IDS 561 – Analytics for Big Data

Project Report

Yelp Review Recommendation System through Sentiment Analysis

2nd May 2023

Group 16

Hoang Nha Nguyen

Nithin Gowda

Prajwal Chidri

Purva Ladge

Contents

Abstract

In this big data analytics project, we developed a recommendation system for Yelp using sentiment analysis techniques to obtain insights into user reviews. Our tasks include collecting and preprocessing the Yelp dataset, performing exploratory data analysis to identify significant features, and lastly applying various machine learning models such as logistic regression, decision trees, and neural networks for sentiment classification. We attained high accuracy in classifying the positive and negative sentiment. We also conducted competitor analysis and keyword analysis to identify trends and patterns in the data. The proposed recommendation system can help businesses improve their customer experience and increase traffic to the Yelp app. Overall, this project demonstrates the power of big data analytics in providing valuable insights for decision-making in the restaurant industry.

1. Introduction

Yelp is a popular online platform where users can share their opinions about local businesses such as restaurants, cafes, and shops. In addition to its review and rating features, Yelp also offers various services such as online reservations, social events, and business data insights. This makes it a valuable source of information for both users and businesses.

With millions of reviews and ratings available on the platform, it can be challenging for users to find the most relevant and trustworthy information. This makes it difficult for users to make informed decisions about where to spend their time and money.

To address this problem, we propose the development of a Yelp review recommendation system that uses sentiment analysis to provide users with personalized recommendations based on their preferences. Our system will analyze the sentiment of each review and rating and use machine learning algorithms to identify patterns and trends in the data.

By leveraging this information, our system will provide users with recommendations that align with their preferences and interests. This will enable users to make more informed decisions about where to spend their time and money and help businesses to receive more accurate feedback on their services.

Overall, our project aims to provide a more personalized and reliable user experience on Yelp, ultimately improving the platform's usefulness for both users and businesses.

2. Background Study

We also conducted literature review to gain a deeper understanding of the topic, to learn of past and on-going research in this area. There are several related projects online and articles in peer-reviewed journals.

The study by Almoudi and Alghamdi analyzed online reviews, specifically restaurant reviews on Yelp, using sentiment analysis with binary and ternary classifications. The study proposed an unsupervised approach based on semantic similarity, achieving a maximum accuracy of 98.30% using ALBERT model and an accuracy of 83.04% for the proposed aspect extraction method. [2]

Ching and Bulos's study used the AYLIEN Text Analysis API and performed time series forecasting using linear regression to understand the customers' concerns of five restaurants and recommended business strategies based on the results. [3]

Research by Luo et al. aimed to offers a refined approach to analyze wealthy review content, providing scholars and practitioners with a better understanding of the relationship between a reviewer's evaluation of distinct attributes and overall satisfaction. [4]

A study by Yu et al., introduced a machine learning-based method for characterizing different aspects of restaurants based on Yelp reviews. Results indicated that customers tend to express more sentiment regarding service, and different cuisines are associated with certain characteristics such as fresh ingredients for Japanese cuisine and pizzas for Italian restaurants. [5]

Liu's study uses over 350,000 Yelp reviews on 5,000 restaurants to perform an ablation study on text preprocessing techniques and compare the effectiveness of several machine learning and deep learning models on predicting user sentiment. Results show that simpler models such as Logistic Regression and Support Vector Machine are more effective at predicting sentiments than more complex models such as Gradient Boosting, LSTM, and BERT. [6]

3. Experimental Setups

3.1 Data Collection

For our project, we sourced our dataset from Kaggle, a popular platform for finding and sharing data sets. Kaggle is known for hosting a wide range of high-quality data sets, and we selected

one that we felt was well-suited to our needs. In this project, we used the Yelp Open Kaggle Dataset [1], a subset of Yelp's businesses, reviews, and user data. It was originally put together for the “Yelp Dataset Challenge”. These datasets include information about businesses across 8 metropolitan areas in the USA and Canada.

This project utilities three JSON files:

- **business.json** : business data including location data, attributes, and categories.
- **review.json** : full review text data including the user_id that wrote the review and the business_id the review is written for user.json.
- **user.json** : user's friend mapping and all the metadata associated with the user.

3.2 Dataset Description

- Counts of Reviews : 6990280
- Counts of Businesses : 150346
- Counts of Users : 1987897
- 8 metropolitan areas
- Target variable : “stars”

With regression models, our objective is to help businesses predict Yelp star ratings based on past reviews given by the users.

As we aim to build the grading recommendations system, the location of the restaurants is an important factor to consider.

3.3 Exploration Data Analysis (EDA)

We employed PySpark, NLTK libraries for EDA and Natural Language Processing. Preparing and structuring the dataset is crucial for the performance of any machine learning or data mining technique. Hence, data preprocessing and EDA are necessary steps in these approaches.

3.3.1 Data Types

We examined the data types of the three datasets including JSON files of “business”, “Review”, and “Users”. Fig 1 displays the features in the dataset and their datatype.

```
[('average_stars', 'double'),
 ('compliment_cool', 'bigint'),
 ('compliment_cute', 'bigint'),
 ('compliment_funny', 'bigint'),
 ('compliment_hot', 'bigint'),
 ('compliment_list', 'bigint'),
 ('compliment_more', 'bigint'),
 ('compliment_note', 'bigint'),
 ('compliment_photos', 'bigint'),
 ('compliment_plain', 'bigint'),
 ('compliment_profile', 'bigint')
 ('compliment_writer', 'bigint'),
 ('cool', 'bigint'),
 ('elite', 'string'),
 ('fans', 'bigint'),
 ('friends', 'string'),
 ('funny', 'bigint'),
 ('name', 'string'),
 ('review_count', 'bigint'),
 ('useful', 'bigint'),
 ('user_id', 'string'),
 ('yelping_since', 'string')]
```

