RECOSYS (REstaurant reCOmmendation SYStem)

॥वसुधैव कुटुम्बकम्॥

THESIS SUBMITTED TO
Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M. Sc.
DEGREE

By

SOURAV BALASAHEB KHOT

(Batch 2022-24 / PRN 22070243027)

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University)

5th Floor, Atur Centre, Gokhale Cross Road, Model Colony,
Pune - 411016

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to the individuals who have contributed immensely to the successful completion of my project, "**RECOSYS (REstaurant reCOmmendation SYStem)**" Their unwavering support, guidance, and encouragement have been pivotal in shaping this study. First and foremost, I am indebted to my internal guide, **Dr. Rajesh Dhumal**, for his invaluable expertise, relentless guidance, and unwavering support throughout this research journey. His profound knowledge and insightful feedback have been instrumental in shaping the direction and outcomes of this study.

I would also like to extend my heartfelt appreciation to **Dr. Vidya Patkar**, the Deputy Director of Symbiosis Institute of Geoinformatics, for her invaluable guidance, encouragement, and constructive suggestions. Her mentorship has immensely contributed to the refinement of the research methodology and the overall quality of the paper.

I express my sincere thanks to **Prof. T.P. Singh**, the Director of Symbiosis Institute of Geoinformatics, for providing a conducive academic environment that fosters research and innovation. His leadership and support have been instrumental in creating a platform for pursuing this research endeavor. I am grateful to my colleagues for their collaboration and willingness to assist me whenever I faced doubts or challenges. Their valuable insights and discussions have enriched my understanding and improved the overall quality of this research.

Finally, I would like to express my deepest appreciation to my parents for their unwavering support, encouragement, and understanding throughout this journey. Their love, belief in my abilities, and sacrifices have been the cornerstone of my success. I extend my heartfelt gratitude to all these individuals for their contributions, without which this research paper would not have been possible.

- **Sourav Khot**

LIST OF ABBREVIATIONS

RECOSYS	Restaurant Recommendation System
NLP	Natural Language Processing
RNN	Recurrent Neural Networks
CNN	Convolutional Neural Networks
npm	Node Package Manager
nltk	Natural Language Toolkit
TfidfVectorizer	Term Frequency - Inverse Document Frequency Vectorizer
RFC	Random Forest Classifier
RPF	Random Partition Forest
NB	Naïve Bayes
TF-IDF	Term frequency-inverse document frequency
SVM	Support Vector Machine
SVC	Support Vector Classifier

GLOSSARY

RECOSYS – Restaurant Recommendation System This is one of the systems which gives customers, the restaurant’s recommendations according to other pre-users’ experience and their opinions for that here we used the sentimental analysis method.

Sentimental Analysis – Sentimental Analysis is one of the NLP techniques which finds the text in positive, negative, or neutral forms. This text is collected from different sources like authors or customers, according to their opinions. and on that basis sentimental analysis is performed thus sentimental analysis is also called opinion-based mining.

Subjective Analysis - Instead of depending simply on factual facts, subjective analysis entails evaluating information based on individual ideas, sentiments, and biases. It uses individual viewpoints and assessments to analyze and comprehend a topic, yet it may lack impartiality and be swayed by personal prejudices.

PREFACE

The RECOSYS (REstaurant reCOmmendation SYStem) project's main aim is to develop an efficient and user-friendly restaurant recommendation system for Pune City. The main objective was to scrape restaurant reviews from Google Maps using the Apify software. The scraped data included essential or worthwhile information such as review text, restaurant titles, location coordinates (latitude and longitude), and ratings.

The purpose of this study is to describe the development and implementation of a recommendation system developed specifically for the Symbiosis Institute of Geoinformatics. By providing instructors and students with personalized recommendations based on their preferences and interests, this technology aims to enhance the user experience.

The importance of utilizing technological advances to streamline and improve various processes has increased as the field of geoinformatics continues to expand significantly. Thanks to the usefulness of recommendation systems in providing customized content, users may locate relevant information and services more rapidly.

To check the recommendation system's accuracy, sentiment analysis techniques were applied to the extracted reviews. The sentiment analysis process involved calculating the polarity and subjectivity of the reviews. A formula was written in this project to calculate the sentimental analysis rating, primarily based on the polarity of the reviews. This rating served as a crucial factor in determining the recommendation scores for each restaurant.

The scope of this article covers the whole life cycle of the recommendation system, from its initial idea and design through its final deployment and evaluation. We also go through the many tools, methods, and algorithms employed in the creation of a dependable and effective system.

When designing the system, we took into account the varied needs and interests of the users, including students and professors. Because the system is designed to respond to user interactions and learn from them, suggestions are always being refined.

We encourage readers to read this study, which describes the creation, application, and assessment of the Symbiosis Institute of Geoinformatics Recommendation System. May this paper be a useful tool for comprehending the operation of the system and its prospective effects on the geoinformatics community at the institute.

1. INTRODUCTION

Food is one of the important factors in our life, everyone works hard to provide healthy food for their family and their loved ones. To meet their cravings, it's important to provide knowledge of different cuisines, the best food options, and their famous restaurants that aim to provide the best quality delicious food.

A normal user finds the best and most affordable restaurants based on location. Currently, google and such a like platform suggest restaurants based only on the stars that every restaurant receives. Even though star reviews are collected from users, they don't represent the actual quality of the food. Hence, I propose RECOSYS (REstaurant reCOmmendation SYStem using sentiment analysis), which provides the top 10 restaurants based on customer reviews and incorporates the taste of food, ambiance, hygiene, customer service, and guest experience. This not only helps the user find better restaurants based on cuisine but also allows gives assurance of the food quality, and hygiene.

Google Maps is the platform where all users upload their local business-related information with their address and all related information on that. Also in that shop, customers/ users visit and express their experience in the form of comments or some text on that site. This way, experiences or feelings about that restaurant or shop will be shared today. For that just need the internet and mobile its one method like mouth-publicity was done in past time. On this site, we get star-based ratings and some comments regarding any shop given by different customers who visited that shop. this star-based rating is actually calculated as all users' average given rating so sometimes we cannot understand how good a restaurant this is e.g. if two customers give ratings as 5 stars then we get their average rating as 5 stars for that shop. So I introduced RECOSYS which shows restaurants whose reviews are more than 100, that restaurants' suggestions are given only although these restaurant's rating is different than normal ones here rating is calculated by the sentimental analysis method.

The main aim of these projects is to provide the best restaurant recommendations to customers using the sentimental analysis method and also provide different star ratings than traditional ones these ratings are dependent on food price, quality, and ambiance like these factors. Also, we provide here one option: to choose cuisine so they will understand which cuisine is getting best in local restaurants. This system works in the right way or not for that, we applied different models like the Multinomial model, Random forest, and Extra Tress classifier model also for checking recommendations shown is right or not for that subjective analysis is performed.

2. OBJECTIVE

1) To develop a robust restaurant recommendation system for Pune City. And applying sentiment analysis techniques to the extracted reviews for calculating the polarity and subjectivity of the reviews.
2) Comparative analysis of sentimental analysis techniques.
3) Formulate the criteria for rating based on polarity.
4) Create a user interface and provide it to users for ease of use.

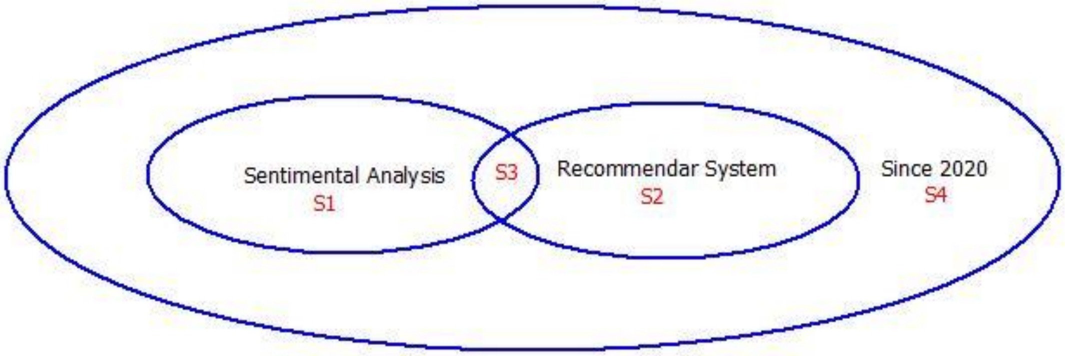
3. LITERATURE REVIEW METHODOLOGY

This section will include information on how relevant studies were found and evaluated for their content. This section will be organized as follows:

Concept maps are used to organize essential concepts and to categorize datasets.

Venn diagrams and the use of search tables to find relevant research

From the Venn diagram below we can understand which keywords are used to search research papers



3.0.1 Venn Diagram for keyword analysis

3.1 LITERATURE REVIEW:

An investigation was conducted on "Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews." This was completed in 2023 by the authors Goonjan Jain and Neha Punetha. Today, a sizable portion of web data originates from text sentiments. For that study, sentiment and emotion analysis of review data was conducted using an unsupervised mathematical optimization framework. The MCDM approach and a mathematical framework based on game theory were used in this experiment on the Yelp dataset and the Tripadvisor dataset to perform sentimental analysis, and the model exhibits lower error rates than Nave Bayes, SVM, XGBoost, and other deep learning models. However, the MCDM-based model faces various difficulties, such as an inadequate collection of terms and some words with inaccurate emotion ratings. It is also unable to handle emoji evaluations and sentences with negation that are projected to be negative responses.

Paulo Rita, Celeste Vong, Flavio Pinheiro, and Joao Mimoso authored a research article titled "A sentiment analysis of Michelin-starred restaurants" in 2022. Their primary concern is looking at how internet reviewers feel about four important factors (food, service, ambiance, and pricing), and then using that information to provide star ratings for choosing Michelin-starred restaurants. They do so by leveraging the Tripadvisor dataset, which consists of 8,871 reviews in total that were retrieved using a web crawler created by Beautiful Soup from 87 eateries throughout Europe. They used a semantrica (lexalytics) technique and emotional analysis for it.

The 2022 article "Arabic Sentiment Analysis of Eateries' Reviews Using Deep Learning" authors are Leen Muteb Alharbi and Ali Mustafa Qamar. This study makes use of a dataset that includes 1371 user reviews of several restaurants in the Qassim area. The reviews are written in Arabic, the local language, and deep learning methodology was utilized to obtain correct results in this article. The LSTM classifier, which is a recurrent neural network (RNN) extension, was employed. The authors obtain the greatest accuracy for these datasets using SVM, followed by logistic regression, KNN, and then the LSTM technique, which yields an accuracy of 83%.

Authors Jing Tian, Wushour Slamu Miaomiao Xu, Chunbo Xu, and Xue Wang added "Research on Aspect-Level Sentiment Analysis Based on Text Comments" in 2022, which incorporates the most recent NLP development trend, merges capsule network, and BERT, and offers the basic model CapsNet-BERT. The dataset used in this study is the SemEval 14 Restaurant Review dataset, and the model used is the CapsNet-BERT model, which has four layers (the embedding layer, the encoding layer, the primary capsule layer, and the category capsule layer). This model substitutes the embedding and encoding layers of CapsNet with pre-trained BERT, and the output is for laptop reviews because it has an accuracy of 79.53%. There are two subtasks for aspect-based sentiment analysis: ACSA and ATSA. The CapsNet-BERT model has accomplished good results in user review sentiment feature extraction and sentiment analysis.

"Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews" In 2023, Hengyun Li a, Bruce X.B. Yu b, Gang Li, and Huicai Gao published ABSA-based emotional analysis, which is a method utilized for restaurant datasets of two world-famous tourism sites in the United States. As a consequence, the ABSA model outperforms models that use general online review sentiment in terms of predictive power.

Ahmad Adel Abu-Shareha, Qusai Y. Shambour, and Mosleh M. Abualhaj wrote the following paper. The work was titled "Restaurant Recommendations Based on Multi-Criteria Recommendation Algorithm" and was published in 2022. The major goal of the article is that selecting restaurants takes time for users, thus they provide multi-criteria recommender systems that can use the multi-criteria ratings of users to understand their tastes and suggest the most suitable places for them to explore. This was used for The following real-world datasets are used: 1) The TripAdvisor dataset contains 14,633 multi-criteria reviews for 205 restaurants from 1,254 individuals. 2) In the TripAdvisor dataset, 1039 customers rated 693 hotels. 3) There were 1716 individuals in the Yahoo! Movies dataset who evaluated 965 films. They employ a multi-criteria recommender algorithm in this research. And the results of the proposed method vs the baseline techniques on the dataset indicate varied levels of sparsity in terms of MAE and RMSE.

" Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes " This document was written in 2021 by Murtadha M. Hamad, Mohanad Ahmed Salih, and Refed Adnan Jaleel and is beneficial for customer satisfaction study on restaurant customer feedback. The main algorithm used in this paper is naive Bayes classification, as well as some preprocessing techniques such as trimming lowercase, removing punctuation, and stop word removal for the Kaggle dataset, which contains 1000 tweets from different restaurants. With this algorithm, they get 73% accuracy,a 27% error rate, 68% precision, and 80.07% recall.

"Restaurant recommender system based on sentimental analysis" This research paper was written by lham Asani, Hamed Vahdat-Nejad, and Javad Sadri in 2021 the consistent approach of recommender systems their main purpose is to cluster the names of foods extracted from user's comments and analyze their sentiments about them for that they use TripAdvisor website data where 100 users were randomly selected and their reviews were collected between January and October 2018. Tokenization, stop word removal, stemming, noun extraction, and noun filtering (wordnet) were all used in this research. Also, for these projects, clustering approaches like hierarchical agglomerative clustering is utilized, and the Wu-Palmer method has demonstrated improved accuracy. The suggested system may provide users with 92.8% accuracy in the Top5 mode, according to the results. A recommender system has also been suggested, which identifies local eateries that fit the user's culinary tastes.

The title of the study is "Sentimental Analysis On Restaurant Reviews." In 2021, writers Parandham G and Mr. Raghavendra R wrote a book that provides some beneficial information on several techniques of emotional analysis. The goal of this work is to develop a Sentiment Analysis method for customer review categorization, which may be useful in analyzing data where opinions are extremely unstructured and either positive or negative. They acquired data for this study from various Google reviews, tweets, and other sources, and they used multiple algorithms for categorization such as linear regression and the naive multinomial algorithm. Lexical analysis, syntactic analysis, semantic analysis, disclosure integration, and pragmatic analysis are examples of natural language processing (NLP) approaches. Furthermore, this work presents research on essential ways for identifying sentiment analysis of reviews. We can infer that the main approaches are appropriate for identifying sentiment analysis of the review text to be analyzed.

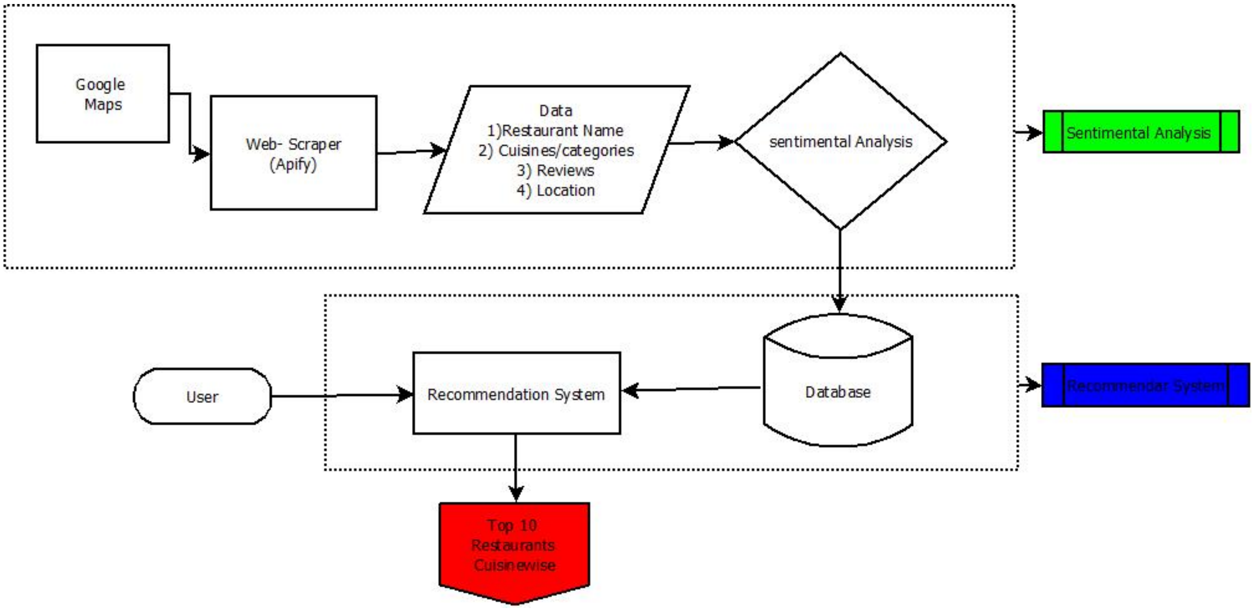
"SENTIMENT ANALYSIS USING BERT ON YELP RESTAURANT REVIEWS" This paper provides detailed knowledge of the BERT model presented by author Sunmin Lee in 2022. The purpose of this is to be a BERT model that determines whether a customer's Yelp review is positive or negative, as well as the degree of said positivity or negativity. For this, they use Yelp's businesses, reviews, and user data, which is accessible via Yelp.com/dataset. And the techniques used in these, such as tokenization and removing stop words, keyword analysis, and BERT model preparation, produce results such as epoch 0 having an average training loss of 1.61 and an accuracy of 0.27, epoch 1 having an average training loss of 1.60 and an accuracy of 0.27, and epoch 2 having an average training loss of 1.60 and an accuracy of 0.28. The BERT model was tested, and the best result produced had an accuracy of 0.28.

3.2 Background of sentimental analysis:

Sentimental analysis is also known as opinion mining this technique is first of all used in early 2000 when researchers like Pang and Lee used terms like subjectivity and polarity and for that algorithms used like SVM and naïve Bayes before that semantic orientation were used for stock market reports for telling bill and bears are opposite to each other using adjectives that were done by Hatzivassiloglou and McKeown and their model gets 90% precision. Then sentimental analysis grows in different ways so many machine learning, lexicon-based, and deep learning approaches are coming in front like decision trees, Bayesian networks, maximum entropy, and so many others. Even deep learning approaches also come like recurrent neural networks (RNNs) and convolutional neural networks (CNNs). So even sentimental analysis can be said a branch of NLP.

4. METHODOLOGY

The study of the RECOSYS system is divided into mainly two sections first one is the sentimental analysis and the second part or section is the recommendation system the main motive of introducing this system is to provide the best options for restaurants to users based on sentimental analysis and provide different type of star rating to them according to the past user's experience and their feedback. For that here is the main structure of the RECOSYS system shown in this figure.



4.0.1 RECOSYS architecture

From this system, you will get a rough idea of the process. In the second part recommendation user will select their appropriate or desirable Cuisine and according to that system shows them the top 10 restaurant options as a result. The deep methodology is discussed below.

4.1 **Problem Statement:**

Provide Restaurants recommendations based on customer reviews, with the help of sentiment analysis.

4.2 **Programming Method or Technical Things:**

For this study, we used Python programming language with the Jupyter Notebook platform also for the frontend page we used visual studio code. The main reason behind using this language is it provides different libraries like Numpy, Pandas, matplotlib, seaborn, Sci-kit learn, Plotely, nltk, streamlit, geopy, etc.

Also for web scraping, we used Apify web scraper one of the web scraping toolkits. And the web page we used ngrok software that allows you to expose a local server running on your machine to the public internet, making it accessible from anywhere in the world.

4.3 **Data:**

This section is related to how the data collection process is done, the source of the data, the main features of the data, and also how data was analyzed these things are deeply explained here. This data is Google Maps review data of Pune city restaurants which are located in Maharashtra, India.

Data consists of 1258 restaurants and around 196 unique cuisine types present. Also, this data shape is 754019 rows × 21 columns.

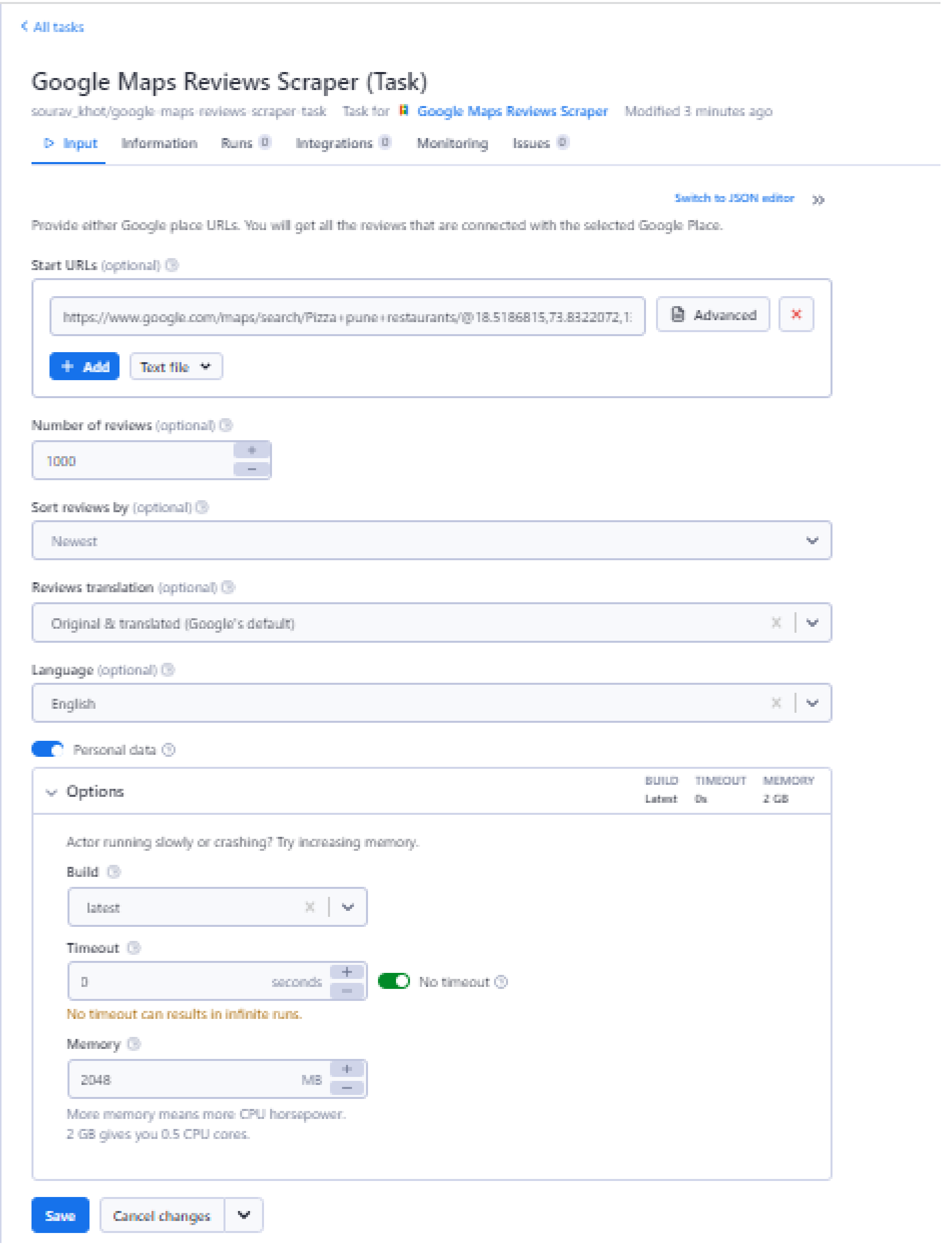
4.4 **Source of the Data:**

For this study data is collected from Google maps. Google Maps is the platform where all business-related shops are presented, from the user's location it gives the nearest shop options to users as well as directions towards that shop. these data are scrapped using the Apify web scraper. This data consists of nearly all Pune city restaurant information like the restaurant's name, locations, reviews, etc.

4.5 Data Collection:

This dataset is scrapped using the Apify web scraper their link is https://apify.com/apify/web-scraper this application contains different scraping options like Amazon product scraper,

YouTube scraper, yelp scraper, Facebook Likes scraper, etc. like these options are available from that for these projects we used google maps review scraper in this section there are some search options present according to that we can scrap reviews. Here is one view of that scrapping tool.



Google Maps Reviews Scraper (Task)

sourav_khot/google-maps-reviews-scraper-task Task for Google Maps Reviews Scraper Modified 3 minutes ago

Provide either Google place URLs. You will get all the reviews that are connected with the selected Google Place.

