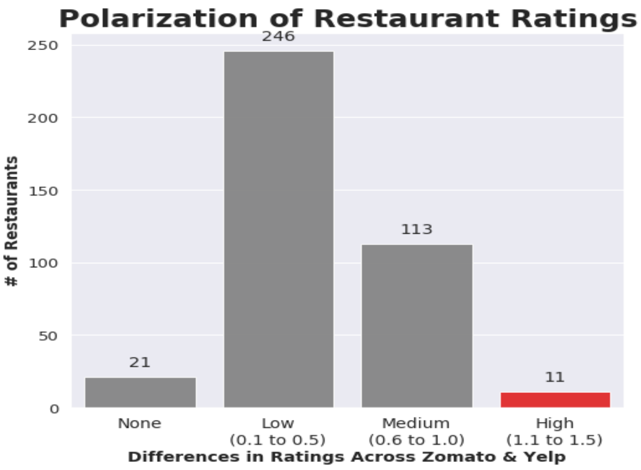
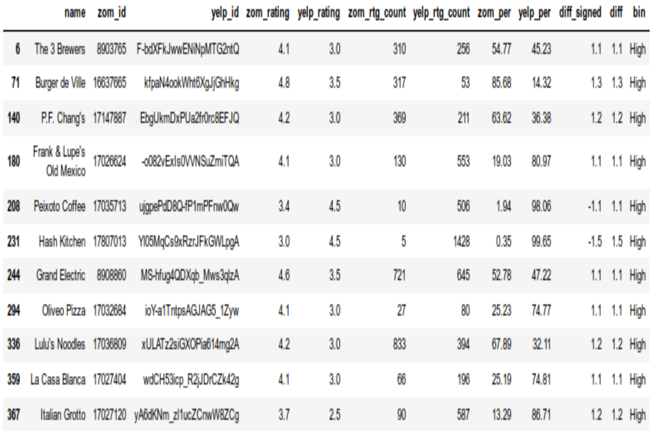


Fig. 6. Polarisation of Restaurannt Ratings

From the above plot we can see that 11 out of the total 391 restaurants have a high rating difference of 1.1 to 1.5 stars. The fact that there are more 'Low' polarized restaurants caters to the assumption that, in general, the ratings are the same across different review sites like Yelp and Zomato.



	name	zom_id	yelp_id	zom_rating	yelp_rating	zom_rtg_count	yelp_rtg_count	zom_per	yelp_per	diff_signed	diff	bin
6	The 3 Brewers	8903765	F-bdXFkJwwENiNpMTG2ntQ	4.1	3.0	310	256	54.77	45.23	1.1	1.1	High
71	Burger de Ville	16637665	kfpaN4ookWht6XgJjGhHkg	4.8	3.5	317	53	85.68	14.32	1.3	1.3	High
140	P.F. Chang's	17147887	EbgUkmDxPUa2fr0rc8EFJQ	4.2	3.0	369	211	63.62	36.38	1.2	1.2	High
180	Frank & Lupe's Old Mexico	17026624	-o082vExIs0VVNSuZmiTQA	4.1	3.0	130	553	19.03	80.97	1.1	1.1	High
208	Peixoto Coffee	17035713	ujgpePdD8Q-fP1mPFnw0Qw	3.4	4.5	10	506	1.94	98.06	-1.1	1.1	High
231	Hash Kitchen	17807013	YI05MqCs9xRzrJFkGWLpgA	3.0	4.5	5	1428	0.35	99.65	-1.5	1.5	High
244	Grand Electric	8908860	MS-hfug4QDXqb_Mws3qlzA	4.6	3.5	721	645	52.78	47.22	1.1	1.1	High
294	Oliveo Pizza	17032684	ioY-a1TntpsAGJAG5_1Zyw	4.1	3.0	27	80	25.23	74.77	1.1	1.1	High
336	Lulu's Noodles	17036809	xULATz2siGXOPia614mg2A	4.2	3.0	833	394	67.89	32.11	1.2	1.2	High
359	La Casa Blanca	17027404	wdCH53icp_R2jJDrCZk42g	4.1	3.0	66	196	25.19	74.81	1.1	1.1	High
367	Italian Grotto	17027120	yA6dKNm_zl1ucZCnwW8ZCg	3.7	2.5	90	587	13.29	86.71	1.2	1.2	High

