**TORONTO RESTAURANT RECOMMENDATION SYSTEM**

Group Members:

|  |  |  |
| --- | --- | --- |
| First name | Last Name | Student number |
| Ashvin | Alex | C0880169 |
| Fadernel | Bedoya | C0872455 |
| Bastian | Castillo | C0872274 |
| Marcelo | Munoz | C0873813 |
| Darshan | Ruparel | C0872322 |

Professor: Vahid Hadavi

April 16, 2023

TABLE OF CONTENTS

[INTRODUCTION 5](#_Toc132596206)

[DATA COLLECTION 6](#_Toc132596207)

[DATA UNDERSTANDING 6](#_Toc132596208)

[TEXT WRANGLING 7](#_Toc132596209)

[EDA + VISUALIZATION 8](#_Toc132596210)

[GeoMap Of Restaurants 8](#_Toc132596211)

[Rating Counts 9](#_Toc132596212)

[Top 10 Restaurants By Review Count 9](#_Toc132596213)

[Distribution Of User Activity By Reviews Over Years 10](#_Toc132596214)

[Top 5 Neighborhoods As per Restaurant Count 10](#_Toc132596215)

[Word Cloud Of Top 5 Neighborhood’s Cuisine 11](#_Toc132596216)

[Average Star Rating Of Reviews By Month and Year 11](#_Toc132596217)

[Location Of Users Visiting Toronto Restaurants 12](#_Toc132596218)

[NAMED ENTITY RECOGNITION 13](#_Toc132596219)

[TF-IDF 14](#_Toc132596220)

[SENTIMENT PREDICTION 14](#_Toc132596221)

[SVM 14](#_Toc132596222)

[RandomForestClassifier 15](#_Toc132596223)

[Comparison Of Both 15](#_Toc132596224)

[VADER SENTIMENT ANALYSIS 15](#_Toc132596225)

[SENITMENT POLARITY PREDICTION USING PRE-TRAINED MODEL: RoBERTa 17](#_Toc132596226)

[RESTAURANT RECOMMENDATION 18](#_Toc132596227)

[PROJECT FLOW 20](#_Toc132596228)

[CONCLUSION 20](#_Toc132596229)

[REFERENCES 21](#_Toc132596230)

**TABLE OF FIGURES**

[Figure 1:Collaborative Filtering vs Content Based Filtering 5](#_Toc132597223)

[Figure 2: Web Scrapping Of Toronto Restaurants 6](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597224)

[Figure 3: Map Of Restaurants in Toronto 8](#_Toc132597225)

[Figure 4: Percentage Of Star Ratings 9](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597226)

[Figure 5: Top 10 Restaurants in Toronto 9](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597227)

[Figure 6: User Activity Over Years 10](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597228)

[Figure 7: Top 5 Neighborhoods 10](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597229)

[Figure 8: Word Cloud Of Cuisines Of Restaurants With Top 5 Neighborhoods 11](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597230)

[Figure 9: Average Star Rating Of Reviews By Month and Year 11](#_Toc132597231)

[Figure 10: Location Of Users outside Toronto 12](#_Toc132597232)

[Figure 11: NER On Positive Reviews 13](#_Toc132597233)

[Figure 12: NER on Negative Reviews 13](#_Toc132597234)

[Figure 13: Vader Compound Score of Yelp Review 16](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597235)

[Figure 14: Vader Scores vs Customer Ratings 16](#_Toc132597236)

[Figure 15: RoBERTa Sentiment Scores 17](#_Toc132597237)

[Figure 16: Cosine Similarity 18](file:////Users/darshanruparel/Desktop/AIML/Term%202/NLP/Final%20Project/Toronto%20Restaurant%20Recommendation%20System%20Report.docx#_Toc132597238)

[Figure 17: Restaurant Recommendation Based On NER, TF-IDF and Cosine Similarity 19](#_Toc132597239)

[Figure 18: Project Flow 20](#_Toc132597240)

# **INTRODUCTION**

Data is increasing day by day and it is leading to a new world with the advancement in Artificial Intelligence and Machine Learning. One of the use cases of reliable data is to create efficient systems using recommendation. Recommendation systems are based on filtering wherein users are recommended different items based on their user history or providing items which are more relevant to their search. These systems are widely used to recommend items to buy, movies, menus, restaurants, places to visit etc.