Start URLs (optional): https://www.google.com/maps/search/Pizza+pune+restaurants/@18.5186815,73.8322072,1:

Number of reviews (optional): 1000

Sort reviews by (optional): Newest

Reviews translation (optional): Original & translated (Google's default)

Language (optional): English

Personal data

Options

BUILD	TIMEOUT	MEMORY
Latest	0s	2 GB

Actor running slowly or crashing? Try increasing memory.

Build: latest

Timeout: 0 seconds No timeout

No timeout can results in infinite runs.

Memory: 2048 MB

More memory means more CPU horsepower.
2 GB gives you 0.5 CPU cores.

4.5.1 Apify software webpage

This is a page from which we scraped reviews there are some options present from that by our convenience we can scrap different restaurant names, restaurant reviews, and like these data.

Here are some options and how that help us to scrap review their explanation as below:

1) Start URLs:

   This section is useful for adding which site reviews users need for these projects we need restaurant data with cuisine so we added it according to that URL. Here are some URLs that were added in this section.

   - https://www.google.com/maps/search/misal+pune+restaurants/@18.515845,73.8584729,15z/data=!3m1!4b1
   - https://www.google.com/maps/search/chinese+pune+restaurants/@18.5158047,73.8584729,15z/data=!4m2!2m1!6e5
   - https://www.google.com/maps/search/Biryani+pune+restaurants/@18.5158853,73.8584729,15z/data=!4m2!2m1!6e5
   - https://www.google.com/maps/search/Bakery+pune+restaurants/@18.5159256,73.8584729,15z/data=!3m1!4b1
   - https://www.google.com/maps/search/japnese+pune+restaurants/@18.5161802,73.8322079,13z/data=!3m1!4b1

   these are some examples of URLs we used for scrapping.

2) Number of reviews:

   This section is useful for how many reviews from one restaurant are needed according to that Apify web scraper will be scraped. for that, we can set a limit of 1000 or 5000 reviews per restaurant for these projects we used around 3000 reviews per restaurant.

3) Sort reviews by:

   This section is useful for which type of reviews the user wants according to that software will scrap. options for these sections present as the highest rating, lowest rating, or newest. From these, we selected the newest reviews for the project.

4) Reviews translated:

   This is useful for getting local language reviews to translated reviews means if there is any review in the local language then it converts to the English language. In our case, some of the reviews are in the Marathi language so they converted into the English language.

5) Language:

   According to the user, they will select the language from which they will get all data, so here we selected English as the language.

6) Timeout:

   A timeout is an option that is useful for how much time scraping should work for that purpose it is available there is one option is present no timeout that is selected for these projects this option takes unlimited time for all scrap.

If we are done with all of the filters, then just save the information and run the actor and reviews scrapping will be started for these data to scrap all we need around 3 hours 15 min time then we stored these data into .CSV (comma-separated values) format file. Before saving these files they are given an option like which features are needed, then we applied some filters and stored around 22 features and stored it. The total size of this file is 250 MB. This file is saved as new_reviews_data.csv in the c folder.

4.6 <u>Apify software working:</u>

Apify software's main work is crawling and scraping data from different websites using the Chrome browser with JavaScript code and converting it into pages. For scraping apify's actors supports working with the crawling process as well as a list of URLs process from which they get optimum performance. In these tools for crawling and scraping, they used Crawlee. It is a browse automation and web scraping library available in npm package. The crawler library is one of the Node.js modules and their coding is done in the JavaScript language.

4.7 **Data Description:**

The new_reviews_data.csv file includes restaurant data such as the title of the restaurant, text or reviews, location, categories, stars, total score, etc. The description of the features and their data type used in the new_reviews_data.csv file are listed below.

- reviewId: This column contains the unique review id of the users
- categoryName: which type of restaurant is told in this column
- location/lat: location of the restaurant in Latitude
- location/lng: location of the restaurant in Longitude
- reviewsCount: Total count of the reviews
- address: address of the restaurant
- text: text or reviews of the restaurant as a string.
- title: restaurant name
- categories/n: different cuisine names in the restaurants (where n = 0,1,2,3,4,5,6,7,8 9)
- stars: star rating given by any customer
- totalScore: average star rating of any restaurant

4.8 **Exploratory data analysis (EDA):**

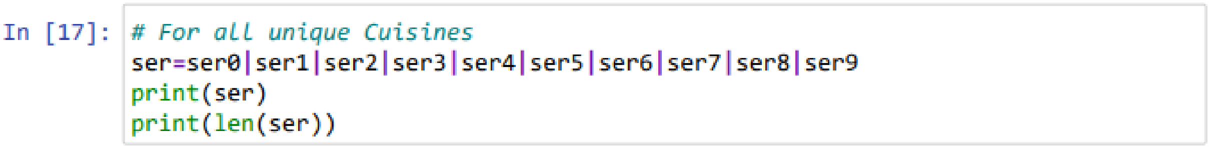
Exploratory data analysis is a pre-part of any project it is important to understand how data goes or how much amount of data is there also to know the datatypes of the features for that purpose Exploratory data analysis is performed. Or we can say that EDA is useful for finding some basics insides from raw data and understanding the data well. This method is performed in almost every project.

4.9 **Data Preprocessing:**

data preprocessing plays a crucial role in our project, as combining and finding unique Cuisines, removing nan value, and converting text or reviews to string values like these small things role is more in these projects.

Combining and finding unique cuisine:

In our data, we have about ten columns of categories/cuisines from which we have to find unique values and save them into one variable. for that, we used a set datatype that helps us to store all data into a variable. firstly, we have to find unique values from each column so we used the .unique() function that helps to find all unique values, then the same function applied to all categories columns and get unique values from all of them. Then for combining we use the Bitwise Or operator (|) that helps us to find all combined unique cuisine into one variable and apply the length function for understanding the count of unique cuisines.



```
In [17]: # For all unique Cuisines
         ser=ser0|ser1|ser2|ser3|ser4|ser5|ser6|ser7|ser8|ser9
         print(ser)
         print(len(ser))
```

4.9.1 Unique cuisine code

In these codes, ser0 to ser9 are the 10 variables of unique cuisines, and ser variable is their universal set of all of them. As a result, we get 200 unique Cusines from the ser variable. after that from this variable, we removed the ‘nan’ value contained because there is at some places nan category type is present which does not make any sense so we removed that, and our total Cuisine count comes as 196.

Converting datatypes:

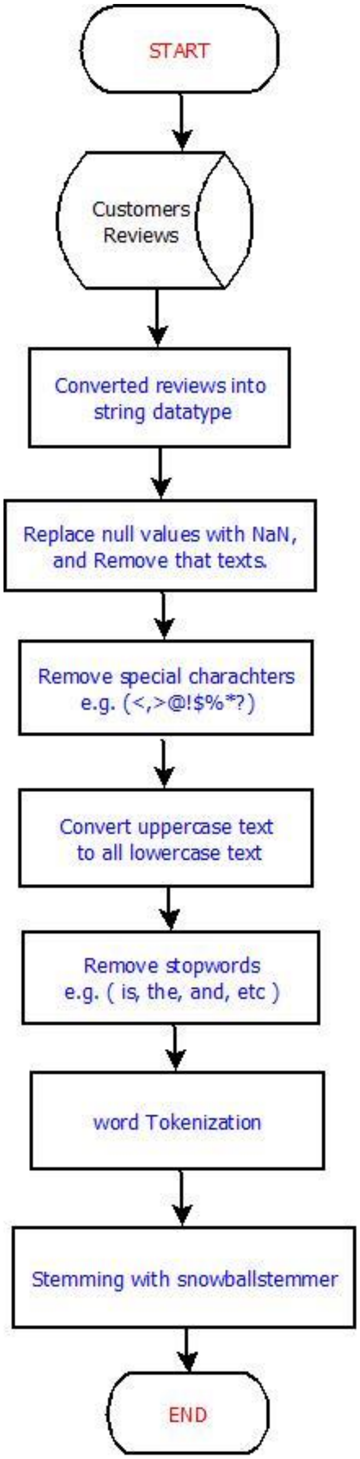
Converting datatypes is one of the preprocessing techniques that users use to get any data in

an appropriate or usable way in these projects we have text data in object data type so we converted text data column to string data type.

Handling missing text:

Handling missing values is one of the important tasks in any project because the missing values or none type text is not useful. In our case also it is not useful so we have to omit those for that we used .dropna() function that removes all nan values and gets all text data without missing values.

Here, are some preprocessing techniques and methods used on text data to convert it in useful manner.


START
Customers Reviews
Converted reviews into string datatype
Replace null values with NaN, and Remove that texts.
Remove special charachters e.g. (<,>@!$%*?)
Convert uppercase text to all lowercase text
Remove stopwords e.g. (is, the, and, etc)
word Tokenization
Stemming with snowballstemmer
END

4.9.2 Text Preprocessing

4.10 Models and concepts used in this Study:

- **Removing Special Characters:**

  It is frequently beneficial to remove special characters from text data before doing sentiment analysis to clean and normalize the input. A frequent preprocessing step is eliminating special characters from text. This procedure includes removing non-alphanumeric or whitespace characters. Punctuation marks, symbols, and emojis are frequently unrelated to the mood communicated in the text and might contribute noise to the analysis. You may reduce the text and focus on the important words and context by deleting special characters. This aids in the standardization of input data and the accuracy of sentiment analysis algorithms. It is crucial to remember, however, that eliminating special characters may result in some information loss, especially if the characters convey sentimental value, such as emoticons or specialized symbols used in specific areas. Using regular expressions or built-in string manipulation routines, several computer languages and libraries provide various methods for removing special characters. The particular implementation will differ based on the language or library used.

- **Converting upper case to lower case:**

  Text conversion to lowercase is a frequent preprocessing step in sentiment analysis. This procedure entails converting all capital characters in the text to lowercase counterparts. By changing the text to lowercase, you ensure that words with various casings are treated as the same term by the analysis. Text conversion to lowercase aids in standardizing input data and lowering vocabulary size. It guarantees that terms such as "happy" and "Happy" are recognized as synonymous, allowing the sentiment analysis algorithm to capture the sentiment regardless of the case used in the text. This stage is crucial since sentiment may be expressed regardless of the letter case. Python programming language and libraries provide built-in functions or methods to convert text to lowercase. In these projects, we create one built-in function as to_lower and applied a lower method to get text in lowercase.

- **Removing Stopwords:**

  Stopwords are popular words that have no important significance and are frequently filtered out to decrease noise and enhance analysis accuracy. Stop words are words that appear often in the English language but add nothing to the feeling or meaning of a phrase, such as "the," "is," "and," "in," and so on. By eliminating stopwords, the emphasis is moved to more informative words that can communicate emotion. Removing stopwords helps in reducing the vocabulary size, improve computational efficiency, and eliminating noise from the text data. It allows sentiment analysis algorithms to prioritize and weigh the sentiment-bearing words more effectively. However, it's important to note that the list of stopwords may vary depending on the specific language, domain, or context of the analysis. There are some approaches to remove stopwords like firstly importing stopwords from nltk.corpus library which gives English language stopwords and stores them in one set, then creates one function for removing stopwords and applying them to that.

- **Word Tokenization:**

  The act of breaking down a sequence of text, such as a one sentence or an article, into individual words or tokens is known as word tokenization. It's an important stage in natural language processing (NLP) and text analysis activities. From nltk.tokenize we import word_tokenize which helps us to break the sentence into small chunks and store it.

  e.g. “I love Python!” These text tokens are “I”, “love” and “Python”. Many NLP applications, such as text classification, machine translation, and sentiment analysis, rely on word tokenization. It aids in the conversion of raw text into a structured format that machines can interpret, analyze, and understand.

- **Snowball stemming:**

  Snowball stemming, also known as the Porter stemming method, is a popular natural language processing (NLP) stemming approach. The act of reducing words to their base means root form or natural form, which aids in normalizing text data and enhancing text analysis, is known as stemming. The Snowball stemmer is an expansion of Martin Porter's original Porter stemming method. It supports stemming in various languages, making it adaptable to a wide range of applications. The main goal of the Snowball stemming algorithm is to remove suffixes from words to obtain their root form. By removing suffixes, different word forms that share a common stem are reduced to the same base word. For example, words like "running," "runs," and "ran" would all be stemmed from "run."

Snowball stemming is a rule-based approach that does not use any outside resources or machine learning. It is rather quick and efficient, making it suited for a wide range of NLP jobs. The stemmed words produced by the Snowball stemmer can be utilized for applications such as text categorization, search engines, and sentimental analysis. Stemming reduces text data dimensionality, consolidates comparable word forms, and improves the accuracy and efficiency of text analysis and retrieval systems.