Fig 1. Features and their datatypes

3.3.2 Distribution and variable analysis

We are interested in businesses whose reviews are rated high. Fig 2 outlines the businesses with its different features:

business_id	name	categories	count
sk2lZI4zmuGAccd3D...	Boyd Hill Nature ...	Active Life, Park...	33
83IeQHroXEctmMpK1...	The Richel D'Ambr...	Hair Salons, Day ...	39
6e85By5Jy7MMnW2cE...	Wanderwell	Tax Services, Boo...	5
WKMJwqnfZKsAae75R...	Roast Coffeehouse...	Coffee & Tea, Foo...	31
jIBjZcqVOU4pOVT-s...	IRB Sushi	Sushi Bars, Resta...	34
NQhyMw8SOU1HB-V9X...	Champion Chevrole...	Automotive, Auto ...	20
q6661I3CGW0UB740E...	India House	Home & Garden, Ho...	32
lpbt16sSm4BTcfeq4...	Super Wok	Restaurants, Chinese	38
AwmeLVLEfdFoCa0La...	The Beer Store	Food, Beer, Wine ...	16
RZ-FNTXvqHKngyLGD...	Gaetano's of West...	Italian, Food, Re...	15

Fig 2. Details of the different businesses

Fig 3. captures the total number of businesses in different categories that are not limited to restaurants, food, shopping, home services, nightlife, beauty & spas, bars, etc.

category	count
Restaurants	36978
Food	20998
Shopping	18915
Restaurants	15290
Home Services	10563
Nightlife	9990
Beauty & Spas	9907
Bars	9130
Health & Medical	8832
Local Services	8556

only showing top 10 rows

Fig 3. Total number of businesses

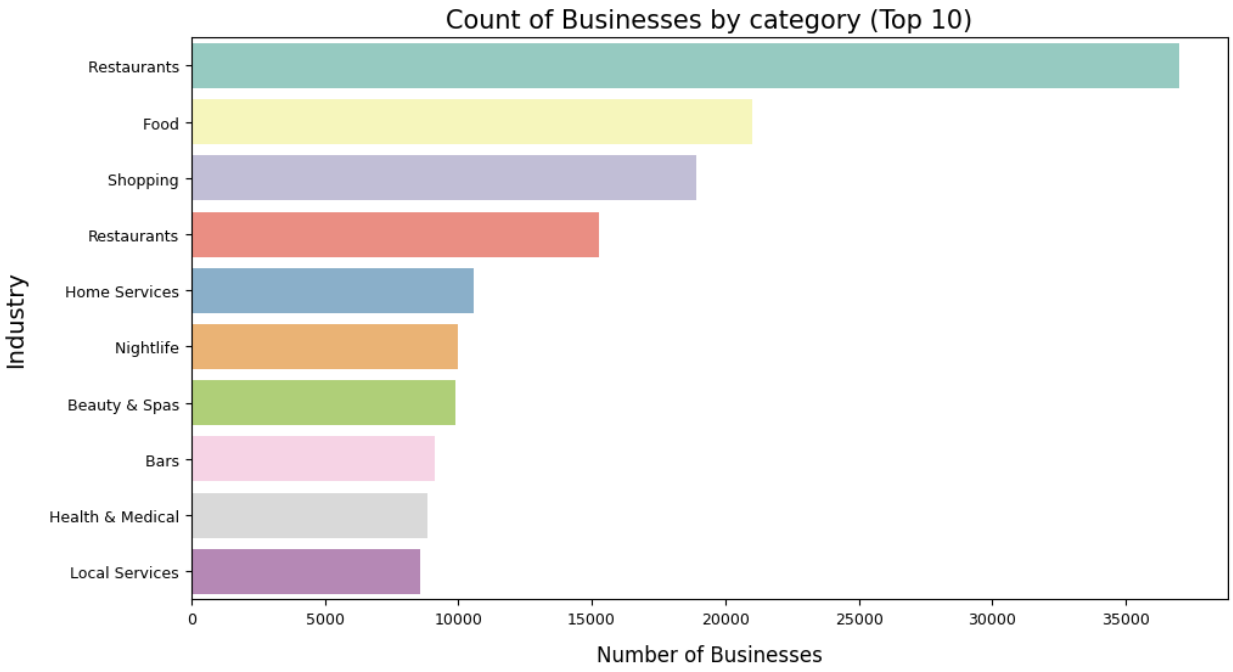


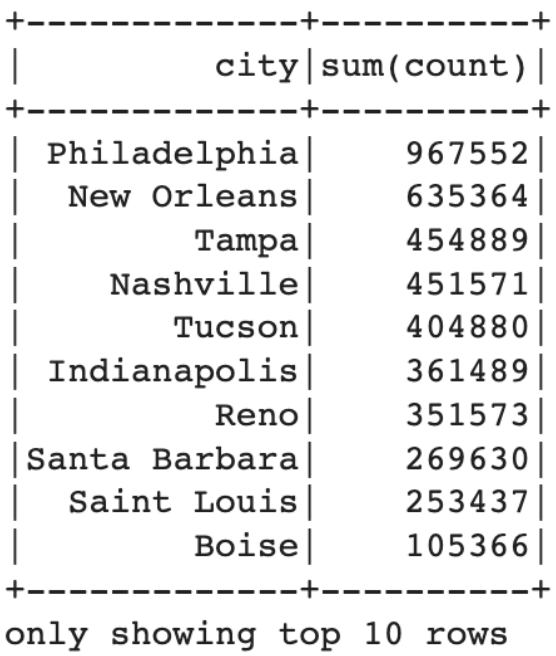
Fig 4. Visualization of count of businesses

Count of Businesses for different star ratings, however, we are only curious about extracting insights into more than 3-star-rated businesses. Fig 5 illustrates this.

stars	count
4.0	31125
4.5	27181
3.5	26519
3.0	18453
5.0	16307
2.5	14316
2.0	9527
1.5	4932
1.0	1986

Fig 5. Count of business ordered by star rating.

Total number of businesses launched in different cities is shown below in Fig 6.