There are 2 types of recommendation systems:

1. Content Based filtering
2. Collaborative filtering

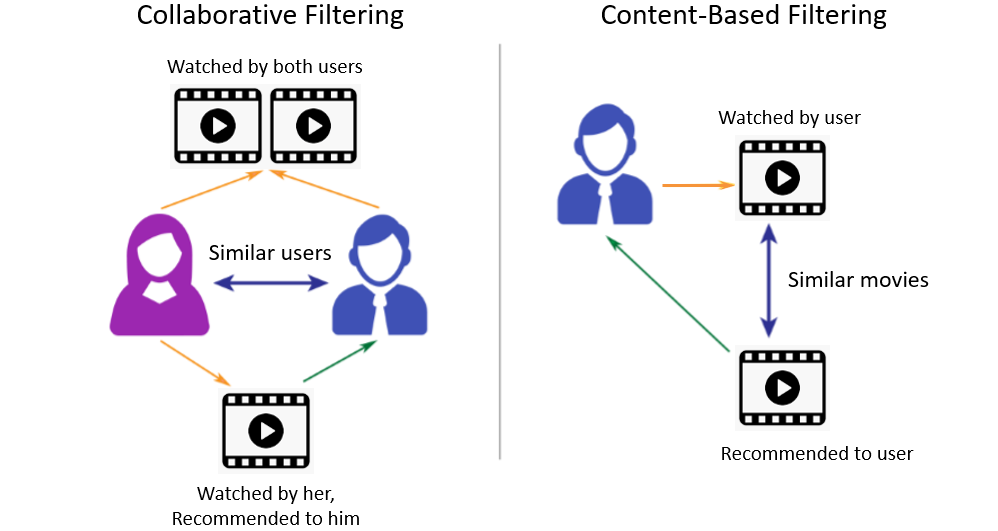


Figure 1:Collaborative Filtering vs Content Based Filtering

This report aims to inform the reader about the restaurant recommendation system and how content-based filtering can be used to recommend the restaurant to the user based on his interest.

# **DATA COLLECTION**

Yelp is a website which publishes crowd-source reviews about business. It contains data of restaurants registered and its customer reviews. We have scrapped the data using Beautiful Soup and Requests library in python. The data scrapped belongs to restaurants in Toronto. There are around 240 restaurants in Toronto with different cuisines and approximately 35,000 customer reviews.

Figure 2: Web Scrapping Of Toronto Restaurants

After collection of restaurant data, latitude and longitude coordinates were obtained using MapQuest API based on restaurant address.

# **DATA UNDERSTANDING**

There are 2 csv files containing restaurant data. One file is created using 240 different files having customer reviews of restaurants. This file contains reviews of all customers from all restaurants, user location, date, and all review ratings. Other file contains data of 240 restaurants having name, neighborhood, review count, rating, URL, menu, address, cuisine, latitude, and longitude.

# **TEXT WRANGLING**

Text wrangling is one of the important steps as it will pre-process the data and get it ready to be converted into vectors. These vectors can then be used for sentiment prediction by machine learning, or it can be helpful for restaurant recommendation based on vectors similarity. Below are the steps followed to make that the data is properly extracted from reviews.

1. Removing the new line character from all reviews and replacing with full stop.
2. Replacing the words like won’t with will not so that the stop words can be detected and removed.
3. Using Regular expression to find only the characters.
4. Tokenizing the words using NLTK.
5. Removing Stop words.
6. Calculating the Part of speech tag for each token.
7. Lemmatizing the words.
8. Removing string “star rating” from rating column so that it contains numbers only.
9. Creating new column ‘label’ which will contain 1 if review rating >= 4 else 0. This will be the target label for the machine learning models.

## 

# **EDA + VISUALIZATION**

Exploratory data analysis will help us to understand the data and discover some trends which can be used for further tasks.

## **GeoMap Of Restaurants**

Below map is created using plotly express wherein the visualization is done using latitude and longitude of restaurants. Here, the dots highlight the location of restaurant. It is clear that more restaurants are located in Downtown.

Map

Description automatically generated

Figure 3: Map Of Restaurants in Toronto

## Chart, pie chart Description automatically generated**Rating Counts**

Figure 4: Percentage Of Star Ratings

It seems most of the customer are satisfied with the service and food of the restaurants as majority of them having given 4 or 5 rating which contributes 80%. This also should be considered while building machine learning models for predicting the sentiments as the data would be less for negative ratings.

## Chart, histogram Description automatically generated**Top 10 Restaurants By Review Count**

Based on the customer review counts these are the top 10 restaurants and all of them are doing well with Pai Northern Thai Kitchen at Duncan Street having the highest reviews.

