Sentiment-Driven Personalized Restaurant Recommendations: A DistilBert Approach

A Project Report

submitted in fulfillment of the
requirements for the award of the degree
of

Bachelor of Technology

in
COMPUTER SCIENCE & ENGINEERING
by

Shilpa Sahu

Roll No. 20115091

Under the guidance of

Dr. Dilip Singh Sisodia

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY RAIPUR
RAIPUR, CG (INDIA)

DECLARATION

We hereby declare that the work described in this report, entitled "**City & Cuisine-Based Restaurant Recommendation System**" which is being submitted by us in fulfillment of the award of the degree of Bachelor of Technology in Computer Science & Engineering to the Department of CSE, National Institute of Technology Raipur is the result of investigations carried out by us under the guidance of Dr. Dilip Singh Sisodia.

The work is original and has not been submitted for any Degree/Diploma of this or any other Institute/university.

Shilpa Sahu

Roll No.: 20115091

Place: RAIPUR

Date:

ACKNOWLEDGEMENT

I would like to acknowledge my college **National Institute of Technology, Raipur** for providing a holistic environment that nurtures creativity and research-based activities.

I express our sincere thanks to **Dr. Dilip Singh Sisodia**, Associate Professor CSE Department, NIT Raipur, the supervisor of the project for guiding and correcting throughout the process with attention and care. He has frequently suggested creative ideas and guided through the major hurdles that occurred during the duration of the project.

I would also thank **Dr. Pradeep Singh**, Head of the department and all the faculty members without whom this project would be a distant reality. I also extend my heartfelt thanks to my family and friends who supported me.

Thank You

Shilpa Sahu

Roll No. 20115091

CERTIFICATE

This is to certify that the project entitled "City & Cuisine-Based Restaurant Recommendation System", that is being submitted by Shilpa

Sahu(Roll No. 20115091), in fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science & Engineering to National Institute of Technology Raipur is a record of bonafide work carried out by them under my guidance and supervision.

The matter presented in this project document has not been submitted by them for the award of any other degree/diploma elsewhere.

Dr. Dilip Singh Sisodia

Associate Professor
Department of Computer Science & Engineering
National Institute of Technology, Raipur (CG.)

H.O.D

Dr. Pradeep Singh

Department of C.S.E
NIT Raipur (CG.)

ABSTRACT

Recommender systems are used to provide users with information and services.Opinions are important in recommendation systems because they affect the success or failure of applications, restaurants, and other tech-driven services in today's world. Positive or negative reviews can

have a big impact on user engagement. Analyzing these reviews is essential to understand user experiences and enhance service quality.
In traditional recommendation system star ratings provide a basic overview, while delving into text-based reviews offers valuable insights into the reasons behind the ratings. However, manually processing numerous reviews is impractical. Leveraging machine learning algorithms for sentiment analysis becomes a pragmatic solution.
Various methods, such as TF-IDF and Word2Vec, have been employed for sentiment analysis. However, these methods faced challenges in comprehending emotions and the genuine meaning embedded in sentences. Moreover, earlier methods lacked personalization, making it crucial to evolve the approach.This work employs the DistilBERT model to analyze sentiments. The method integrates two datasets: a business dataset with details about shops, addresses, and features, and a review dataset containing user-generated review texts. Through the combination of these datasets, the framework aims to provide personalized recommen dations considering both city and cuisine preferences. The proposed method combines Sentiment Analysis and Recommendation Systems to provide users with tailored suggestions. Sentiment Analysis is used to classify the reviews of restaurants based on the text. This method addresses limitations of previous methods and improves the overall user experience. The results from Sentiment Analysis are then sent to a recommendation system, which generates a list of the top-n restaurants for the user. This approach has shown better results compared to existing recommendation process. Consequently, the proposed system enhances the accuracy of recommended items by examining the sentiment expressed in user reviews. The implementation of above proposed model has yielded promising results, with an accuracy of 83%. Additionally, the Mean Squared Error (MSE) stands at 0.34, the Root Mean Squared Error (RMSE) at 0.58, and the Mean Absolute Error (MAE) at 0.45. These metrics collectively indicate a high level of precision and reliability in the model's predictions.