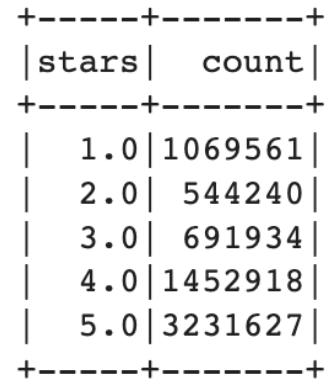


city	sum(count)
Philadelphia	967552
New Orleans	635364
Tampa	454889
Nashville	451571
Tucson	404880
Indianapolis	361489
Reno	351573
Santa Barbara	269630
Saint Louis	253437
Boise	105366

only showing top 10 rows

Fig 6. Count of business in different cities

Total number of businesses for a range of star ratings 1-5 is illustrated in Fig 7. We can see that 4-star reviews have the highest number of reviews and 2-star ratings has the lowest count.



stars	count
1.0	1069561
2.0	544240
3.0	691934
4.0	1452918
5.0	3231627

Fig 7. Total businesses

3.3.3 Univariate Analysis

Total number of open businesses is illustrated below in Fig 8.

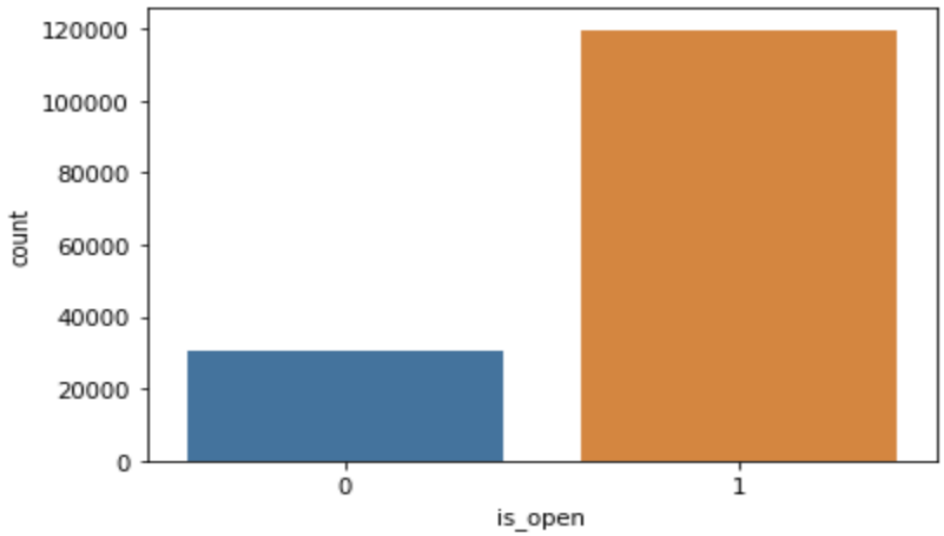


Fig 8. Total open businesses

Visualization of total count of businesses in different cities is shown in Fig 9. We can observe that Philadelphia has the highest number of businesses and Boise has the lowest number of businesses.

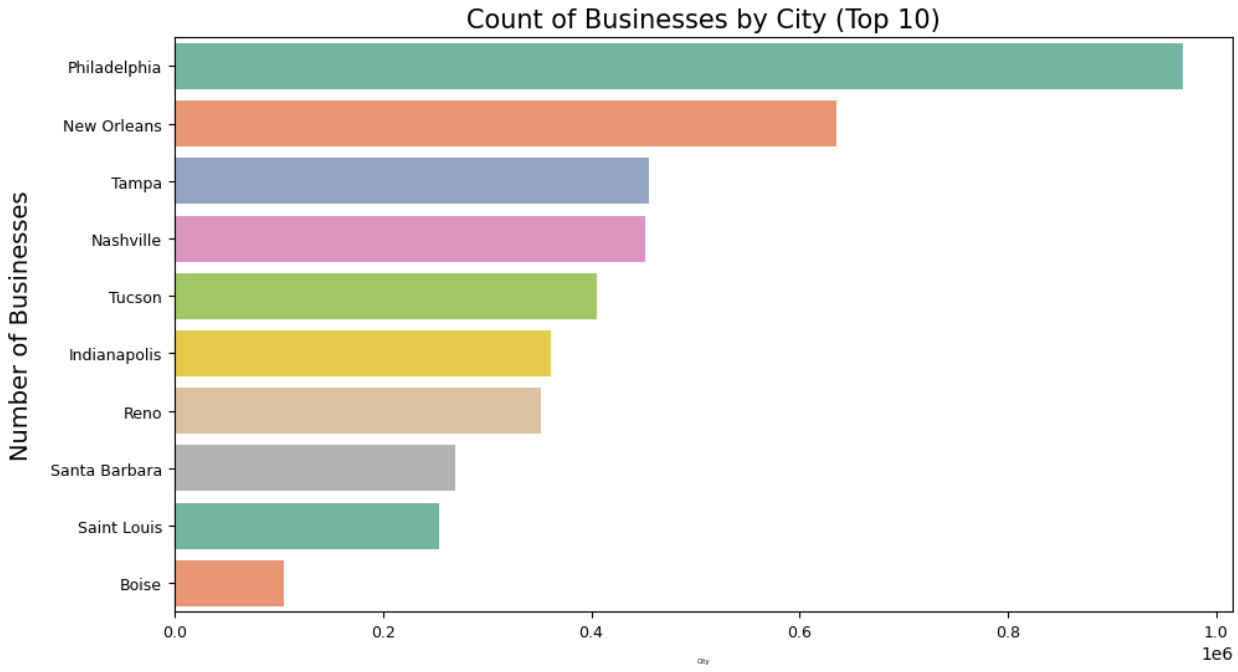


Fig 9. Businesses by City

Fig 10 illustrates the Total reviews received over time. We learn that the number of reviews reaches its peak around 2019 and then there is a decline. This could possibly be explained due to saturation, or the pandemic caused due to COVID-19.