- **Polarity:**

  In sentiment analysis, polarity relates to a text's sentiment or emotional orientation, which is commonly described as positive, negative, or neutral. It is a critical component of sentiment analysis, which seeks to ascertain the underlying sentiment or opinion communicated by a written document. Polarity calculation is done by using Textblob. Textblob is a Python package based on NLTK (Natural Language Toolkit) and provides a higher-level API for working with text data for tasks such as sentiment analysis. Polarity classification is the task of automatically assigning sentiment labels to text data based on its emotional tone. The sentiment labels typically include positive, negative, and sometimes neutral. Polarity classification can be performed at various levels, such as document level, sentence level, or aspect level (targeted sentiment analysis). The polarity score range is in [-1,1].

- **Subjectivity:**

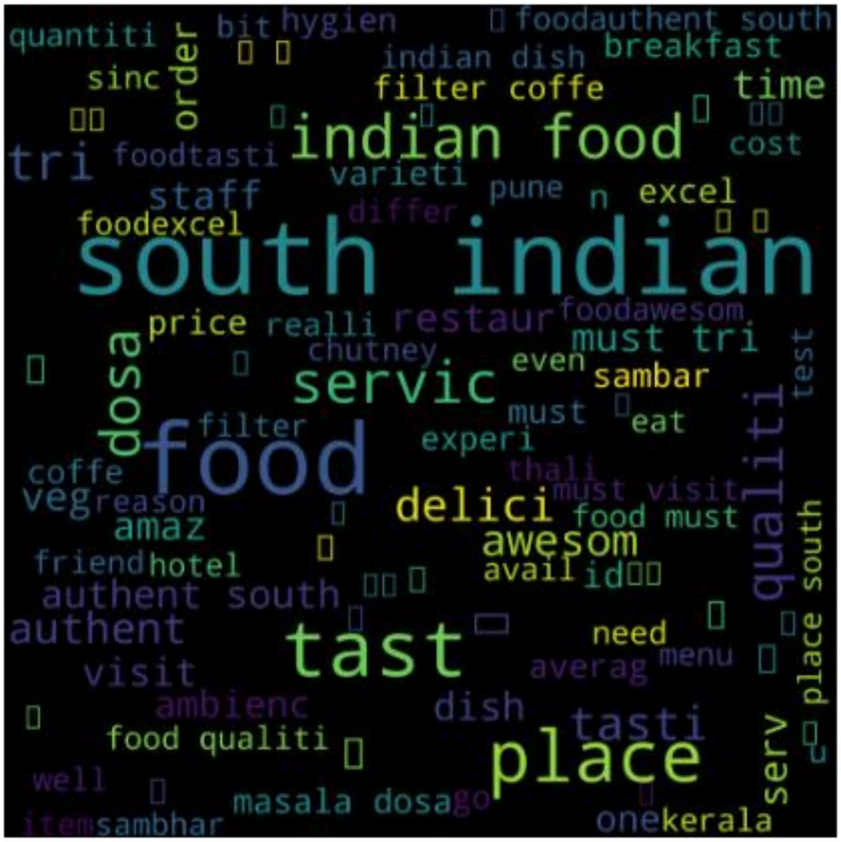
  Subjectivity classification in sentiment analysis seeks to evaluate whether a text represents subjective or objective information. Subjective writing expresses personal beliefs, sentiments, or emotions, whereas objective text conveys factual information or neutral statements. Also, Subjectivity categorization can be accomplished using a variety of methods, including rule-based approaches, machine learning algorithms, or a mix of the two. To identify subjective language signals, these strategies depend on numerous aspects such as linguistic patterns, sentiment lexicons, syntactic structures, or context. Subjectivity is calculated using the Pythons Textblob package their range is [0,1]. Where 0 is objective and 1 is subjective.

- **Segmentation with worldcloud:**

  In sentiment analysis, segmentation refers to the act of separating a text into smaller segments or units to analyze sentiment on a more granular level. Segmentation entails breaking down the text into smaller parts, such as paragraphs, phrases, or even individual sentences or clauses, rather than analyzing sentiment at the document or sentence level. For segmentation concept of lexical cohesion was used. The ability to compare sentiment across segments is enabled via segmentation. It is possible to find patterns, trends, or variances in sentiment expression by comparing sentiment ratings or polarity across segments. Segmentation classifies text as either positive, negative, or neutral.

  A word cloud is a type of data visualization approach that displays words in various sizes and places based on their frequency or relevance within a particular corpus. It provides a simple and aesthetically appealing method of grasping essential phrases or concepts within a body of material. There are various phases involved in making a word cloud. The text data is first preprocessed to eliminate unnecessary words, punctuation, and special characters. This keeps the emphasis on the most important terms. Following that, the frequency of each word is computed to determine its prominence. The larger a term appears in the word cloud, the more frequently it appears. Words are often randomized or organized in an aesthetically attractive pattern. When the computations are finished, the word cloud is created as a graphical representation, with words shown in various colors, typefaces, or styles to improve visual impact. Word clouds are often utilized in a variety of industries, including social media analysis, content analysis, and market research since they give a fast summary of the most important phrases in a dataset. However, it is critical to understand that word clouds have limits. They do not capture contextual information or word connections, and their efficacy is affected by preprocessing decisions and representation limits.

  Here we created a word cloud for neutral segmentation data so here it gives neutral words that have no values, and positive or negative impacts on text :



4.10.1 Word cloud for segmentation neutral reviews

- **Balancing data with the Resampling method:**
  Data balancing in sentiment analysis is a key step in addressing the class imbalance, which occurs when one sentiment class dominates the dataset while others are underrepresented. Resampling is one method for balancing the data, which entails altering the dataset to achieve a more equal distribution of sentiment classifications. To generate a more balanced dataset, resampling approaches try to either raise the minority class instances (oversampling) or reduce the majority class instances (undersampling). This helps to avoid bias towards the majority class and guarantees that the sentiment analysis model learns from a diverse group of samples from each sentiment class. The main benefits of sampling are Addressing Class Imbalance, Improved Performance, and better generalization.

- **Train-test split:**

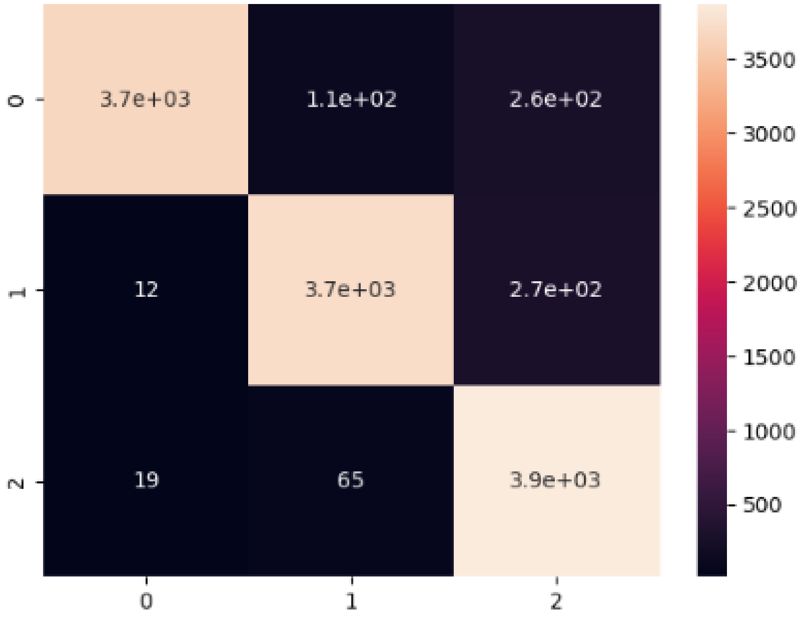
  A train-test split is a frequent practice in machine learning that is used to evaluate a model's performance on unseen data. It entails separating the available dataset into two subsets: training and testing. The training set allows the model to be trained by exposing it to labeled samples and allowing it to understand patterns and correlations in the data. The testing set, on the other hand, is used to evaluate the model's performance on previously unknown data. We may assess how effectively the model generalizes to new, unseen cases by evaluating it on the testing set. The train-test split is critical to avoiding overfitting, which happens when a model performs well on training data but fails to perform on new data. We can determine the model's capacity to generalize and generate accurate predictions on unknown data by assessing it on the testing set. The train-test split is usually done at random to guarantee an impartial representation of the data. The ratio of the training set to the testing set might vary based on the size of the dataset and the unique requirements of the study. Common split ratios are 80:20 (80% for training, 20% for testing) or 70:30 (70% for training data, 30% for testing data), or another way is split into 75:25 (75% for training data, 25% for testing data)

- **Model Building:**
  Model building is one of the most important tasks in machine learning without model building machine learning project will not be complete. So here for checking how many classes segmentation classifies right, we used different machine learning, probabilistic, and ensemble learning models like Random forest classifier, Logistic regression, Extra trees classifier, and Multinomial model, SVM. So here we tried model building on South Indian restaurant's cuisines text.

  - Random forest Classifier:
    Random Forest Classifier is approach of ensemble learning that makes predictions by combining numerous decision trees. It is a supervised learning method that is often used for classification jobs. The Random Forest Classifier works by creating an ensemble of decision trees, where each of which is trained on a randomly

selected portion of the training data. Each decision tree in the ensemble learns to forecast using a subset of characteristics and divides the data into multiple branches based on these attributes. the RFC introduces randomness in two main ways Random Sampling and a Random Subset of Features. The Random Forest Classifier integrates the forecasts of all the decision trees in the ensemble while making predictions. The most frequent prediction for classification problems is determined via majority voting, in which the class with the most votes from individual trees is chosen as the final forecast. Individual tree forecasts are averaged to generate the final prediction in regression problems. The main benefits of random forest are Robustness to Overfitting, Feature Importance, and it can handle missing data by utilizing other features to make predictions.

RPF is widely used in various domains due to its ability to handle high-dimensional datasets, deal with noisy data, and provide accurate predictions. It is regarded as a strong and adaptable classification algorithm that provides a good mix of predicted performance and computational economy. For RF we get a confusion matrix as follows:



4.10.2 Confusion matrix for RF

- Multinomial Model:

  Multinomial Naive Bayes is a probabilistic Model mostly used in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email, book, article, or newspaper story, using the Bayes theorem. It computes the likelihood of each tag for a given sample and outputs the tag with the highest likelihood. And it is mostly used for text data to classify and find probability its mostly used in sentimental analysis.

  The Bayes hypothesis, created by Thomas Bayes, decides the likelihood of an event occurring based on prior knowledge of event-related factors. It is calculated using the following formula:

$$\boldsymbol{P(S|K) = P(S) * P(K|S)/P(K)}$$

  Where,

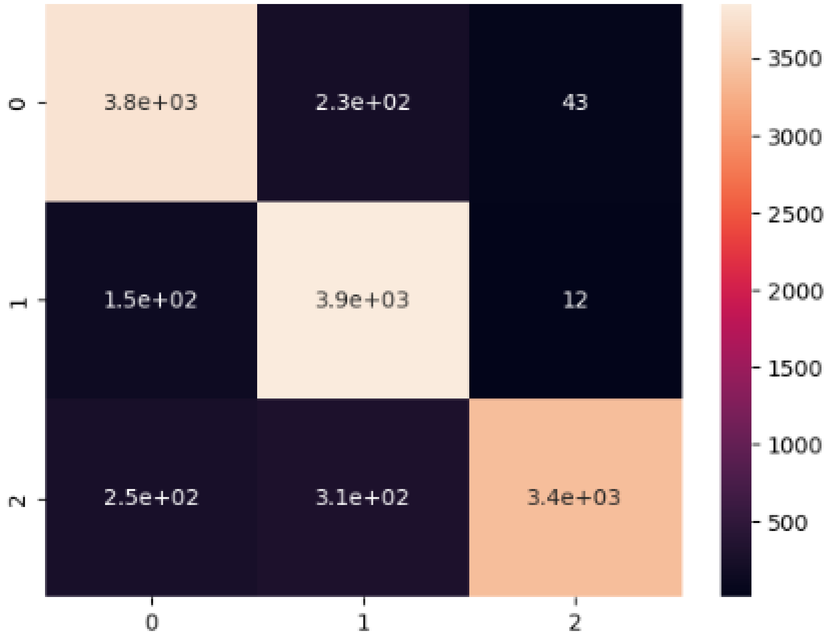
  $P(K)$ = prior probability of K

  $P(S)$ = prior probability of class S

  $P(K|S)$ = occurrence of predictor K given class S probability

  The class with the highest probability is picked as the projected class label for the given input feature vector x when using the Multinomial Classifier to generate a prediction. This method comprises comparing the probabilities for each class and selecting the one with the greatest likelihood.