CONTENT

Title	Page No.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

With the exponential growth of social media platforms, the amount of data generated has increased tremendously. In a conventional recommendation system, star ratings provide a summary of the product, but analysing text-based reviews provides insightful information about the factors that influence the ratings. This vast amount of unstructured and heterogeneous data can be overwhelming for users and make it difficult to find relevant information. This project aims to address this challenge by developing a personalized, fine-grained user preference-oriented framework that utilizes user reviews, comments, and sentiment analysis to extract relevant information and provide personalized recommendations.

1.2 Project Objectives and Significance

Developing a framework that can efficiently extract relevant details from social media data and offer tailored recommendations based on user preferences is the main goal of this project. This framework aims to:

- Predict the rating of restaurants listed in the Yelp dataset [14] based on analyzing the sentiment of review text given by the user to a particular restaurant. Classification techniques are used.

- Improve user experience by offering graphical User Interface, in particular, requires two inputs from users or customers in order to predict the top restaurants in a given city for a specific cuisine that the user provides.

- Determine which services the restaurant offers, including price range, bike parking, and business card acceptance.

- Reduce information overload by filtering out irrelevant recommendations.

1.3 Scope

Our focus is on developing a framework that carefully analyzes user reviews and comments on restaurants to provide personalized city and cuisine-based recommendations. We leverage the power of the transfer learning model.

To achieve this, we will merge two datasets: a business dataset containing information about restaurants and a review dataset containing user reviews and comments. This combined dataset will be preprocessed, tokenized, and subjected to sentiment analysis using the classification model. The extracted information will then be used to generate personalized recommendations for each user.

1.4 Motivation

Social media has become an integral part of our daily lives, yet it often presents a challenge in extracting relevant information and providing personalized recommendations amidst the vast amount of data generated. Our project aims to bridge this gap by developing a framework that empowers users to navigate the social media landscape with ease, tailored recommendations, and a delightful user experience. We believe that everyone should have access to a personalized recommendation system that caters to their unique dining preferences, making every restaurant search a breeze.

CHAPTER 2
LITERATURE REVIEW

2.1 Existing work

Research on sentiment analysis in restaurant reviews has been extensive. In [1] ,author examine how sentiment analysis techniques are used to determine customers' sentiments on several aspects of dining experiences, like meal quality, service, and ambience.

on fact, texting remains one of the most popular ways to communicate on social networks, despite the availability of other channels. The goal of the work described is to identify, analyse, and utilise the sentiment and emotion that people express through text in their Twitter messages in order to generate recommendations. In [2] ,author gathered tweets and responses on a few specific subjects and created a dataset containing text, user, emotion, sentiment, and other data.

In [3], authors employ cosine similarity and term frequency-inverse document frequency (TF-IDF) and Cosine Similarity to offer other hotel selections based on their reviews. The weight value of terms or documents is found using TF-IDF, and comparable types of values are extracted from the sample set using cosine similarity. Customers are frequently observed to display inconsistent rating behaviour, which results in less accurate preferences being gathered for the recommendation task. In order to alleviate this issue,in [4] authors take into consideration using the sentiment data gathered from user-posted comments on the channels as a stand-in for user ratings. in [4],authors experimented with different sentiment analysis classifiers, such as the responsive neural network-based sentiment analysis, to assess the effectiveness of substituting sentiment data for user ratings.

[5] research dining establishment recommendation systems for patron preference and services according to amenities and rating. The proposed sentiment score measure natural language processing technology is utilized to determine the sentiments and perspectives of user comments. Instead, it is much more practical to train an algorithm to perform this task, and advancements in machine learning make this possible. Many machine learning algorithms, such as Random Forest Classifier, Multinomial Naïve Bayes, and Bernoulli's Naïve Bayes, have been analyzed, and their behavior has been examined in [6].