Fig. 7. 11 High Polarised Restaurants

When we investigate further (as shown in the above table), it becomes clear that, for 2 of the 11 restaurants (i.e. Peixoto Coffee and Hash Kitchen), the rating count of Zomato accounts for less than 2 percent of the total rating counts from both review sites. Since there is an inadequate amount of Zomato ratings, we cannot rely on the rating difference of the restaurants. It is also a similar situation with Burger de Ville where less than 15 percent of the total rating counts are from Yelp.

Cuisines in each Category of Rating Difference:
Can the cuisine of a restaurant explain the extent of difference in rating between the sites?

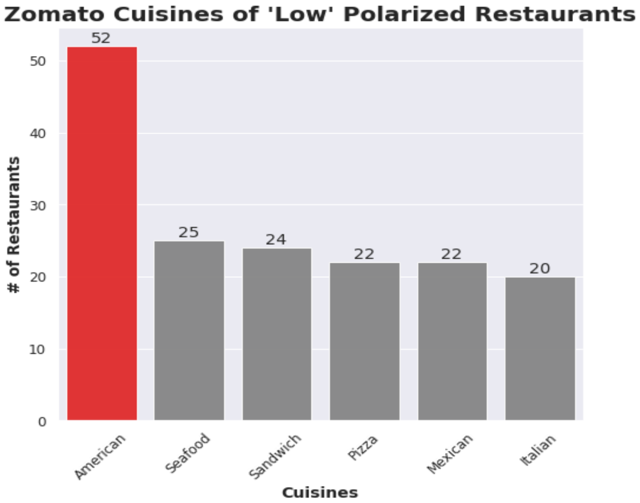


Fig. 8. Zomato cuisines of 'Low' Polarised Restaurants.

American cuisine restaurants are the most among the restaurants that have a low rating difference (i.e. between 0.1 to 0.5) between Zomato and Yelp. There are 52 American cuisine restaurants which is more than twice the next category (i.e. Seafood). This would lead us to believe that American cuisine restaurants usually tend to have a low rating difference between both the sites.

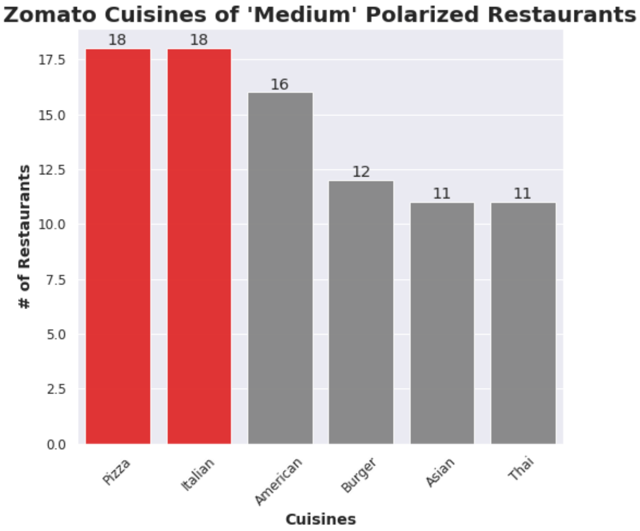


Fig. 9. Number of votes by Restaurant Id

From the above plot, we can see that Italian cuisine restaurants and Pizzerias are the most among the restaurants that have a medium rating difference (i.e. between 0.6 to 1.0) between Zomato and Yelp.