  The binary logistic regression model is expanded by the multinomial classifier, a multi-class classification model. It offers a mathematical framework for computing class probabilities based on the given features. The model selects the most likely class label for a given input by using the softmax function to assign probabilities to each class. This mathematical structure makes the Multinomial Classifier an excellent tool for addressing classification problems with several classes. For the Multinomial Model we get a confusion matrix as follows:



4.10.3 Confusion matrix for multinomial model

- ExtraTreesClassifier Model:

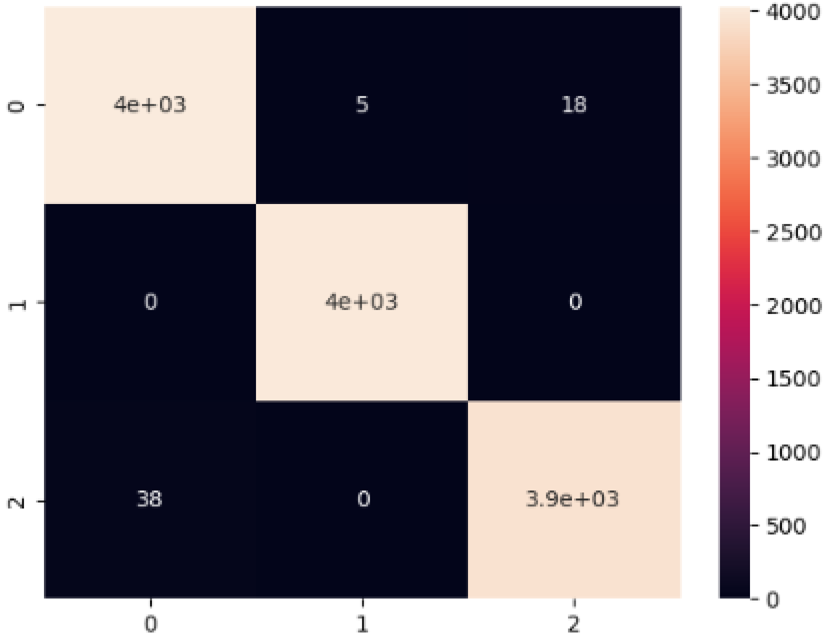
  The ExtraTreesClassifier is an ensemble learning model that makes predictions by combining numerous decision trees. It is a Random Forest method modification that offers more randomness to further vary the individual trees.

  The ExtraTreesClassifier is based on the notion of decision trees and the aggregation of predictions from them. The ExtraTreesClassifier generates a forest of decision trees given a training dataset D of N samples (x_i, y_i), where x_i represents the input features and y_i represents the matching class labels.

  The ExtraTreesClassifier grows each decision tree by randomly picking a subset of characteristics at each split point. This randomization aids in reducing overfitting and increasing tree variety. To select the optimum attribute and threshold for dividing the data, the splitting criterion used for each tree is dependent on several parameters such as Gini impurity or entropy.

  During prediction, each decision tree in the ExtraTreesClassifier provides a class label to an input feature vector x separately based on a majority vote or an averaging of the class labels in the leaf nodes. After then, the final forecast is determined by combining the predictions of all the trees in the forest.

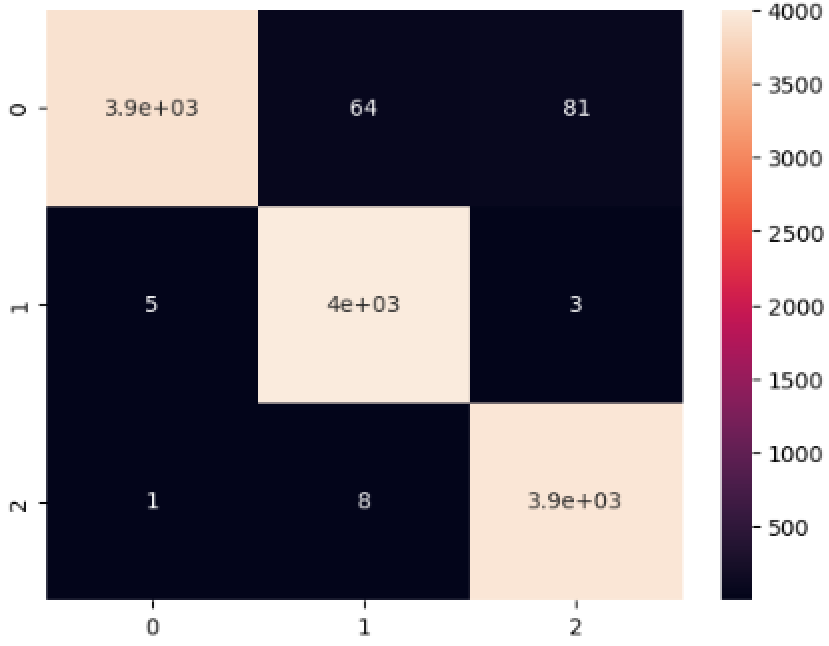
The ExtraTreesClassifier's mathematical formulation entails the building and aggregation of decision trees using random feature subsets and splitting criteria. However, the real ExtraTreesClassifier implementation includes complicated computations and algorithms to efficiently generate and merge the trees. For ExtraTreesClassifier Model we get a confusion matrix as follows:



4.10.4 Confusion matrix for ExtraTreesClassifier

- Logistic Regression:

A statistical model used for binary classification tasks is logistic regression. Based on the supplied characteristics, it calculates the likelihood of an instance belonging to a specific class. The model assumes a linear relationship between the characteristics and the probability log odds, and it employs a logistic (sigmoid) function to transfer the linear combination of features to a probability value. The logarithmic loss function is used to optimize logistic regression, which is learned via maximum likelihood estimation. It can take both numerical and categorical inputs and produces interpretable coefficients for each feature. While logistic regression is efficient and easy to understand, it may underperform in scenarios with complicated decision boundaries or non-linear correlations between features and the target variable. For logistic regression, we get a confusion matrix as follows:



4.10.5 Confusion Matrix for Logistic Regression

- KNN classifier:

  The K-Nearest Neighbours (KNN) classifier may be used to categorize text input into positive, negative, or neutral sentiment categories. The KNN technique is based on the idea that examples are similarly likely to belong to the same class.

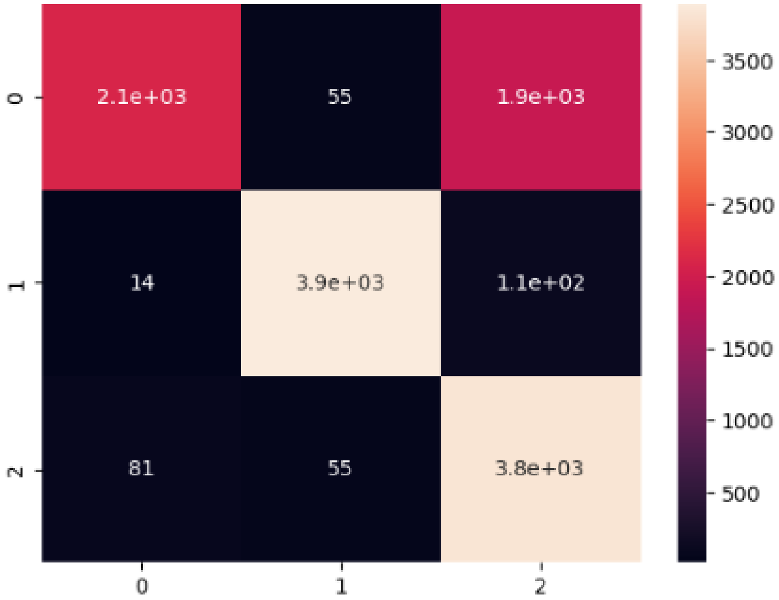
  To begin, the dataset must be constructed, which consists of text samples with accompanying sentiment labels. The next step is to convert the textual data into a numerical representation that the KNN classifier can use. This is possible using approaches such as bag-of-words or TF-IDF, which turn text into numerical vectors based on word frequency or relevance.

Once the text has been numerically represented, the KNN classifier uses a distance metric such as Euclidean or cosine distance to determine the similarity between the input text and the training examples. Based on their numerical representations, this similarity assessment assesses how closely connected two bits of text are.

Based on the determined similarities, the KNN classifier then chooses the k closest neighbors to the input text. The emotion of the supplied text is determined using the nearest neighbors. The KNN classifier assigns a sentiment label to the input text by conducting a majority vote among the sentiment labels of the k closest neighbors. In other words, the anticipated emotion is picked from the sentiment category with the highest representation among the neighbors

The KNN classifier's performance may be assessed by comparing the predicted sentiment labels to the actual sentiment labels in the dataset. Precision, recall, and F1 score are accuracy measures that may be used to evaluate the classifier's efficacy in sentiment categorization.

In emotional analysis, the KNN classifier provides a basic and intuitive technique for categorizing text into sentiment categories. However, numerous parameters must be addressed, such as the text representation approach utilized, the distance metric used for similarity computation, and the value of k, which affects the number of neighbors examined. These judgments should be based on the dataset's particular properties and the intended performance of the classifier. For KNN classifier, we get a confusion matrix as follows:



4.10.6 Confusion matrix for KNN classifier

- SVM:

  The Support Vector Machine (SVM) is a supervised machine learning model used for classification and regression tasks. The Support Vector Classifier (SVC) is a variant of SVM used specifically for classification. In SVC, the model aims to find a hyperplane in the feature space that maximizes the margin between different classes while minimizing classification errors. The hyperplane is determined by a subset of the training data called support vectors. For binary classification, let's consider a training dataset with input vectors $X = \{x_1, x_2, ..., xn\}$ and corresponding class labels $Y = \{y_1, y_2, ..., yn\}$, where $yi \in \{-1, +1\}$. The goal of SVM is to find a hyperplane defined by the equation:

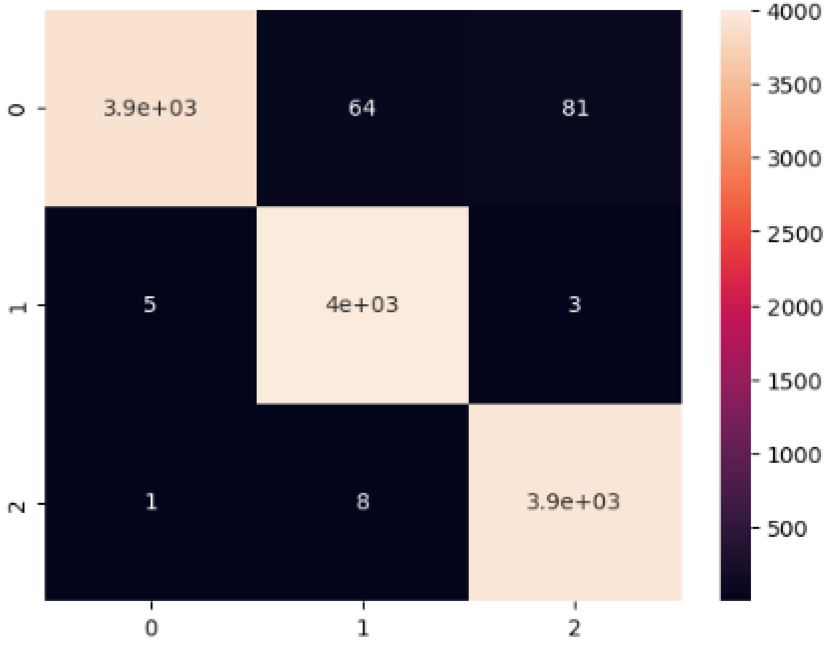
$$w.x - b = 0$$

Here, w represents the normal vector to the hyperplane, x represents the input vector, and b is the bias term. The objective of SVM is to maximize the margin, which is the distance between the hyperplane and the nearest data points, known as support vectors. Mathematically, this can be formulated as an optimization problem:

minimize $1/2\ ||w||\text{^}2$ subject to $yi(w.xi - b) \geq 1\ for\ all\ i$

In this formulation, $||w||\text{^}2$ represents the squared norm of the weight vector w, and the constraint $yi(w.xi - b) \geq 1$ ensures that the data points are correctly classified on the appropriate side of the hyperplane.