The study of feelings towards an object or entity is known as sentiment analysis (SA) or opinion mining [7]. sentiment analysis Classification problems can be addressed; sentiment analysis will determine whether the sentiment expresses a positive or negative opinion [7][8].Product reviews are the most significant use of sentiment analysis; these reviews are crucial for business owners because they allow them to make decisions based on customer feedback, as well as for users because they can be advised for products based on the feedback left by other customers [7].
Numerous researchers have attempted to address the issue of personalized social media research using a variety of techniques. [8] gave a thorough overview of the work being done in the field of social media sentiment analysis. Improvements in a number of recently proposed algorithms and sentiment analysis (SA) applications were thoroughly examined and presented in this survey effort. Many strategies, including non-machine learning-based techniques and machine learning-based techniques (probabilistic classifiers, such as Naive Bayes, Maximum Entropy and Linear, Support Vector Machines (SVM), and neural networks) were looked into.

Developing a recommendation system is an interesting concept, which includes techniques that combine the short text content with existing knowledge, based on the sentiments and opinions that are now available. [10] Investigate a novel use of Recursive Neural Networks (RNN) in conjunction with deep learning systems for sentiment analysis of interviews. provide a deep learning model to process user feedback and generate a potential user rating for user recommendations [11]. Ultimately, data learning for the recommendations is achieved by a deep belief network and sentiment analysis (DBNSA). A variety of research has been actively conducted on sentiment analysis techniques such as an approach using word frequency or morphological analysis, and the method of using a complex neural network. Study of Convolutional Neural Networks and Recurrent Neural Networks is done to find out if deep learning algorithms perform better. Numerous studies have been actively carried out on sentiment analysis techniques, including the use of a

complex neural network and word frequency or morphological analysis. Convolutional and recurrent neural networks are studied to determine whether deep learning algorithms work better. [12] Provide a four-step procedure for users to recommend the best book. The levels are referred to as collections of related sentences by the semantic network, sentiment analysis (SA), recommendation system, and reviewers. A two-step efficient resource recommendation model is introduced in order to raise the level of learning resource efficiency [13].The model relies on unsupervised deep learning machines to identify learning styles and user examples, as well as a sentiment analysis bonus system based on user experience to enhance or modify the recommendation system's list classification of items.

2.2 Summary of Existing work

S. No.	Research paper title	Method Used	Remarks
1.	Sentiment Analysis of Restaurant Reviews [1]	SVM, Naïve Bayesian Classifier algorithm	The SVM method is the most effective in utilising the hidden relationships between parameters that the ladder method is unable to identify. This is due to the fact that Naïve Bayesian solves the problem using a probabilistic approach, whereas SVM solves the problem geometrically.
2.	Emotion and sentiment analysis from Twitter text [2]	Naïve Bayes classifier	The examination of various emotions and sentiments revealed some intriguing human traits.
3.	A Sentiment analysis-based hotel recommendation using TF-IDF Approach [3]	TF-IDF and Cosine Similarity	The TF-IDF helps determine the weight value of terms or frequency of documents, and the cosine similarity helps extract comparable types of values from the sentiment data set.
4.	Comparison of Sentiment Analysis and User	Matrix factorization and GeoSoCa multi	The findings indicate that a basic binary rating

	Ratings in Venue Recommendation [4]	feature-based venue recommendation models	consisting of two options—"like" and "dislike"—is a sufficient replacement for the multi-rating scales currently in use for video recommendation in location-based social networks.
5.	Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities [5]	NLP[26] algorithm	When compared to existing algorithms, the NLP[26] algorithm performs better.
6.	A Machine Learning Approach to Building a Tourism Recommendation System using Sentiment Analysis [6]	Naive Bayes and Random Forest classifier	In general, LSTM RNNs perform better than traditional ones. RNNs for relationship learning.
7.	Application of Deep Learning to Sentiment Analysis for recommender system on cloud [10]	RNN-based Deep-learning Sentiment Analysis	In-depth education sentiment analysis (RDSA) helps to identify a specific position in accordance with the user's needs by improving the accuracy of the senstiment analysis, which in turn yields better recommendations to the user.
8.	User Rating Classification via Deep Belief Network Learning and Sentiment Analysis [11]	Deep Belief Network and Sentiment Analysis (DBNSA)	The DBSNA method outperforms conventional methods in classification accuracy by employing a discriminative loss function to predict user ratings based on a late representation derived from user comments.