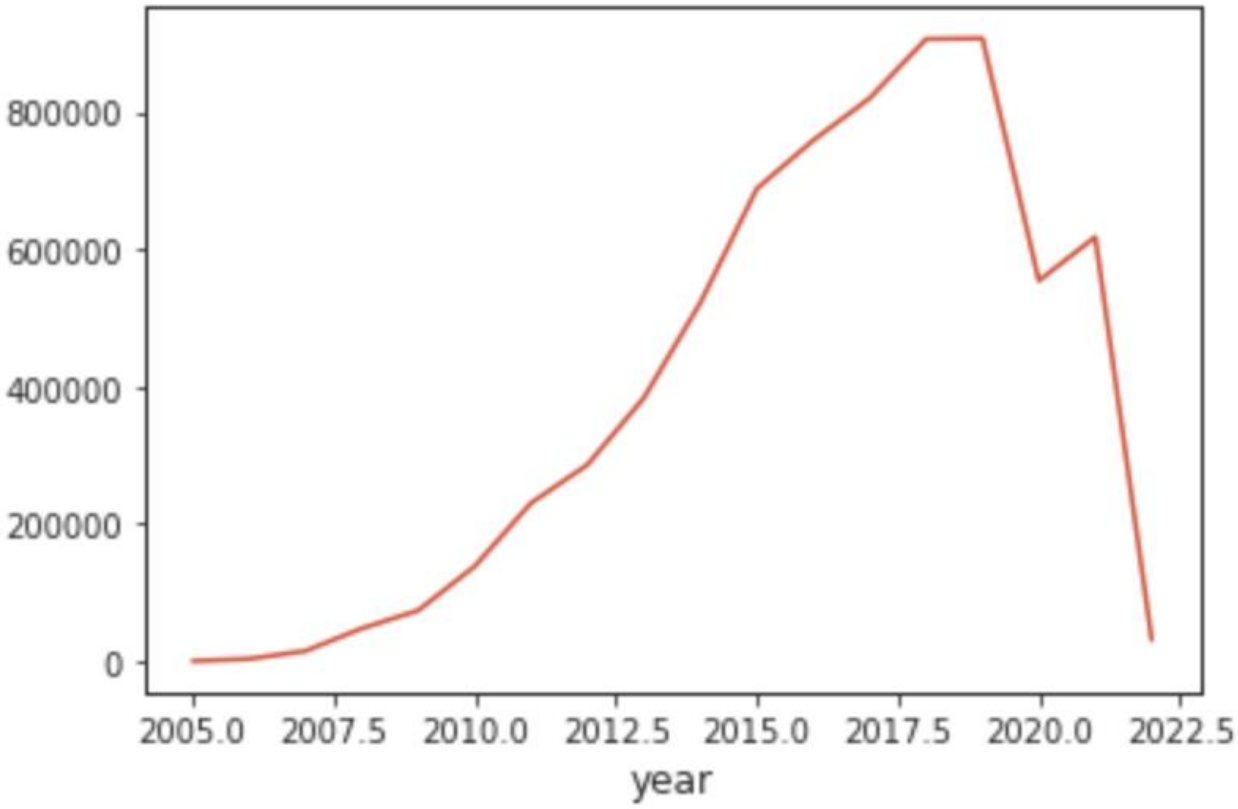


Fig 10. Total reviews over Time

3.3.4 Bivariate Analysis

Count of reviews and length of the reviews is illustrated in Fig 11. There are higher number of lower length of reviews and fewer lengthy reviews.

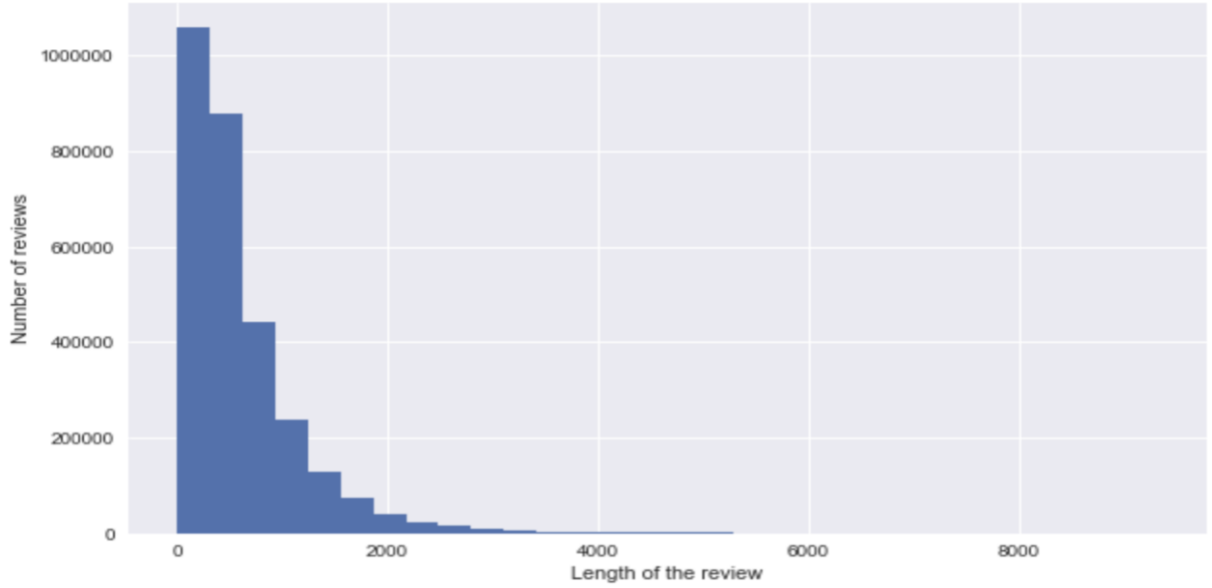


Fig 11. Reviews count vs reviews length

There are higher number of positive reviews and fewer number of negative reviews as displayed below in Fig 12.