SVM can find the ideal hyperplane that maximizes the margin and produces high classification performance by solving this optimization issue. To effectively handle this issue, a variety of optimization approaches, including quadratic programming, can be applied. By analyzing the equation $w.x - b$, fresh, unobserved data points can be categorized after the ideal hyperplane has been found. The data point belongs to one class if the outcome is good, and to the other class if it is negative. By employing kernel functions, which implicitly translate the input vectors into a higher-dimensional space, SVM is also able to solve nonlinear classification problems. This makes it feasible to identify nonlinear decision boundaries.



4.10.7 Confusion matrix for SVM classifier

- Classification report:

  A classification report is a summary of a classification model's performance. It gives numerous evaluation measures, including accuracy, recall, F1-score, and support, for each class in the classification task. The precision of a class is the ratio of accurately predicted instances to the total expected instances of that class. It shows how well the model avoids false positives. The ratio of properly predicted occurrences of a class to the total instances of that class in the dataset is known as recall (also known as sensitivity or true positive rate). It assesses the model's ability to detect false positives. The harmonic mean of accuracy and recall is the F1-score. It gives a balanced metric that takes accuracy and memory into account.

4.11 New Star Rating:

This rating is dependent on the text, for that we applied the below formula on the polarity column that gives us a new star rating based on sentimental analysis so we can call this rating the sentiment star rating. The formula for these ratings is as below:

$$\textbf{Rating} = \frac{(\boldsymbol{Score - lower\ limit}) * (\boldsymbol{new_{score}\ upper\ \ limit\ \ - new_{score} lower\ limit}\ )}{(\boldsymbol{upper\ limit - lower\ limit})} + \boldsymbol{new_{score} lower\ limit}$$

Here,

Score: Polarity score

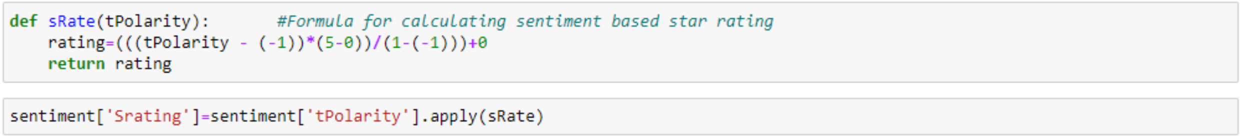
Lower limit: polarity lower limit which is -1.

Upper limit: polarity upper limit which is 1.

$\text{new}_{\text{score}}$ lower limit: It is a new rating lower score which is 0.

$\text{new}_{\text{score}}$ upper limit: It is a new rating upper score which is 5.

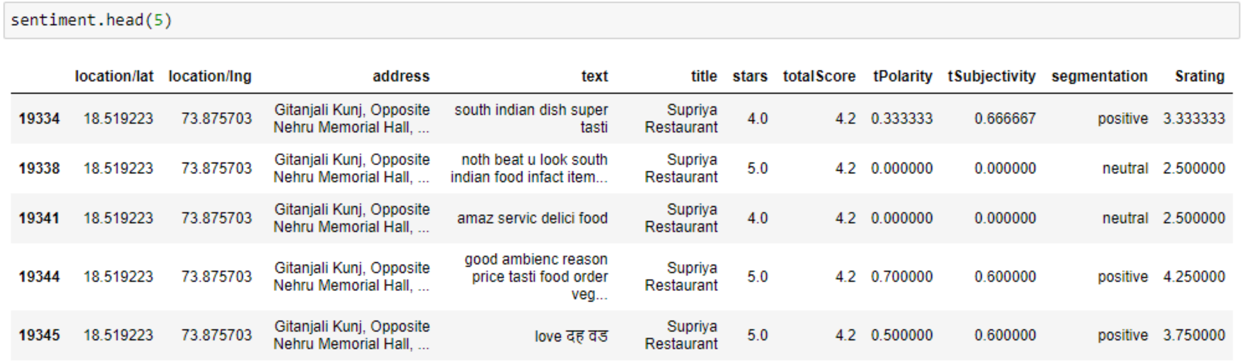
After applying this formula with function on data we get new results column as follows:



```
def sRate(tPolarity):       #Formula for calculating sentiment based star rating
    rating=(((tPolarity - (-1))*(5-0))/(1-(-1)))+0
    return rating
```

```
sentiment['Srating']=sentiment['tPolarity'].apply(sRate)
```

4.11.1 New rating Formula



```
sentiment.head(5)
```

	location/lat	location/lng	address	text	title	stars	totalScore	tPolarity	tSubjectivity	segmentation	Srating
19334	18.519223	73.875703	Gitanjali Kunj, Opposite Nehru Memorial Hall, ...	south indian dish super tasti	Supriya Restaurant	4.0	4.2	0.333333	0.666667	positive	3.333333
19338	18.519223	73.875703	Gitanjali Kunj, Opposite Nehru Memorial Hall, ...	noth beat u look south indian food infact item...	Supriya Restaurant	5.0	4.2	0.000000	0.000000	neutral	2.500000
19341	18.519223	73.875703	Gitanjali Kunj, Opposite Nehru Memorial Hall, ...	amaz servic delici food	Supriya Restaurant	4.0	4.2	0.000000	0.000000	neutral	2.500000
19344	18.519223	73.875703	Gitanjali Kunj, Opposite Nehru Memorial Hall, ...	good ambienc reason price tasti food order veg...	Supriya Restaurant	5.0	4.2	0.700000	0.600000	positive	4.250000
19345	18.519223	73.875703	Gitanjali Kunj, Opposite Nehru Memorial Hall, ...	love दह वड	Supriya Restaurant	5.0	4.2	0.500000	0.600000	positive	3.750000

4.11.2 Sample data

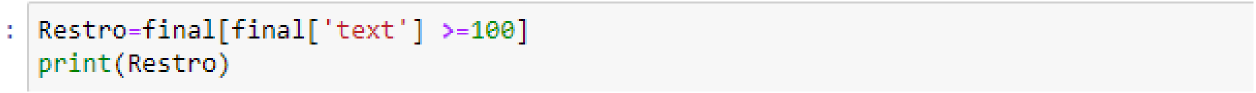
Here Srating is the column that gives rating reviews on the basis of sentimental analysis.

4.12 Recommender system:

After getting a new star rating created one recommender system that helps any customer to get top restaurant according to sentiment analysis rating and their nearby location for that we applied the group by function on data and calculated the total texts count as well as the mean of 'Srating' column concerning different restaurants. The 'Srating' column is nothing but new star rating based on sentimental analysis then arrange into ascending order and stored into one variable.

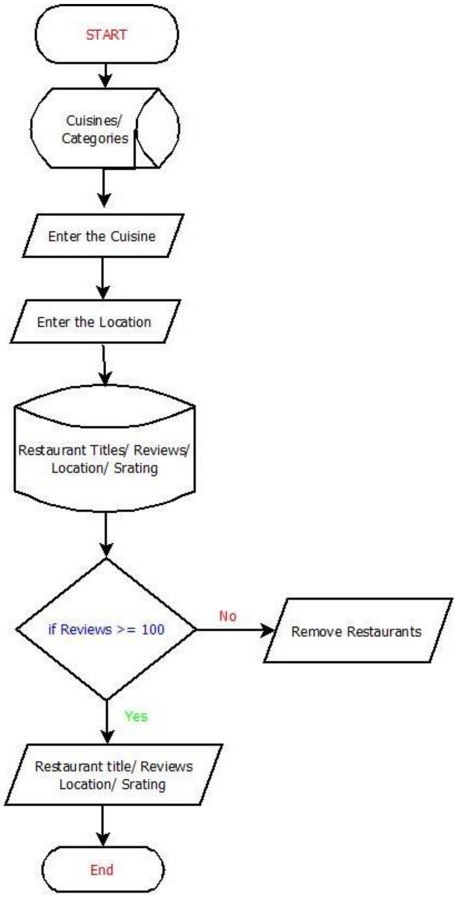
Some restaurants have lower reviews so that restaurant is not as popular than others so we removed those restaurants and applied one condition which is if the restaurant have more than 100 reviews then only it comes into the system.

Condition 1: Restaurant reviews/ Text >= 100



```
: Restro=final[final['text'] >=100]
  print(Restro)
```

4.12.1 Condition 1 code



4.12.2 Condition 1 Flowchart

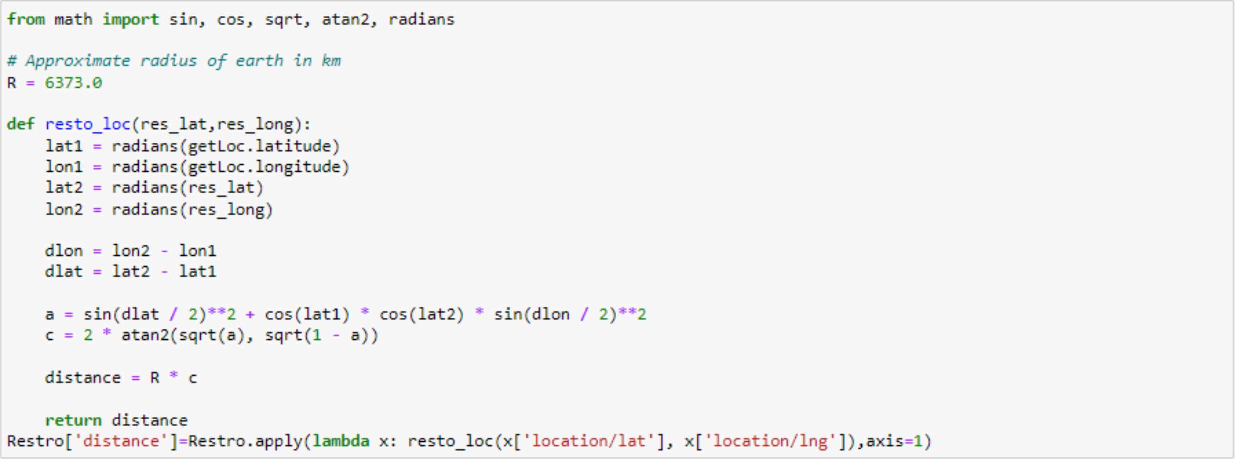
Calculating Distance:

For that thing, we created one user-defined function that gives the distance between restaurants and your current location or the user’s given location and finds a nearer restaurant. For that first we take users’ location from the input function and then from geopy.geocoders module we selected Nominatim is a geocoding service that allows you to convert addresses into geographic coordinates so whenever the user will select their location according to that their latitude and longitude we get it.

For the difference between restaurants and users we applied one function that calculated the difference between two points

For the calculation of distances between the user's location and restaurant locations, we defined the resto_loc function which takes latitude and longitude values as input. Inside the function, the latitude and longitude of the user's location (getLoc) and the latitude and longitude of the restaurant location are converted to radians. The differences in longitude and latitude are calculated, and the Haversine formula is applied to determine the central angle between the two locations. Using the central angle, the distance between the locations is calculated by multiplying the radius of the Earth by the central angle.

Finally, the resto_loc function is applied to each row of the Restro DataFrame using the apply() method. It calculates the distance between the user's location and each restaurant's location, and the resulting distances are stored in a new column named 'distance' in the Restro DataFrame.