9.	Design of Book Recommendation System Using Sentiment Analysis [12]	Convolution Neural Network (CNN), K-nearest neighbor (KNN)	The system employs a hybrid methodology that blends collaborative filtering techniques with content-based filtering techniques. sentiment analysis is used to extract the sentiment associated with each book from user reviews.
10	Recommender E-Learning platform using sentiment analysis aggregation [13]	Unsupervised deep learning	Proposed a two-step efficient resource recommendation model for online learning platforms. It does this by combining teacher involvement, sentiment analysis, and unsupervised deep learning.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Dataset Description

The information used in this project is a component of the Yelp Data Challenge [14]. The dataset [14] comprises a collection of JSON files containing business information, reviews, tips (condensed reviews), user data, and check-ins. The business's objectives include its name, address, opening hours, category, average star rating, number of reviews, and a number of other characteristics like noise level or service policy. A star rating, the review text, the review date, and the total number of votes the review has received are all listed in the review objects. These two types of objectives have been our main focus in this project. The dataset[14] is broken down into six subdatasets, each of which provides a brief description of the data.

The size of the Data [14] is 6.84 Gb including the sub files:

1. Dataset of “Business” (139 MB)
2. Dataset of “Check-In” (50.3 MB)
3. Dataset of “Photo” (34.9 MB)
4. Dataset of “Review” (4.39 GB)
5. Dataset of “Tips” (203 MB)
6. Dataset of “Users” (2.03 GB)

We have concentrated on the following two goals in this project:

- Business objectives include name, address, opening hours, category, average star rating, number of reviews, and several attributes like noise level or services policy.
- A star rating, the review text, the review date, and the total number of votes the review has received are all listed in the review objects.

3.1.1 Attributes of ‘Business.json’

Here's example of how the information presented in a table format:

Table 1. Attributes of “business” dataset[14]

Field	Description	Example
business id		tnhfDv5Il8EaGSXZGiuQGg
name	business name/restaurant name	Garaje
city ,state	address	475 4rd St, san francisco, CA, 94106
latitude,longitude	coordinate	Latitude 37.7817529521, Longitude -122.39612197
stars	stars by user	4.5 stars (rounded to half-stars)
review count	total reviews no.	1198 reviews
is_open	open or closed status	Open(is_open:1)
attributes	services offered by business	Business Parking: - Garage: No -Street: Yes -Validated: No -Lot: No -Valet: No
categories	categories of business type	-Mexican -Burgers -Gastropubs
hours	operating hours of each day of a week	Monday to Friday: 10:00 AM to 9:00 PM Saturday: 10:00 AM to 9:00 PM Sunday: 11:00 AM to 6:00 PM

3.1.2 Attributes of ‘Review.json’

An organized summary of the review data is given in this table, which also includes the review text, date, star rating, unique IDs, and the total number of insightful, humorous, and cool votes that were cast.

Table 2 .Attributes of “review” dataset[14]

Field	Description	Example
Review ID	encrypted review id	zdSx_SD6obEhz9VrW9uAWA
user_id	encrypted user id	Ha3iJu77CxlrFm-vQRs_8g
business_ID	encrypted business id	nhfDv5Il8EaGSXZGiuQGg
stars	rating by user	4
date	date of rating	2016-03-09
text	review text of restaurant	Excellent spot to hang out after work because of the enjoyable atmosphere and reasonable prices. It's quite lively but a little loud. The food is good and the staff is friendly. They have a decent assortment of beverages.
useful votes	number of useful votes received	0
funny Votes	total number of funny votes	0
cool Votes	total number of cool votes	0

3.2 Data Preprocessing

3.2.1 Flowchart:

The Figure 1 provides a visual representation of the preprocessing steps undertaken in this study. The process involves three key stages: Business Dataset Preprocessing, Review Dataset Processing, and the subsequent Merging of Business and Review Datasets. Each of these steps plays a crucial role in preparing and organizing the data for further analysis.