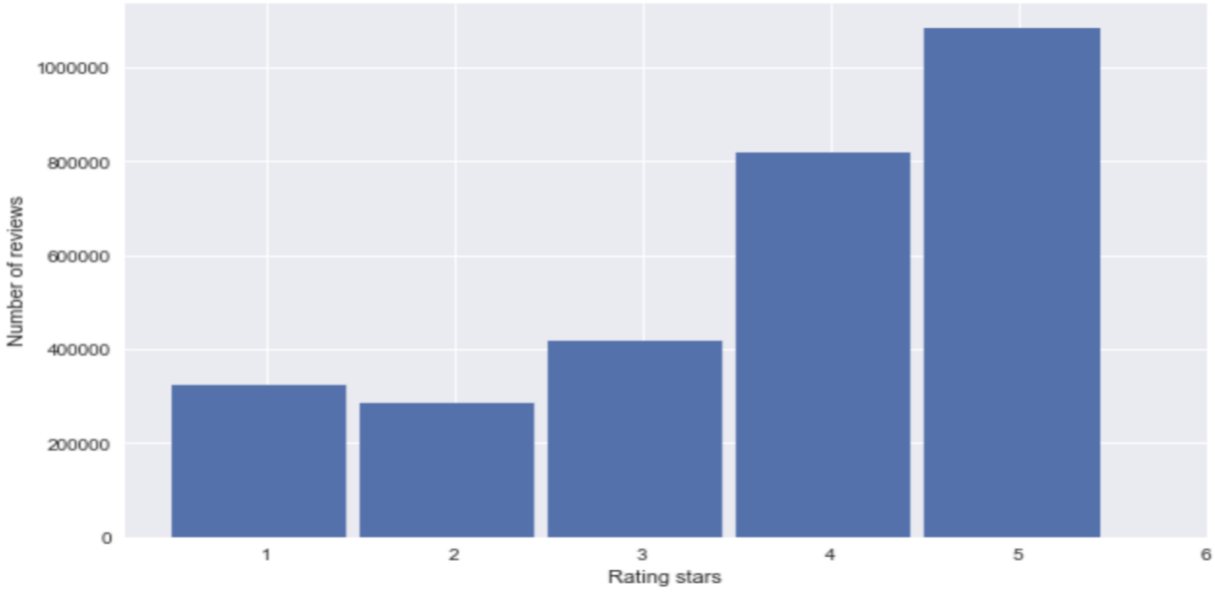


Fig 12. Reviews count vs star rating

3.4 Data Pre-processing

- We dropped the rows with null values.
- Filtered the businesses with reviews less than a particular threshold (we are interested in gaining insights into businesses with reviews greater than 3.0-star ratings)
- Cleaned the reviews text data by removing stop words, punctuations, and white spaces.
- Used label encoding for “stars” columns (categorizing the numerical star ratings)

  *Consider normalization if necessary

3.5 Text Pre-processing

Text preprocessing is the process of transforming raw text data into a format that is suitable for analysis, machine learning, or other natural language processing tasks. It involves several steps such as tokenization, stemming, lemmatization, stop-word removal, and punctuation removal.

Tokenization refers to the process of breaking down the text into individual words or tokens. Stemming involves reducing words to their root form by removing suffixes, while lemmatization does the same but ensures that the resulting word is still a valid word in the language. Stop-word removal involves removing common words such as "the", "a", "an" which are not meaningful in the analysis, and punctuation removal involves removing symbols such as commas, periods, and quotation marks.

Text preprocessing is necessary for several reasons: It reduces the size of the text corpus, making it easier and faster to process. It removes noise from the text and makes it more consistent and standardized. It helps in improving the accuracy of text analysis by reducing the complexity and variability of the text. It enhances the performance of machine learning algorithms by improving their ability to extract meaningful information from the text.

Step 1 Text data after removal of punctuations & label defined using the threshold value = 3

```
+--------------------+--------------------+-----+
|           review_id|                text|label|
+--------------------+--------------------+-----+
|KU_O5udG6zpxOg-Vc...|If you decide to ...|    1|
|BiTunyQ73aT9WBnpR...|I ve taken a lot ...|    1|
|saUsX_uimxRlCVr67...|Family diner  Had...|    1|
|AqPFMleE6RsU23_au...|Wow   Yummy  diff...|    1|
|Sx8TMOWLNuJBWer-0...|Cute interior and...|    1|
+--------------------+--------------------+-----+
only showing top 5 rows
```

Step 2 Text data after tokenization and removal of stop words

```
+--------------------+--------------------+-----+--------------------+--------------------+
|           review_id|                text|label|               words|           words_new|
+--------------------+--------------------+-----+--------------------+--------------------+
|KU_O5udG6zpxOg-Vc...|If you decide to ...|    1|[if, you, decide,...|[decide, eat, , a...|
|BiTunyQ73aT9WBnpR...|I ve taken a lot ...|    1|[i, ve, taken, a,...|[ve, taken, lot, ...|
|saUsX_uimxRlCVr67...|Family diner  Had...|    1|[family, diner, ,...|[family, diner, ,...|
|AqPFMleE6RsU23_au...|Wow   Yummy  diff...|    1|[wow, , , yummy, ...|[wow, , , yummy, ...|
|Sx8TMOWLNuJBWer-0...|Cute interior and...|    1|[cute, interior, ...|[cute, interior, ...|
+--------------------+--------------------+-----+--------------------+--------------------+
only showing top 5 rows
```

Step 3 Text data after feature extraction:

1. Count Vectorization

2. Term Frequency and Inverse Document Frequency (tf-idf)

```
+--------------------+--------------------+-----+--------------------+--------------------+--------------------+--------------------+
|           review_id|                text|label|               words|           words_new|                  tf|              tf_idf|
+--------------------+--------------------+-----+--------------------+--------------------+--------------------+--------------------+
|KU_O5udG6zpxOg-Vc...|If you decide to ...|    1|[if, you, decide,...|[decide, eat, , a...|(202063,[0,1,2,6,...|(202063,[0,1,2,6,...|
|BiTunyQ73aT9WBnpR...|I ve taken a lot ...|    1|[i, ve, taken, a,...|[ve, taken, lot, ...|(202063,[0,7,14,1...|(202063,[0,7,14,1...|
|saUsX_uimxRlCVr67...|Family diner  Had...|    1|[family, diner, ,...|[family, diner, ,...|(202063,[0,2,3,13...|(202063,[0,2,3,13...|
|AqPFMleE6RsU23_au...|Wow   Yummy  diff...|    1|[wow, , , yummy, ...|[wow, , , yummy, ...|(202063,[0,11,28,...|(202063,[0,11,28,...|
|Sx8TMOWLNuJBWer-0...|Cute interior and...|    1|[cute, interior, ...|[cute, interior, ...|(202063,[0,2,4,6,...|(202063,[0,2,4,6,...|
+--------------------+--------------------+-----+--------------------+--------------------+--------------------+--------------------+
only showing top 5 rows
```

4. Modeling

4.1 Logistic Regression (Baseline Model)

We chose Logistic Regression as our baseline Model for the following reasons:

Interpretability: Logistic regression models are easy to interpret,

Baseline comparison: By using logistic regression as a baseline model, we can compare the performance of more complex models to this simple model. This can help us to determine whether more complex models provide significant improvements in prediction accuracy.

Overall, logistic regression is a useful baseline model because it is simple, interpretable, and computationally efficient, making it a good starting point for many data analysis tasks.

4.2 Random Forest

Random Forest can handle large datasets with high-dimensional features, such as text data, making it an appropriate choice for our project.

4.3 Naive Bayes

Naïve Bayes is a simple yet effective algorithm for classification tasks, particularly in the domain of natural language processing. Naive Bayes assumes that the features are conditionally independent given the class label, so it is used to train and allowing for scalable and real-time classification.

4.4 SVM

SVM is a powerful algorithm that can handle high-dimensional data and can effectively classify data into different classes based on their features.

In the context of sentiment analysis and opinion mining, SVM can be used to classify customer reviews as either positive or negative based on the features extracted from the text data.

Moreover, SVM can handle both binary and multi-class classification tasks, making it a flexible algorithm that can be used for a variety of classification tasks.

Furthermore, SVM can provide good generalization performance and can be easily adapted to handle imbalanced datasets, which is often the case with sentiment analysis. Therefore, SVM is

a good option for an alternate model in this project because it can provide reliable classification results.

5. Findings

Model Comparison is shown below in Table 1

	Naive Bayes	**Logistic Regression**	**Random Forest**	**SVM**
F1 Score	0.846	0.8816	0.766	0.9122
Parameters used	Smoothing = 1.0 Model Type = 'Multinomial'	Lambda = 0.02 Alpha = 0.3	Max Depth = 2	Regular Parameter = 0.3 Number of iterations = 50

Table 1

We can observe that SVM is the best model with the highest F1 Score.

6. Yelp Review Recommendation System - Alternating Least Square (ALS) Algorithm

We implemented a recommender system that could suggest a user his/her preferred business around his or her location. Since we found that Restaurants were the top businesses and Philadelphia was the most reviewed city, we chose both these categories to build the recommender system.

We built the recommender system using Alternating Least Square (ALS) algorithm.
The ALS algorithm is a popular collaborative filtering algorithm that is widely used for building recommender systems. It is particularly well-suited for handling large and sparse datasets, making it a good option for this project, which involves analyzing Yelp reviews and recommending restaurants in Philadelphia.

The ALS algorithm works by factorizing the user-item interaction matrix into two lower-dimensional matrices: a user matrix and an item matrix. These matrices are then iteratively optimized using alternating least squares until the predicted ratings match the actual ratings as closely as possible. This allows the algorithm to make personalized recommendations to users based on their past behavior and preferences, which is essential for building an effective recommender system.

Additionally, ALS is relatively computationally efficient and can scale to handle large datasets with millions of users and items. This makes it a practical and scalable solution for the task of recommending restaurants to users based on their location and preferences.

The following are the parameters used in the recommender system:

- maxIter = 5
- regParam = 0.09
- rank = 25

Our Model was evaluated on Root Mean Square Error (RMSE).

A lower RMSE of 1.44 indicates that the model is more accurately predicting the user's preferences and interests, which leads to more personalized and relevant recommendations. Support Vector Machine gave the highest accuracy of 92% in predicting the sentiment of a business.

The Recommender system gives the user = 952 to following restaurant recommendations in the city "Philadelphia".

7. Conclusion

In this project, we leveraged machine learning techniques to improve customer experience and satisfaction in the restaurant industry. We extracted insights from Yelp reviews and developed

a recommender system using the Alternating Least Square algorithm to suggest users their preferred restaurants in Philadelphia. The system achieved an RMSE score of 1.44 and provided the user with 952 restaurant recommendations. Additionally, we used Support Vector Machine to predict business sentiment with a high accuracy of 92%. This project demonstrates the potential of deploying machine learning models to personalize and improve customer experience in the restaurant industry.

8. Future Scope

There are several potential future directions for this project. One possible avenue is to incorporate more data sources and features to further improve the recommender system's accuracy and personalization. For example, demographic data or social media activity could be used to better understand the user's preferences and suggest more relevant recommendations.

The recommender system could be extended to other cities and categories beyond restaurants to provide a more comprehensive and diverse set of recommendations. Overall, there are many exciting opportunities to further enhance and develop this project to improve customer experience and satisfaction in the restaurant industry.

Acknowledgement

We would like to thank Professor Yuheng Hu for the valuable advice and lectures during the entire course of the project. We would like to thank TA Aida Sanatizadeh for the lab sessions during the semester which aided us in completing this project successfully.

Roles of each member

Throughout the duration of the group project, every member of the team made a fair and balanced contribution to the project. We made a concerted effort to hold regular meetings, in which we discussed and planned the next steps to be taken to ensure the project's success.

We placed a great emphasis on ensuring that every member of the team had an equal say in shaping the project, and that no single individual's ideas or contributions were given undue weight or preference over others.