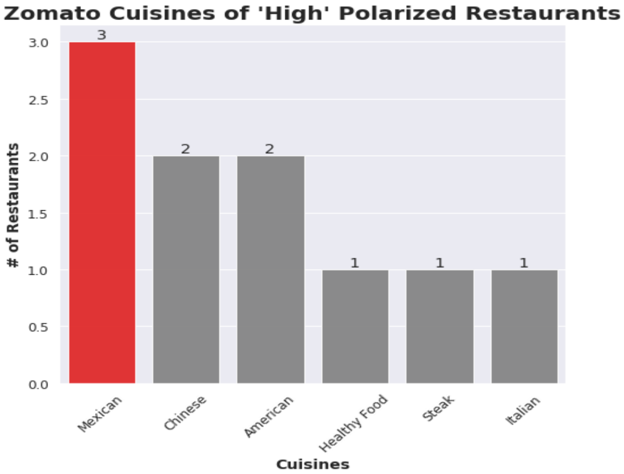


Fig. 10. Zomato cuisines of 'High' Polarised Restaurants

Mexican cuisine restaurants are the most among the restaurants that have a high rating difference (i.e. between 1.1 to 1.5) between Zomato and Yelp. Given that Mexican cuisine restaurants do not have relatively high numbers in the 'Low' and 'Medium' polarized rating plots, we can assume that Mexican restaurants tend to have a high rating difference across both sites.

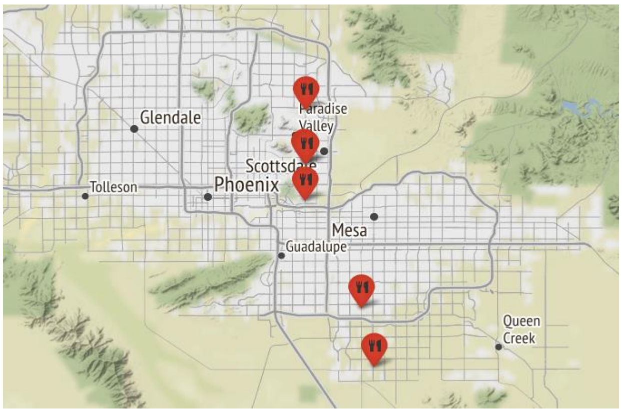


Fig. 11. Geospatial

From the above plot using python's package Folium, it is seen that 6 out of the 11 restaurants that fall under the 'High' polarized rating category are situated in and around the cities nearing Phoenix, Arizona.

Word Cloud for Top reviewed business

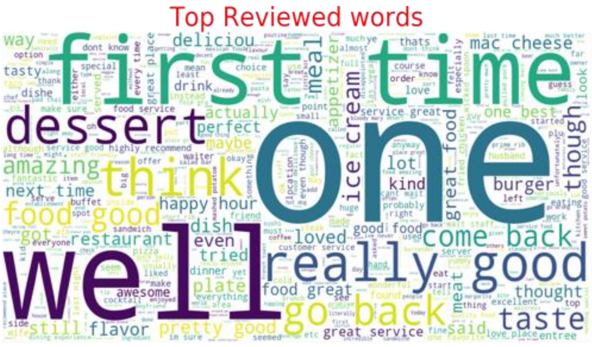


Fig. 12. Top reviewed words

For the list of 391 matched restaurants, we have extracted the 190407 Yelp reviews associated with them. Wordcloud is being used to visualize the set of top reviewed words after removing the stopwords (a, an, the) and punctuation marks to get processed set of words.

As we can observe in the word cloud below, the size of each word represents the maximum number of occurrences of the word in the entire dataset. It is observed that word "One", "well", "first, "time", "dessert" are the top 5 words used by reviewers.

Sentiment Analysis

Using Textblob library in Python we have performed the sentiment analysis on textual data. Polarity score between -1 to 1 is assigned to each review. We have categorized the reviews into three sentiments based on the score bands as follows:

TABLE II. Sentiment categorization by polarity score

Polarised score	Sentiment Category
>= -1 to <0	Negative
=0	Neutral
>0 to <=1	Positive

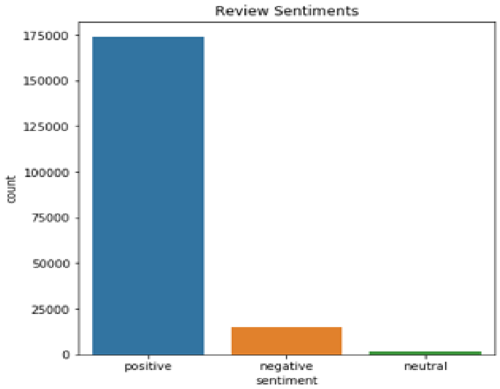


Fig. 13. Review Sentiments

As observed, the sentiment is mostly positive which is evident from the below graph the Yelp restaurants.