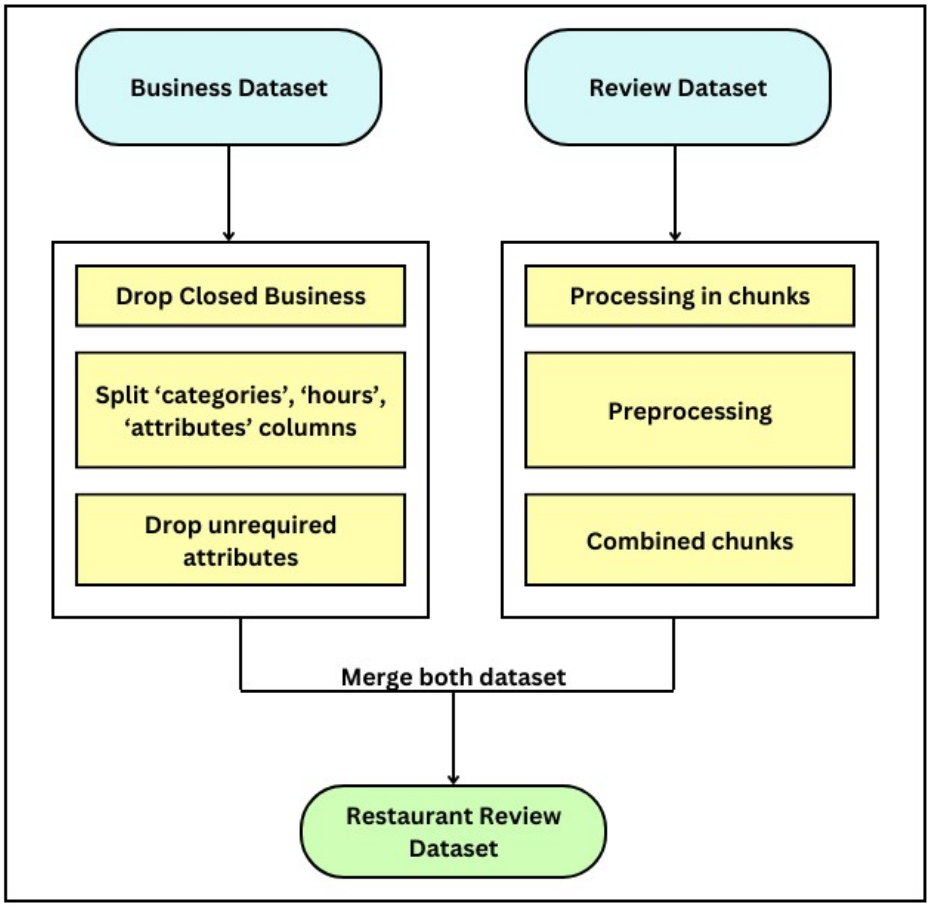

Figure 1. Steps of data preprocessing

Here's step-by-step detailed explanation of the process:

3.2.2 Business Dataset Preprocessing:

3.2.2.1 Drop Closed Businesses

The first step involves filtering out businesses that are currently closed. This is done by selecting only the records where the 'is_open' flag is equal to 1. This ensures that the analysis focuses on actively operating businesses, providing more relevant insights. Figure 2 illustrate the total count of open and closed businesses. We will be dropping closed business here.

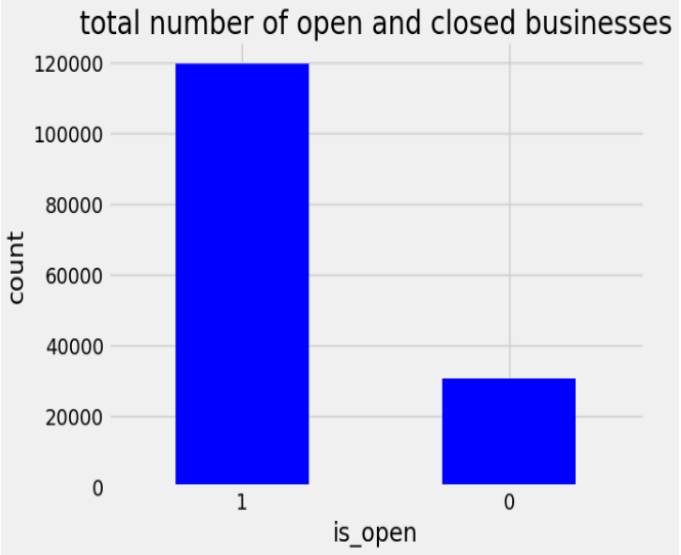

Figure 2 Count of open and closed business

3.2.2.2 Organize Data for Further Analysis

The preprocessing likely aims to organize the data for further analysis or visualization. This includes splitting certain columns into separate columns to make the data more structured and easier to work with.