The project report, which was a culmination of our efforts, was written collaboratively, with each team member contributing equally to its composition. Each member's unique perspective and expertise were incorporated, resulting in a report that was comprehensive, insightful, and reflective of our collective efforts.

References

[1] https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=Dataset_User_Agreement.pdf

[2] Eman Saeed Alamoudi, Norah Saleh Alghamdi. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*. 27 Jan 2021.

[3] Michelle Renee D. Ching, Remedios de Dios Bulos. Improving Restaurants' Business Performance Using Yelp Data Sets through Sentiment Analysis. *ICEEG '19: Proceedings of the 3rd International Conference on E-commerce, E-Business and E-Government*. June 2019

[4] Yi Luo b, Liang (Rebecca) Tang, Eojina Kim, Xi Wang. Finding the reviews on yelp that actually matter to me: Innovative approach of improving recommender systems. *International Journal of Hospitality Management.* 2020

[5] Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao. Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. Center for Urban Science & Progress, New York University

[6] Siqi Liu. Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models. University of Waterloo. 15 April 2020

Appendices

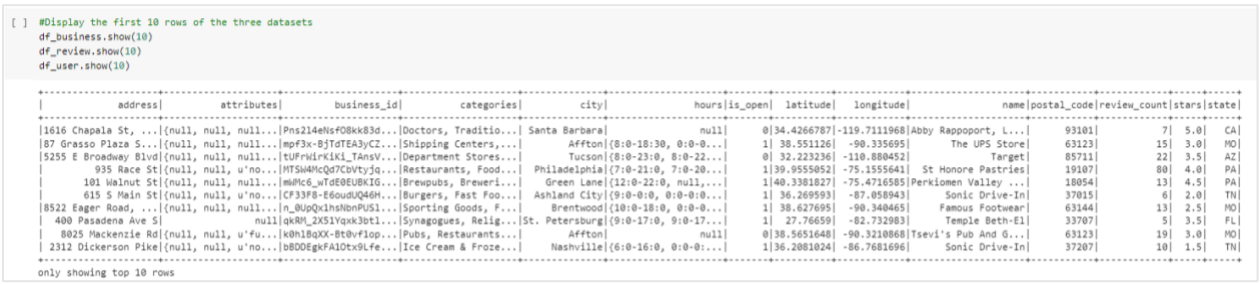
Appendix 1

SN	JSON File	Column Name	Description
1	business.json	address	Address of the business
2		attributes	Attributes used to describe the business
3		business ID	Unique ID of the business
4		categories	Type of Business
5		city	The city in which the business is located
6		hours	The businesses open hours
7		is_open	Indicates whether the business is open or not
8		latitude	Coordinates of the business location
9		longitude	Coordinates of the business location
10		name	Name of the business
11		postal_code	Postal Code of the business
12		review_count	Sum total count of all reviews for the business
13		stars	Rating of the business
14		state	The state in which the business is located
15	review.json	business id	Unique ID of the business
16		date	Date of the review
17		review_id	Unique ID number of the review
18		user_id	Unique ID number of the yelp user
19		Stars	1/2/3/4/5-star rating
20	user.json	average_stars	Average of all star ratings of all users
21		compliment_cool	Refers to how stylish, trendy, or fashionable a business is
22		compliment_cute	Refers to the charm, aesthetics, or attractiveness of a business
23		compliment_funny	Refers to the humor, wit, or entertainment value of a business
24		compliment_hot	Refers to the popularity, buzz, or demand for a business
25		compliment_list	Refers to a curated list of businesses that a user can create and share on Yelp
26		compliment_more	Refers to additional information or details about a business that a user can access on Yelp
27		compliment_note	Refers to a short comment or observation that a user can add to their review of a business
28		compliment_photos	Refers to the images that users can upload and share on Yelp to showcase a business
29		compliment_plain	Refers to the simplicity or lack of flair of a business
30		compliment_profile	Refers to the public page of a business on Yelp that contains information, reviews, photos, and other details
31		compliment_writer	Refers to the users who write reviews and provide feedback on Yelp
32		cool	Refers to the Yelp users who have a lot of followers and whose reviews are considered to be influential or stylish

33	elite	Refers to Yelp users who have been recognized by Yelp for their contributions to the community and who have been granted access to special features and events
34	Fans	Refers to the number of Yelp users who have bookmarked a business on Yelp as one of their favorites
35	Friends	Refers to the number of Yelp users who have added a particular Yelp user as a friend on the platform
36	Funny	Refers to the Yelp users who are known for their entertaining and witty reviews
37	name	Refers to the username or display name of a Yelp user
38	Review_count	Total number of reviews that a business has received
39	Useful	The number of Yelp users who have voted a review as helpful or valuable
40	User_id	Unique ID number of the yelp user
41	Yelping_since	The date and time when the user created an account