Figure 5: Top 10 Restaurants in Toronto

## **Distribution Of User Activity By Reviews Over Years**

Chart, histogram

Description automatically generatedThe users on Yelp have increased significantly from 2008 to 2019. Due to Covid, people isolated themselves and restaurants were closed, and the user activity dropped down drastically in 2020-2021. Yelp and the business registered will have to engage people so that users start sharing their review which in turn will help the business to reach out to other people and increase the popularity hence increase in revenue.

Figure 6: User Activity Over Years

## Chart, bar chart Description automatically generated**Top 5 Neighborhoods As per Restaurant Count**

This bar chart illustrates the top 5 neighborhoods having maximum restaurants. Business who are looking to open new restaurants can plan it accordingly as these areas have high chances of maximum users.

Figure 7: Top 5 Neighborhoods

## **Word Cloud Of Top 5 Neighborhood’s Cuisine**

Text

Description automatically generatedThe word cloud on the left helps us to visualize the popular cuisine of restaurants near top 5 neighborhoods. Businesses can identify that Italian, Japanese, Middle Eastern, Asian Fusion, and Thai as some of the popular cuisines.

Figure 8: Word Cloud Of Cuisines Of Restaurants With Top 5 Neighborhoods

## Chart, line chart Description automatically generated**Average Star Rating Of Reviews By Month and Year**

Chart, line chart

Description automatically generated

Figure 9: Average Star Rating Of Reviews By Month and Year

The above 2 figures should the average star rating as per month and year over the time. Star ratings have decreased a lot as we approached. It seems that we can use the data as time series to predict the future ratings, but this is not part of the scope.

## **Location Of Users Visiting Toronto Restaurants**

Table

Description automatically generatedTable

Description automatically generated

Figure 10: Location Of Users outside Toronto

Above image tells us about the customers visiting Toronto restaurants from locations other than Toronto. If we consider customer from other areas in Toronto, then Markham has highest counts as it is closest. If we consider customer outside Canada, then San Francisco and New York has higher numbers. Businesses should make sure that if they have chains in other areas or countries except Toronto then they should provide best services over all so their business is not impacted as customers are coming from different locations.

# **NAMED ENTITY RECOGNITION**

NER is a technique used for information extraction. Here, we have used NER to extract based on two categories: dish names and non-dish words. On the left we have non dish words and on the right, we have words used to define dish.

Text

Description automatically generated Text

Description automatically generated

Figure 11: NER On Positive Reviews

Text

Description automatically generated Text

Description automatically generated

Figure 12: NER on Negative Reviews

From the above we can conclude below points:

1. Users are happy with the food and service provided by the restaurants, but it seems sometimes the quality of food and service is not consistent.
2. People are more interested in dishes containing cheese, pizza, burger, shrimp.
3. Izakaya is a Japanese word meaning stay-drink-place which is frequent in negative reviews so business should take some action as Japanese people seems to be unhappy.
4. People have mentioned about the waiters in negative reviews, so restaurants should make sure that waiter behave properly with the customers as this can impact the business.
5. Chef, on the other hand, have been praised as per positive reviews and that is a good take away for the restaurant owners.
6. NER will be useful in recommending restaurants to customers.

# **TF-IDF**

TF-IDF is an algorithm used for identifying the words that are most important in a document or corpus. Here, TF stands for term frequency and IDF is inverse document frequency used for finding the rarity of word in the entire document. Here, we have used TF-IDF to convert the preprocessed text into vectors.

# **SENTIMENT PREDICTION**

Once, we have created the vectors using TF-IDF, we can train machine learning models to predict the sentiments of reviews. Here, the label column contains 0 for negative reviews and 1 for positive reviews. We have split the data into train and test wherein the models will be trained on 80% of data.

## **SVM**

Support Vector Machine (SVM) is a supervised machine learning algorithm which best suits for classifying two groups. After training the SVM model, we tried to predict the remaining 20% percent and it is found that SVM achieved accuracy of 89.88%. So, SVM did good job on identifying the reviews as positive or negative.

## **RandomForestClassifier**

A random forest classifier is supervised machine learning algorithm which uses decision tress logic and averaging to achieve higher accuracy. After training the model with 80% of data, model predicted the sentiments of test data with 84.68% accuracy.

## **Comparison Of Both**

SVM did better than RandomForestClassifier and could be used to predict the sentiments based on reviews of customer.

|  |  |
| --- | --- |
| Model Used | Accuracy (%) |
| SVM | 89.88 |
| RandomForestClassifier | 84.68 |

# **VADER SENTIMENT ANALYSIS**

Valence Aware Dictionary and Sentiment Reasoner (VADER) is a rule-based tool to identify the sentiments of the customer reviews. It also provides the polarity score of how positive, negative, and neutral a review is.