a) Splitting 'categories' Column

The 'categories' column contains various strings representing the business's categories. To make this information more accessible, the column is split into separate columns, using commas as separators. This allows for easier analysis of the different categories associated with each business.

b) Splitting 'hours' Column

The 'hours' column is in nested form, meaning it contains a dictionary within each record. To extract the hour information, the column is split into separate columns, each representing a specific day of the

week and its corresponding hours of operation. This makes the hour information more organized and easier to analyze.

c) Splitting 'attributes' Column

Similar to the 'hours' column, the 'attributes' column is also in nested form. It contains various business attributes, such as parking options and takeout availability. To make this information more accessible, the column is split into separate columns, each representing a specific attribute.

d) Dropping Redundant Columns

After splitting the 'attributes' column, two redundant columns, 'BikeParking' and 'BusinessParking', are dropped from the resulting DataFrame. These columns contain overlapping information, and keeping them would introduce unnecessary duplication.

3.2.3 Review Dataset Processing:

3.2.3.1 Efficient Chunk Processing

The review dataset is quite large, with approximately 69,90,280 rows. To handle this large amount of data efficiently, the code employs a chunking approach. This involves processing the dataset in smaller chunks, allowing for better memory management and faster processing.

3.2.3.2 Date-Related Processing

The code performs date-related processing on each chunk of the review dataset. This likely involves extracting and formatting date information from the reviews, such as the date of the review or the reviewer's joining date on the platform.

3.2.3.3: Combining Processed Chunks

After processing each chunk of the review dataset, the chunks are combined into a final data frame. This creates a unified data frame containing all the processed review data for further analysis.

3.2.4 Merging Business and Review Datasets

After preprocessing both datasets, the business and review datasets are merged to create a new merged dataset. This merged dataset combines information about businesses with the corresponding reviews, providing a comprehensive view of each business and its associated customer feedback.

3.3 Exploratory Data Analysis

The location (state and city) and business category were the main business features we used in our data analysis. The analysis of the distribution of reviews regarding the business's category and location is included in Figure [3] of the dataset. The frequency distribution of the Category 1 and Category 2 vertex counts are shown in Figures [4] and [5], respectively. These graphical representations provide a visual understanding of the distribution patterns within each category by illustrating the relationship between the categories and their corresponding counts.