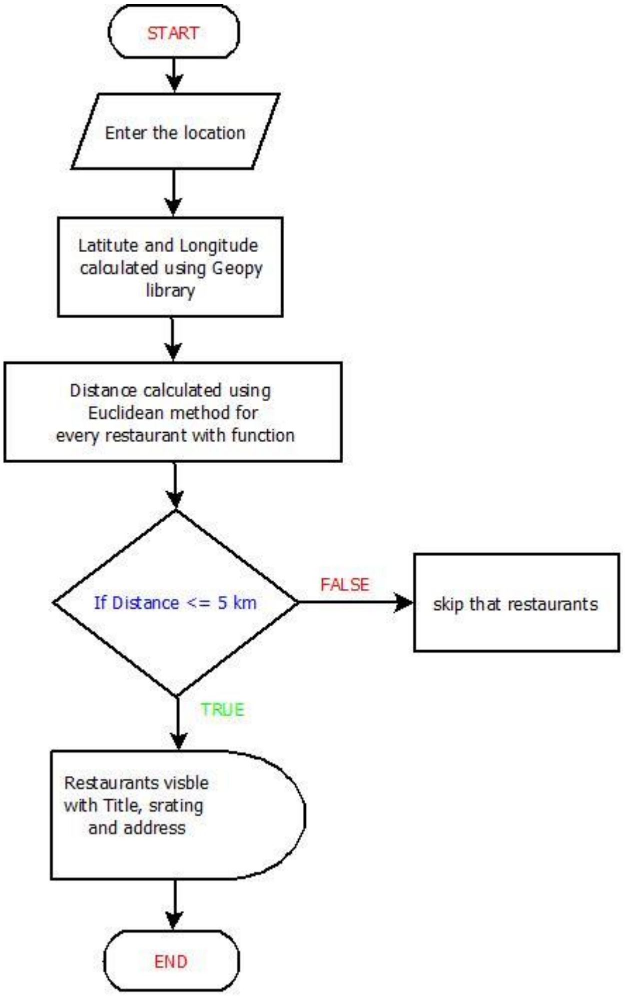


```
from math import sin, cos, sqrt, atan2, radians

# Approximate radius of earth in km
R = 6373.0

def resto_loc(res_lat,res_long):
    lat1 = radians(getLoc.latitude)
    lon1 = radians(getLoc.longitude)
    lat2 = radians(res_lat)
    lon2 = radians(res_long)

    dlon = lon2 - lon1
    dlat = lat2 - lat1

    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    distance = R * c

    return distance
Restro['distance']=Restro.apply(lambda x: resto_loc(x['location/lat'], x['location/lng']),axis=1)
```

4.12.3 Distance Calculation code

Condition 2: Restaurant distance <=5

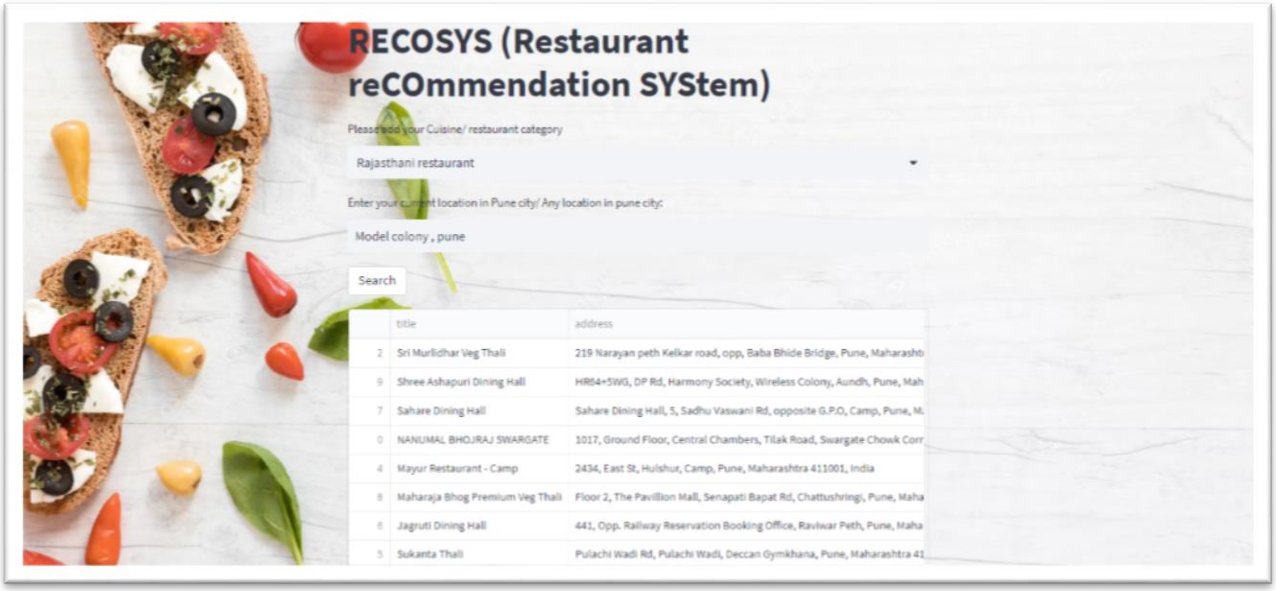
Here we have taken this condition because RECOSYS provides restaurants within the 5-kilometer buffer.


START
Enter the location
Latitute and Longitude calculated using Geopy library
Distance calculated using Euclidean method for every restaurant with function
If Distance <= 5 km
FALSE
skip that restaurants
TRUE
Restaurants visble with Title, srating and address
END

4.12.4 Condition 2

5. HOW RECOSYS WORKS ?

For user convenience created one web page with the help of the Streamlit library and with visual studio software that is created in that convenient way for the user, copy of that page as below:



RECOSYS (Restaurant reCOmmendation SYStem)

Please add your Cuisine/ restaurant category

Rajasthani restaurant

Enter your current location in Pune city/ Any location in pune city:

Model colony , pune

	title	address
2	Sri Murlidhar Veg Thali	219 Narayan peth Kelkar road, opp, Baba Bhide Bridge, Pune, Maharasht
9	Shree Ashapuri Dining Hall	HR64+5WG, DP Rd, Harmony Society, Wireless Colony, Aundh, Pune, Mah
7	Sahare Dining Hall	Sahare Dining Hall, 5, Sadhu Vaswani Rd, opposite G.P.O, Camp, Pune, M.
0	NANUMAL BHOJRAJ SWARGATE	1017, Ground Floor, Central Chambers, Tilak Road, Swargate Chowk Corr
4	Mayur Restaurant - Camp	2434, East St, Hulshur, Camp, Pune, Maharashtra 411001, India
8	Maharaja Bhog Premium Veg Thali	Floor 2, The Pavillion Mall, Senapati Bapat Rd, Chattushringi, Pune, Maha
6	Jagruti Dining Hall	441, Opp. Railway Reservation Booking Office, Raviwar Peth, Pune, Maha
5	Sukanta Thali	Pulachi Wadi Rd, Pulachi Wadi, Deccan Gymkhana, Pune, Maharashtra 41

5.0.1 RECOSYS webpage

From there, we can get the top 10 restaurant's names, addresses, Srating (sentiment analysis rating), and distance from your current location or you have allowed that location.

For the creation of this page, we used Streamlit library and pickle file also for making it interactive and providing to users we used ngrok software.

Streamlit library:

Streamlit is a Python package that aids in the creation of interactive web applications and data visualization dashboards. It simplifies the creation and sharing of apps by providing a simple and user-friendly interface. Streamlit excels in data exploration, visualization, and interactive components.

Streamlit allows you to write code in a logical and linear fashion, making it easy to build apps even for beginners. The library comes with several built-in functions for processing user input, creating visualizations, and displaying data tables. It also supports real-time updates, allowing users to interact with data and view changes in visualizations as they modify input parameters.

The ability of Streamlit to swiftly turn Python scripts into web apps is its core selling point. Stramlit assists in the development of functions that use a decorator-based approach to generate interactive components like as sliders, checkboxes, and dropdown menus.

Finally, we can state that streamlit is simple to use. Streamlit interacts with popular data science tools like Pandas, Matplotlib, and Plotly. You may then leverage these libraries' features to do data analysis, make visualizations, and integrate them directly into your Streamlit application. Streamlit also provides application sharing and deployment across several platforms, making it simple to share information with others.

Pickle file:

A pickle file is a binary-formatted serialized object stored with Python's pickle module. Pickling enables platform-agnostic storing and retrieval of trained machine learning models or large datasets.

Consider a pickle file to be a magic box. Put items into the box, close it, and keep it secure. When you need those items again, simply open the box and take them out. It's a convenient method to store and move data without having to worry about losing crucial data or having to recreate things from the start. Even we can say that the pickle file stores the data or connects with notebook code.

Pickling a Python object creates a byte stream that may be saved to a file or transmitted over a network. This byte stream holds the object's state, including its properties and methods. Pickling is a method that allows objects to be saved to a disc or transferred between computers while keeping their underlying structure and data.

Pickle files are often used in a number of situations. The usage of pickle files in a report may be explained in machine learning projects in terms of their advantages. Pickle files are a simple and fast method to save and transfer complex objects, making them excellent for exchanging trained models, preprocessed datasets, or any other Python objects that must be retained in their entirety.

Pickle files also support object persistence over sessions or settings, ensuring that an object's state may be safely restored at a later time.

While pickle files offer flexibility and simplicity of use, they should be loaded with caution from untrusted sources. Untrusted pickle files may contain malicious code that compromises the security of a system. As a result, always load pickle files from reliable sources or verify their integrity before using them.

ngrok software:

Ngrok eliminates the requirement for complex network setups and application deployment to a public server. It has a simple command-line interface that creates a secure tunnel to your local server, thereby making it available over the internet. This is especially handy for development and testing since it allows you to share your work with clients without requiring a complicated infrastructure setup.

Ngrok's ability to produce a unique URL for each tunnel, which remains persistent until expressly canceled, is one of its primary characteristics. This implies that you may give anyone a public URL and they will be able to easily visit your locally hosted application or website. Ngrok also provides various subdomain options, allowing you to customize the generated URL to some extent. In this project also used ngrok software to share RECOSYS webpage to other devices that provide links to users and from that they will access it through mobile or laptop.

6. SUBJECTIVE ANALYSIS:

Subjective analysis is a method used to check how a system works, which means in which we take people’s opinions about a product or system and on that basis decide whether the product/ system is good or bad. So for RECOSYS also subjective analysis is performed and their result is as follows:

So to take reviews or their opinion, created one Google form and we create one questionary that helps to collect the reviews. Generally, the Google form is used for Surveys and Questionnaires, Event Registrations, Feedback Suggestions and for Quizzes, and Assessments so here we provide some questions like these: -

Q.1 Your Name?

Q.2 Profession?

Q.3 Selected Cuisine

Q.4 How much do you like to rate RECOSYS in stars?

Q.5 Would you like to use this system in the future?

Q.6 Do you have any suggestions to improve the System?

Like these questions are provided to the user also for some questions options are given in the form of buttons for easy use of the user the main reason for using Google form is that helps to create surveys to collect feedback, opinions, or data from a group of respondents. You can create different types of questions, such as multiple-choice, checkboxes, text input, rating scales, etc.

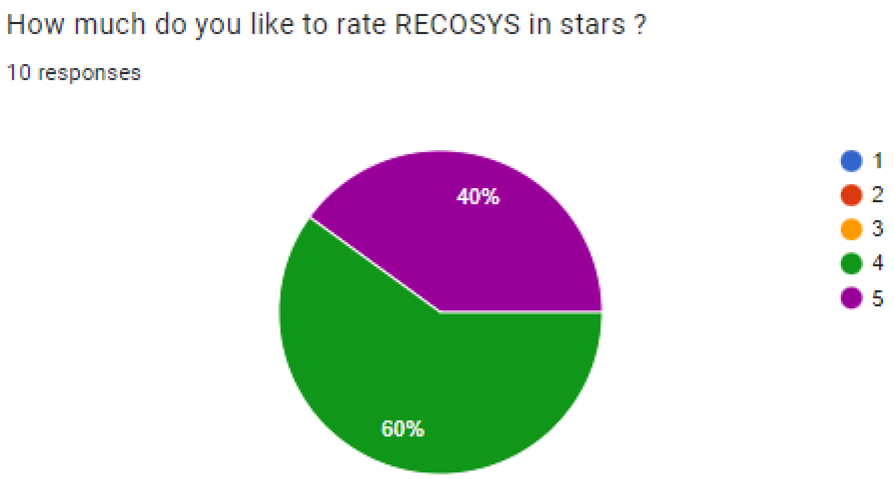
We have taken a total of Pune's local 10 people's responses from that 2 were teachers, 6 students, and 2 other local people who have given their opinion. They had chosen different types of cuisines like Café, Rajasthani, Korean, Biryani, Chinese restaurant, Family Restaurants, Bakery, French, etc

7. DISCUSSION:

Visualizing this distribution in a pie chart, the chart clearly illustrates the breakdown of ratings. The larger portion of the pie, representing 60% of the reviews, is attributed to the 4-star rating. This indicates a positive sentiment among the majority of reviewers toward the RECOSYS system. On the other hand, a significant portion, accounting for 40% of the reviews, bestowed the system with a perfect 5-star rating, which signifies a notably high level of satisfaction with the RECOSYS system.

These findings suggest that the RECOSYS system has received favorable feedback overall. The majority of reviewers were quite pleased with the system, as evidenced by the 4-star ratings. Furthermore, a substantial number of participants found the system to be exceptional, granting it a perfect 5-star rating. This positive sentiment is valuable for the RECOSYS system, indicating its effectiveness and user satisfaction.