VADER Sentiment Analyzer was trained to get the polarity scores of all the reviews. Vader generated 4 scores for each review (negative, neutral, positive, compound). Here, if the compound score >= 0.05 then it is considered positive, <= -0.05 is considered as negative and in between the two range it is considered as neutral. Below is the chart comparing the review ratings with the compound score generated by VADER.

Chart, box and whisker chart

Description automatically generatedFor star rating 1, VADER did good job by generating the compound score below 0. For ratings 2 ,3, 4 and 5, VADER predicted the compound score as neutral and positive. It seems that Vader did perform well but this requires a deeper exploration.

Figure 13: Vader Compound Score of Yelp Review

For exploring it further, below are the charts which helps us in identifying the scores of VADER correctly. As we can see in first graph VADER considered some of the reviews with rating 1 as positive, graph with neutral values looks fine but graph with negative values also has some 5-star ratings.

Chart, bar chart

Description automatically generated

Figure 14: Vader Scores vs Customer Ratings

Below are examples of review with customer rating and wrong VADER prediction:

|  |  |  |
| --- | --- | --- |
| Reviews | Customer Rating | VADER prediction |
| They always tell you "10-15 minutes" but you can be awaiting an hour so just make sure your party is prepared to wait. Good food | 1 | Positive |
| Fried mushroom & spinach dumplings were fire! General Tao's chicken was fire! Fried rice was fire! | 5 | Negative |

VADER generated wrong outputs for 1,314 reviews wherein 1,246 reviews with 1, 2-star rating were identified as positive and 68 reviews with 5-star rating were identified as negative. It calculates the sentiment of each word in review and takes the average of all the words in that review to identify the sentiment. This can be misleading, so we thought of using RoBERTa pretrained model.

# **SENITMENT POLARITY PREDICTION USING PRE-TRAINED MODEL: RoBERTa**

RoBERTa is optimized version of BERT in which the transformer is trained using some changes in hyperparameters. We implemented the model with the available version in Pytorch and used Kaggle services for generating the sentiment scores using the model pretrained on twitter sentiments.

Chart, bar chart

Description automatically generated

Figure 15: RoBERTa Sentiment Scores

As you can see that RoBERTa did well in identifying the sentiments of the users wherein positive chart contains high values with 4,5-star rating and negative chart contains high values with 1,2-star rating. These scores will be useful for recommending the restaurants.

# **RESTAURANT RECOMMENDATION**

To recommend the restaurant it is important to first extract all the important words from the reviews and store in data frame for all restaurants. Now based on the user input we will take the NER’s and convert into vectors using TF-IDF.

Graphical user interface

Description automatically generated with low confidenceNow using cosine similarity, we will compare the user input vector with all other restaurant vector. We will sort the data frame based on similarity and will take the first 20 restaurants based on similarity of NER.

Figure 16: Cosine Similarity

Now we will again find the cosines between the 20 rows based on review count, ratings, location, and sentiment scores. Then we will sort based on similarity and the top 5 rows would become the relevant restaurants.

Below is an example of restaurant recommendation:

Graphical user interface, text, application, email

Description automatically generated

Figure 17: Restaurant Recommendation Based On NER, TF-IDF and Cosine Similarity

# **PROJECT FLOW**

Diagram

Description automatically generated

Figure 18: Project Flow

All the project code was uploaded to the repository located at the following URL:

<https://github.com/PR0YECT0SWEB/Toronto_Restaurant/blob/main/Toronto%20Restaurant%20Recommendation%20System.ipynb>

# **CONCLUSION**

Content based recommendation can be used to recommend the top 5 restaurants. Using NER, TF-IDF and cosine similarity between features, it becomes easy to find the relevant restaurants. Vader did good job in identifying the sentiments, but RoBERTa was very accurate. SVM had more accuracy than Random Forest for classifying the sentiments as negative or positive. This report could help the businesses to find what customer are saying about restaurants in Toronto. It could also help new businesses to find correct place and which cuisine type to invest in. Users can use this project to get the top 5 relevant restaurants based on their interest.

# **REFERENCES**

Ng, J. (n.d.). Content-based Recommender Using Natural Language Processing (NLP). Kdnuggets.com. https://www.kdnuggets.com/2019/11/content-based-recommender-using-natural-language-processing-nlp.html

(n.d.). Toronto Restaurants Yelp. Yelp.ca. https://www.yelp.ca/toronto