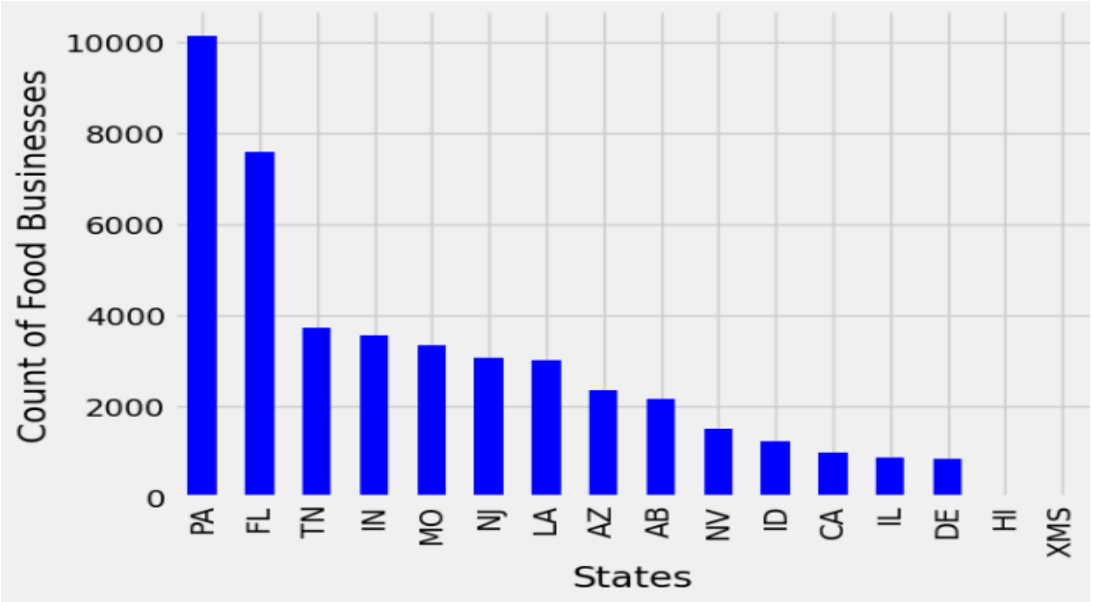

Figure 3 Frequency distribution of State v/s Number of food businesses

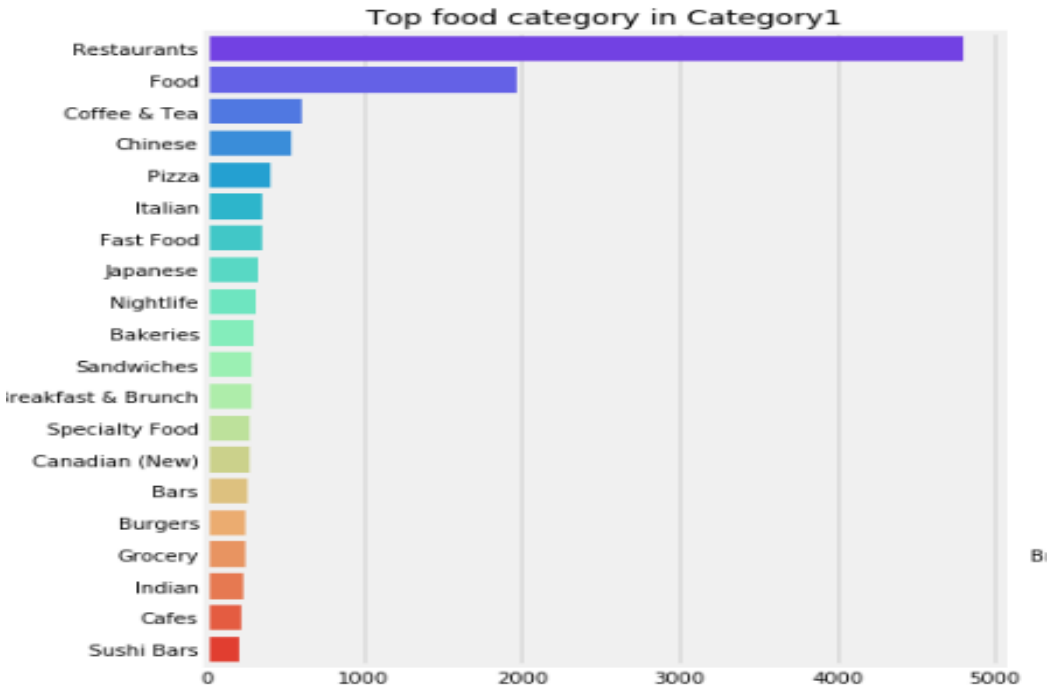

Figure 4 : Frequency distribution of category1 v/s count

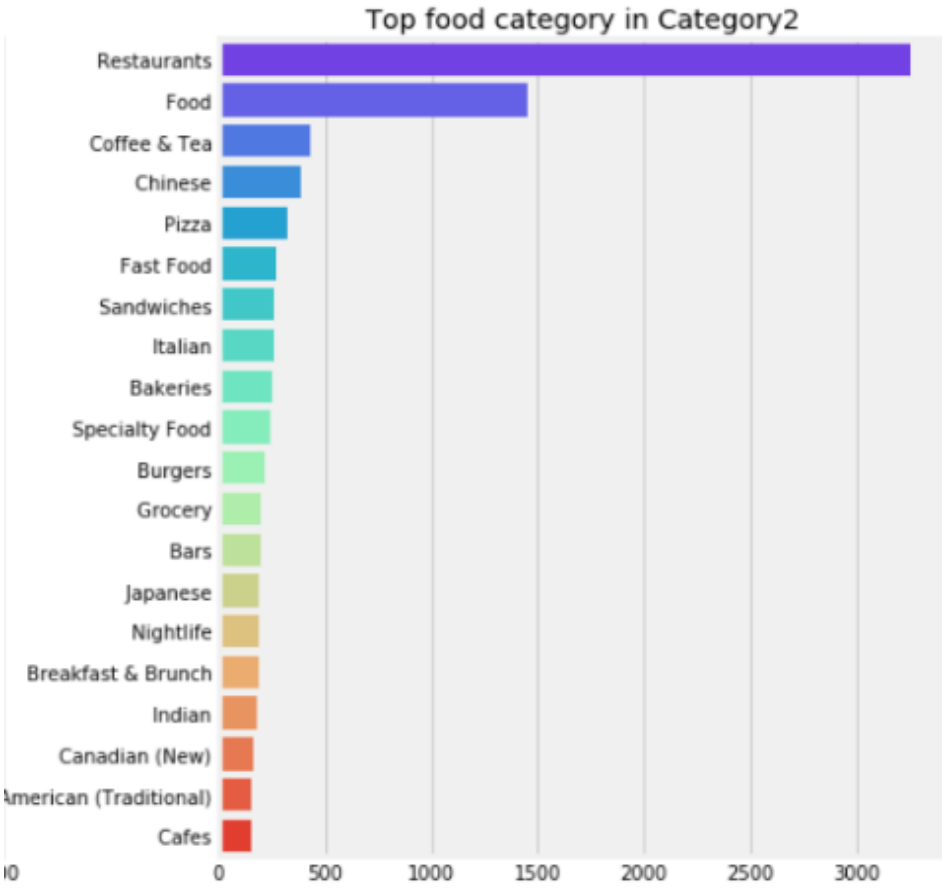

Figure 5 : Frequency distribution of category2 v/s count

3.4 Proposed methodology

3.4.1 Overview

The suggested method incorporates the outcomes of the Sentiment Analysis process. Figure [6] illustrates the primary components and interactions of the proposed system. The Yelp Restaurant Reviews [14] dataset serves as input for the sentiment analyzer, generating a sentiment score as output. This score is then forwarded to the recommender system to generate a top-notch recommendation list.