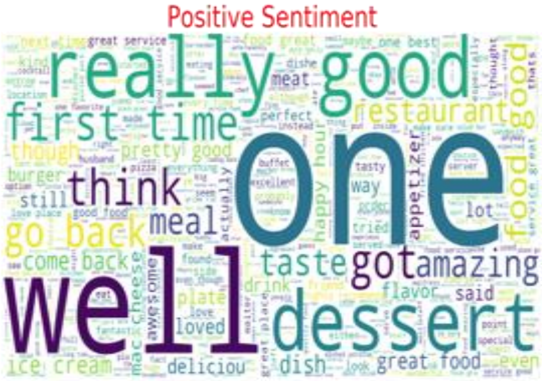


Fig. 14. Positive Sentiment word cloud

From the positive sentiments Word Cloud. Most frequent words used by reviewers is "one"- implying that those review had phrases like "one of the best restaurants", "well", "really", "good", "dessert"-implying most users liked it, "amazing".

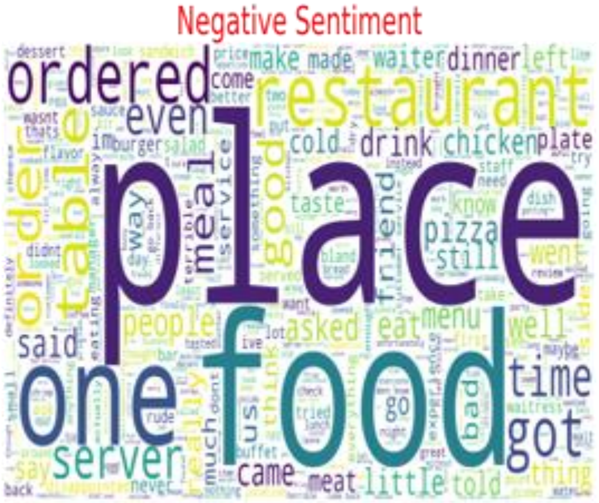


Fig. 15. Negative Sentiment word cloud

From the negative sentiments Word Cloud. Most frequent words used by reviewers is "place"-implying bad outlook, "food", "one", "time".

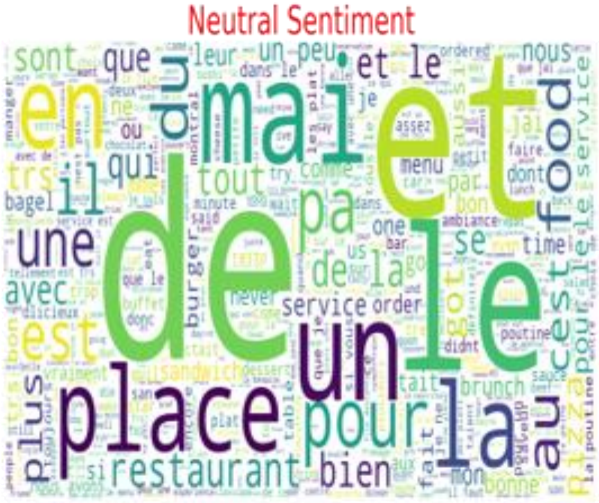


Fig. 16. Neutral Sentiment word cloud

From the neutral sentiments Word. Most frequent words used by reviewers are using Spanish "de", "et" , "le" as pre-processing was performed only on English reviews.

V. CHALLENGES

A. *Time Consuming PostgreSQL Write Speeds:*

- The insertion to PostgreSQL tables were time consuming. For example, the review dataset of 6aset of 6 GB took more than 2 hours to insert into its PostgreSQL table

Solution:

- By adding the 'UNLOGGED' keyword in the CREATE TABLE statement for the review table, the insertion time was reduced to just 20 minutes. This instructs PostgreSQL to prevent logging the operations performed on the review table. While the table's data will be lost in the event of a database crash, this is the tradeoff we make for faster writes to PostgreSQL

B. *Fetching Data from Zomato's API:*

- There was a limitation for only retrieving 20 restaurants in each API request

Solution:

- Zomato API only returns 20 records per query. Looping, firing queries from different API keys all returns same set of 20 records every time. Thus, a start variable was introduced in order to control which set of 20 restaurants are fetched.