It is important to carefully consider both the positive aspects and potential areas for improvement revealed by the review ratings. By taking into account the opinions and feedback provided by reviewers, further enhancements can be made to the RECOSYS system, ensuring an even better user experience and potentially attracting more positive ratings in the future.



How much do you like to rate RECOSYS in stars ?

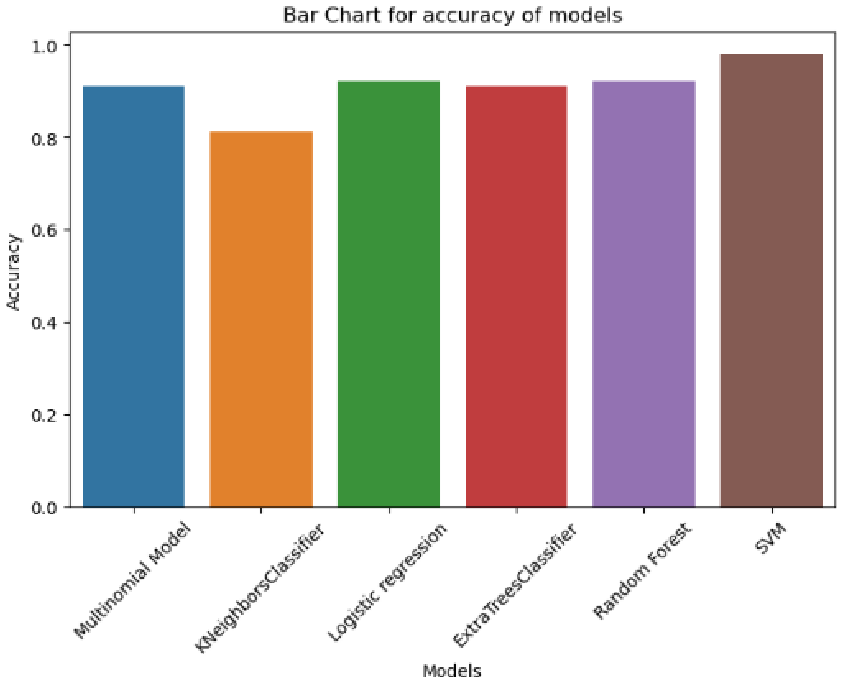
10 responses

7.0.1 Google Form Result

8. RESULTS:

8.1 Model Accuracy:

We applied Six models here from that we got the highest accuracy for SVM which is around 98% then random forest, ExtraTreesclassifier, and Logistic regression which is around 93%, and then for the Multinomial model, we get an accuracy of around 91% also for KNN we get accuracy as 81% all models comparison graph is shown in 8.0.1. From the accuracy and confusion matrix, we can say our model predicts right classes.



8.0.1 Model Accuracy Bar plot

Sr. No.	Model	Precision	Recall	F-1 score	Training Accuracy	Testing Accuracy
1	Multinomial Model	0.92	0.92	0.92	93.54%	91.73%
2	KNN	0.86	0.82	0.82	83.15%	81.52%
3	Random Forest Classifier	0.93	0.93	0.93	94.51%	92.99%
4	ExtraTreesClassifier	1.0	1.0	1.0	92.86%	91.58%
5	Logistic regression	0.99	0.99	0.99	94.51%	92.99%
6	SVM	0.98	0.98	0.98	99.26%	98.37%

8.0.2 Model Accuracy Table

9. CONCLUSION:

So from the above result, we can conclude at this stage, the system looks promising. From subjective analysis also we get positive responses so users like this system and in the future, they will be used this system. The main difference between RECOSYS and Google Maps is, Maps will not give recommendations by rating they give suggestions by distance wise so we have to find top rating restaurants manually whereas RECOSYS gives top restaurants ratings on the basis of sentiment analysis also it provides their distance with an address that really helpful for the user. From the sentimental analysis and recommendation system, we get appropriate results.

Finally, we can say the RECOSYS project successfully implemented a restaurant recommendation system for Pune City. By scraping and analyzing restaurant reviews, we generated sentimental analysis ratings and provided personalized recommendations based on user preferences. The project's frontend interface facilitated seamless user interaction. The evaluation of different models showcased the effectiveness of SVM, Random Forests, ExtraTreesClassifier, and Logistic Regression for sentiment analysis. This project demonstrates the potential of sentiment analysis and recommendation systems in improving user experiences in the restaurant domain.

10. FUTURE WORK:

The Main future work is this system is useful only for Pune city so provide it all over India country, we can add multiple features aspects like intricate average price per meal, and support of multiple locations in data so users also take advantage. and for that, we can add filters. and also work on distance here it gives direct distance so try to calculate navigated distance. We take static data for dynamic data we have to work on it.

REFERENCES:

[1] Goonjan Jain and Neha Punetha, "Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews.", 2023.

[2] Paulo Rita, Celeste Vong, Flavio Pinheiro, and Joao Mimoso, "A sentiment analysis of Michelin-starred restaurants", 2022.

[3] Leen Muteb Alharbi and Ali Mustafa Qamar, "Arabic Sentiment Analysis of Eateries' Reviews Using Deep Learning",2022.

[4] Jing Tian, Wushour Slamu Miaomiao Xu, Chunbo Xu, and Xue Wang, "Research on Aspect-Level Sentiment Analysis Based on Text Comments",2022.

[5] Hengyun Li a, Bruce X.B. Yu b, Gang Li, and Huicai Gao, "Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews",2023.

[6] Ahmad Adel Abu-Shareha, Qusai Y. Shambour, and Mosleh M. Abualhaj, "Restaurant Recommendations Based on Multi-Criteria Recommendation Algorithm",2022.

[7] Murtadha M. Hamad, Mohanad Ahmed Salih, and Refed Adnan Jaleel, " Sentiment Analysis of Restaurant Reviews in Social Media using Naïve Bayes ",2021.

[8] lham Asani, Hamed Vahdat-Nejad, and Javad Sadri, "Restaurant recommender system based on sentimental analysis",2021.

[9] Parandham G and Mr. Raghavendra R, "Sentimental Analysis On Restaurant Reviews.",2021.

[10] Sunmin Lee, "SENTIMENT ANALYSIS USING BERT ON YELP RESTAURANT REVIEWS",2022.

Proforma 4

Undertaking from the PG student while submitting his final dissertation to his respective institute

Ref. No.

I, the following student

Sr. No.	The sequence of student's names on a dissertation	Students name	Name of the Institute & Place	Email & Mobile
1.	First Author	Sourav Khot	SIG	Email: 22070243027@sig.ac.in Mobile: 9892700193

Note: Put additional rows in case of more students

hereby give an undertaking that the dissertation **RECOSYS (REstaurant reCOmmendation SYStem)** been checked for its Similarity Index/Plagiarism through the Turnitin software tool; and that the document has been prepared by me and it is my original work and free of any plagiarism. It was found that:

1.	The Similarity Index (SI) was: *(Note: SI range: 0 to 10%; if SI is >10%, then authors cannot communicate ms; **attachment of SI report is mandatory**)*	7 %
2.	The ethical clearance for research work conducted was obtained from: *(Note: Name the consent obtaining body; if 'not appliable' then write so)*	NA
3.	The source of funding for research was: *(Note: Name the funding agency; or write 'self' if no funding source is involved)*	Self
4.	Conflict of interest: *(Note: Tick √ whichever is applicable)*	No

5.	The material (adopted text, tables, figures, graphs, etc.) as has been obtained from other sources, has been duly acknowledged in the manuscript: *(Note: Tick √ whichever is applicable)*	Yes

In case any of the above-furnished information is found false at any point in time, then the University authorities can take action as deemed fit against all of us.

Sourav Balasaheb Khot

Full Name &
Signature of the student

Dr Rajesh Dhumal

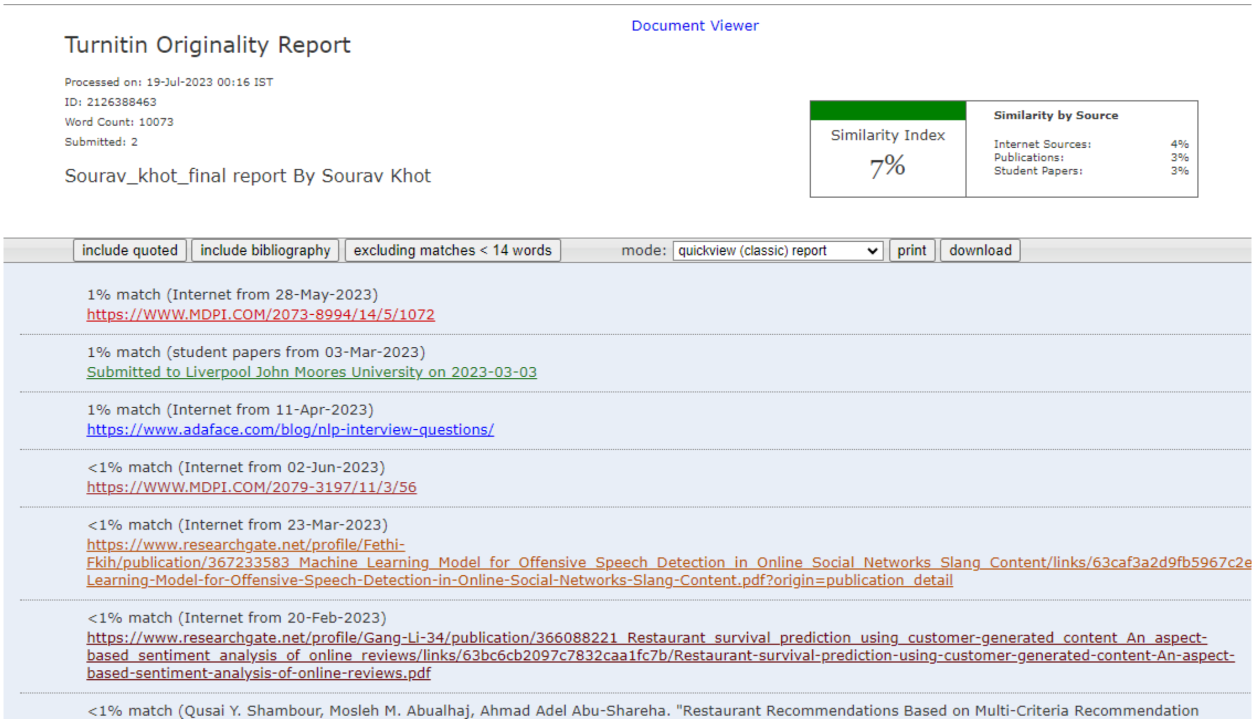
Name &
Signature of SIU Guide/Mentor

Date: 4 July 2023

Endorsement by
Academic Integrity Committee (AIC)

Place: Pune

Note: It is mandatory that the Similarity Index report of plagiarism (only the first page) should be appended to the UG/PG dissertation



Turnitin Originality Report

Processed on: 19-Jul-2023 00:16 IST
ID: 2126388463
Word Count: 10073
Submitted: 2

Sourav_khot_final report By Sourav Khot

Similarity Index
7%

Similarity by Source

Internet Sources: 4%
Publications: 3%
Student Papers: 3%

mode: quickview (classic) report

1% match (Internet from 28-May-2023)
https://WWW.MDPI.COM/2073-8994/14/5/1072

1% match (student papers from 03-Mar-2023)
Submitted to Liverpool John Moores University on 2023-03-03

1% match (Internet from 11-Apr-2023)
https://www.adaface.com/blog/nlp-interview-questions/

<1% match (Internet from 02-Jun-2023)
https://WWW.MDPI.COM/2079-3197/11/3/56

<1% match (Internet from 23-Mar-2023)
https://www.researchgate.net/profile/Fethi-Fkih/publication/367233583_Machine_Learning_Model_for_Offensive_Speech_Detection_in_Online_Social_Networks_Slang_Content/links/63caf3a2d9fb5967c2e
Learning-Model-for-Offensive-Speech-Detection-in-Online-Social-Networks-Slang-Content.pdf?origin=publication_detail

<1% match (Internet from 20-Feb-2023)
https://www.researchgate.net/profile/Gang-Li-34/publication/366088221_Restaurant_survival_prediction_using_customer-generated_content_An_aspect-based_sentiment_analysis_of_online_reviews/links/63bc6cb2097c7832caa1fc7b/Restaurant-survival-prediction-using-customer-generated-content-An-aspect-based-sentiment-analysis-of-online-reviews.pdf

<1% match (Qusai Y. Shambour, Mosleh M. Abualhaj, Ahmad Adel Abu-Shareha. "Restaurant Recommendations Based on Multi-Criteria Recommendation