Appendix 1. Attribute Name and Description

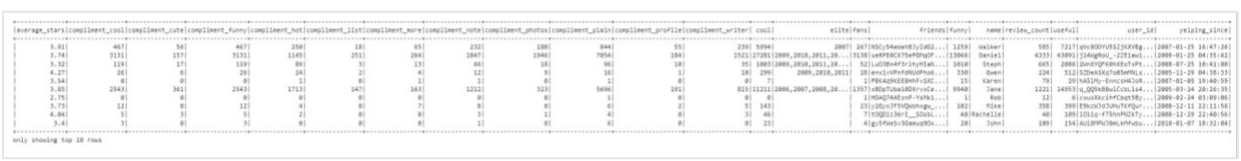
Appendix 2



```
[ ] #Display the first 10 rows of the three datasets
    df_business.show(10)
    df_review.show(10)
    df_user.show(10)
```

```
+--------------------+--------------------+--------------------+--------------------+---------------+-----+--------+----------+-----------+--------------------+-----------+------------+-----+-----+
|             address|          attributes|         business_id|          categories|           city|hours|is_open|  latitude|  longitude|                name|postal_code|review_count|stars|state|
+--------------------+--------------------+--------------------+--------------------+---------------+-----+-------+----------+-----------+--------------------+-----------+------------+-----+-----+
|1616 Chapala St, ...|{null, null, null...|Pns2l4eNsfO8kk83d...|Doctors, Traditio...| Santa Barbara|  null|      0|34.4266787|-119.7111968|Abby Rappoport, L...|      93101|           7|  5.0|   CA|
|87 Grasso Plaza S...|{null, null, null...|mpf3x-BjTdTEA3yCZ...|    Shipping Centers,...|        Affton|{8:0-18:30, 0:0-0...|      1| 38.551126| -90.335695|      The UPS Store|      63123|          15|  3.0|   MO|
|5255 E Broadway Blvd|{null, null, null...|tUFrWirKiKi_TAnsV...|Department Stores,...|         Tucson|{8:0-23:0, 8:0-22...|      0| 32.223236| -110.880452|              Target|      85711|          22|  3.5|   AZ|
|       935 Race St|{null, null, u'no...|MTSW4McQd7CbVtyjq...|Restaurants, Food...|  Philadelphia|{7:0-21:0, 7:0-20...|      1|39.9555052| -75.1555641|  St Honore Pastries|      19107|          80|  4.0|   PA|
|      101 Walnut St|{null, null, null...|mWMc6_wTdE0EUBKIG...|Brewpubs, Breweri...|    Green Lane|{12:0-22:0, null,...|      1|40.3381827| -75.4716585|Perkiomen Valley ...|      18054|          13|  4.5|   PA|
|       615 S Main St|{null, null, u'no...|CF33F8-E6oudUQ46H...|Burgers, Fast Foo...|  Ashland City|{9:0-0:0, 0:0-0:0...|      1| 36.269593|  -87.058943|      Sonic Drive-In|      37015|           6|  2.0|   TN|
|8522 Eager Road, ...|{null, null, null...|n_0UpQx1hsNbnPUSl...|Sporting Goods, F...|     Brentwood|{10:0-18:0, 0:0-0...|      1| 38.627695|  -90.340465|     Famous Footwear|      63144|          13|  2.5|   MO|
|  400 Pasadena Ave S|                null|qkRM_2X51Yqxk3btl...|Synagogues, Relig...|St. Petersburg|{9:0-17:0, 9:0-17...|      1|  27.76659|  -82.732983|      Temple Beth-El|      33707|           5|  3.5|   FL|
|   8025 Mackenzie Rd|{null, null, u'fu...|k0hlBqXX-Bt0vf1op...|        Pubs, Restaurants...|        Affton|                null|      0|38.5651648| -90.3210868|Tsevi's Pub And G...|      63123|          19|  3.0|   MO|
| 2312 Dickerson Pike|{null, null, u'no...|bBDDEgkFA1Otx9Lfe...|Ice Cream & Froze...|      Nashville|{6:0-16:0, 0:0-0:...|      1|36.2081024| -86.7681696|      Sonic Drive-In|      37207|          10|  1.5|   TN|
+--------------------+--------------------+--------------------+--------------------+---------------+-----+-------+----------+-----------+--------------------+-----------+------------+-----+-----+
only showing top 10 rows
```

```
+--------------------+----+-------------------+-----+--------------------+-----+--------------------+------+--------------------+
|         business_id|cool|               date|funny|           review_id|stars|                text|useful|             user_id|
+--------------------+----+-------------------+-----+--------------------+-----+--------------------+------+--------------------+
|XQfwVwDr-v0ZS3_Cb...|   0|2018-07-07 22:09:11|    0|KU_O5udG6zpxOg-Vc...|  3.0|If you decide to ...|     0|mh_-eMZ6K5RLWhZyI...|
|7ATYjTIgM3jUlt4UM...|   1|2012-01-03 15:28:18|    0|BiTunyQ73aT9WBnpR...|  5.0|I've taken a lot ...|     1|OyoGAe7OKpv6SyGZT...|
|YjUWPpI6HXG5301wP...|   0|2014-02-05 20:30:30|    0|saUsX_uimxRlCVr67...|  3.0|Family diner. Had...|     0|8g_iMtfSiwikVnbP2...|
|kxX2SOes4o-D3ZQBk...|   1|2015-01-04 00:01:03|    0|AqPFMleE6RsU23_au...|  5.0|Wow!  Yummy, diff...|     1|_7bHUi9Uuf5__HHc_...|
|e4Vwtrqf-wpJfwesg...|   1|2017-01-14 20:54:15|    0|Sx8TMOWLNuJBWer-0...|  4.0|Cute interior and...|     1|bcjbaE6dDog4jkNY9...|
|04UD14gamNjLY0IDY...|   1|2015-09-23 23:10:31|    2|JrIx1S1TzJ-iCu79u...|  1.0|I am a long term ...|     1|eUta8W_HdHMXPzLBB...|
|gmjsEdUsKpj9Xxu6p...|   0|2015-01-03 23:21:18|    2|6AxgBCNX_PNTOxmbR...|  5.0|Loved this tour! ...|     0|r3zeYsv1XFBRA4dJp...|
|LHSTtnW3YHCeUkRDG...|   0|2015-08-07 02:29:16|    0|_ZeMknuYdlQcUqng_...|  5.0|Amazingly amazing...|     2|yfFzsLmaWF2d4Sr0U...|
|B5XSoSG3SfvQGtKEG...|   0|2016-03-30 22:46:33|    1|ZKvDG2sBvHVdF5oBN...|  3.0|This easter inste...|     1|wSTuiTk-sKNdcFypr...|
|gebiRewfieSdtt17P...|   0|2016-07-25 07:31:06|    0|pUycOfUwM8vqX7KjR...|  3.0|Had a party of 6 ...|     0|59MxRhNVhU9MYndMk...|
+--------------------+----+-------------------+-----+--------------------+-----+--------------------+------+--------------------+
only showing top 10 rows
```



Appendix 2. Illustration of all first 10 rows of all columns from three JSON files

-END OF REPORT-