C. *Identifying Unique Restaurants between Yelp and Zomato using Coordinates:*

- There were many duplicates found when using only the restaurant name to find similar restaurants from Yelp and Zomato.

Solution:

- The restaurants having the same name, and coordinate values were queried. For this, the latitude and longitude coordinates were compared for equality only up to the 3rd decimal place. The 3rd decimal position of a latitude and longitude can accurately identify one residential street from another. Referring this blog [6], 2 and 4 decimal places would have been too generic and too precise respectively. A total of 391 restaurants were found to be on Yelp and Zomato

D. *Handling Large Review Dataset:*

- When using the review dataset, due to low memory and computational power on the local machine, it was challenging to load the entire data of 6 GB into python for analysis

Solution:

- Google Colab was used to load in the review dataset and perform text analytics using the NLTK, Wordcloud, and Textblob packages in Python. This reduced the computational time significantly

VI. CONCLUSION

It is evident that restaurants with certain cuisines like Mexican are more likely to have polarizing ratings between the Yelp and Zomato. However, restaurants with cuisines like American are less likely to have a large difference in ratings between the sites. Most of the restaurants with 'High' polarized ratings are situated in and around the neighboring cities of Phoenix, Arizona.

VII. FUTURE WORK

A thorough analysis of the user data can be analyzed to see if there are any trends in the reviewers of the restaurants with polarizing ratings. Another situation would be to investigate the reviews of such restaurants and note when (i.e. which year or month) such polarization take place. This could tell us the time when the ratings substantially dropped or increased on either site making the rating difference noticeable.

VIII. References

[1] T. Economist, "The world's most valuable resource is no longer oil, but data," 6 May 2017. [Online]. Available: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

[2] V. L. K. J. S. a. K. P. S. M. Prithivirajan, "Analysis of star ratings in consumer reviews: A case study of Yelp," in *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, 2015.

[3] M.Luca, "Reviews, Reputation, and Revenue: The Case of Yelp.com," *SSRN Electronic journal* , 2011.

[4] S. S. a. S. S. S. Hegde, "Restaurant setup business analysis using yelp dataset," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, 2017.

[5] U. P.-S. G. S. P. Fayyad, "From data mining to knowledge discovery in databases," *AI Magazine 17(3),* pp. 37-53, 1996.

[6] "Precision Matters: The Critical Importance of Decimal Places," [Online]. Available: https://blis.com/precision-matters-critical-importance-decimal-places-five-lowest-go/.

[7] A. G. M. a. V. K. M. P. Cervellini, "Finding Trendsetters on Yelp Dataset," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Athens, 2016.

[8] T. D. a. J. Kalita, "Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning," in *Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA),* Anaheim, 2016.

[9] Yelp, "Yelp Open Dataset," [Online]. Available: https://www.yelp.com/dataset. [Accessed December 2019].

[10] Zomato, "Zomato API for Developers," [Online]. Available: https://developers.zomato.com/api#headline2. [Accessed December 2019].

[11] N. C. o. Ireland, "NORMA eResearch @NCI Library," [Online]. Available: http://trap.ncirl.ie/. [Accessed December 2019].

Appendix:

Group members:

Donovan Jeremiah	Student Number: x18181562- 33.33%
Shrey Shah	Student Number: x18192271- 33.33%
Mubeen Ali Mohammed	Student Number; x18180370- 33.33%

In this project every member of the group worked on various datasets from Zomato and Yelp. Shrey worked on the Restaurant dataset of Zomato and the Business dataset of Yelp. Mubeen worked on the Reviews datasets from both Yelp and Zomato. Donovan worked on combined datasets involving Restaurant, Business, and Tip.

Each member of the group fetched the required data programmatically from Zomato API and Yelp in the form of JSON files, stored in MongoDB, processed, cleaned and transformed their datasets. We then performed analysis and visualisations on our data from PostgreSQL.

The group collaborated and worked on the report together.