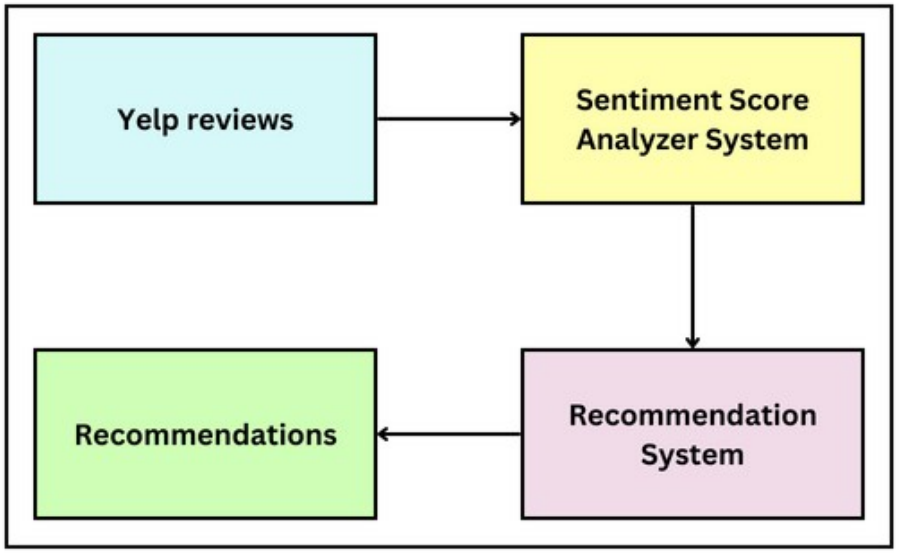

Figure 6. Overall System Architecture

3.4.2 sentiment analysis on Restaurants Reviews

3.4.2.1 Flowchart:

Fig. 7, the visual representation, gives a clear picture of the sequential stages in the sentiment analysis process. The steps involved are as follows: gathering data from the Yelp Restaurant Reviews dataset [14], advancing through filtering based on user preferences, text processing, and calculating sentiment scores using the "Distilbert" model [15], predting ratings based on sentiment scores, and assessing the model's performance using metrics like mean absolute error (MAE) and mean square error (MSE).

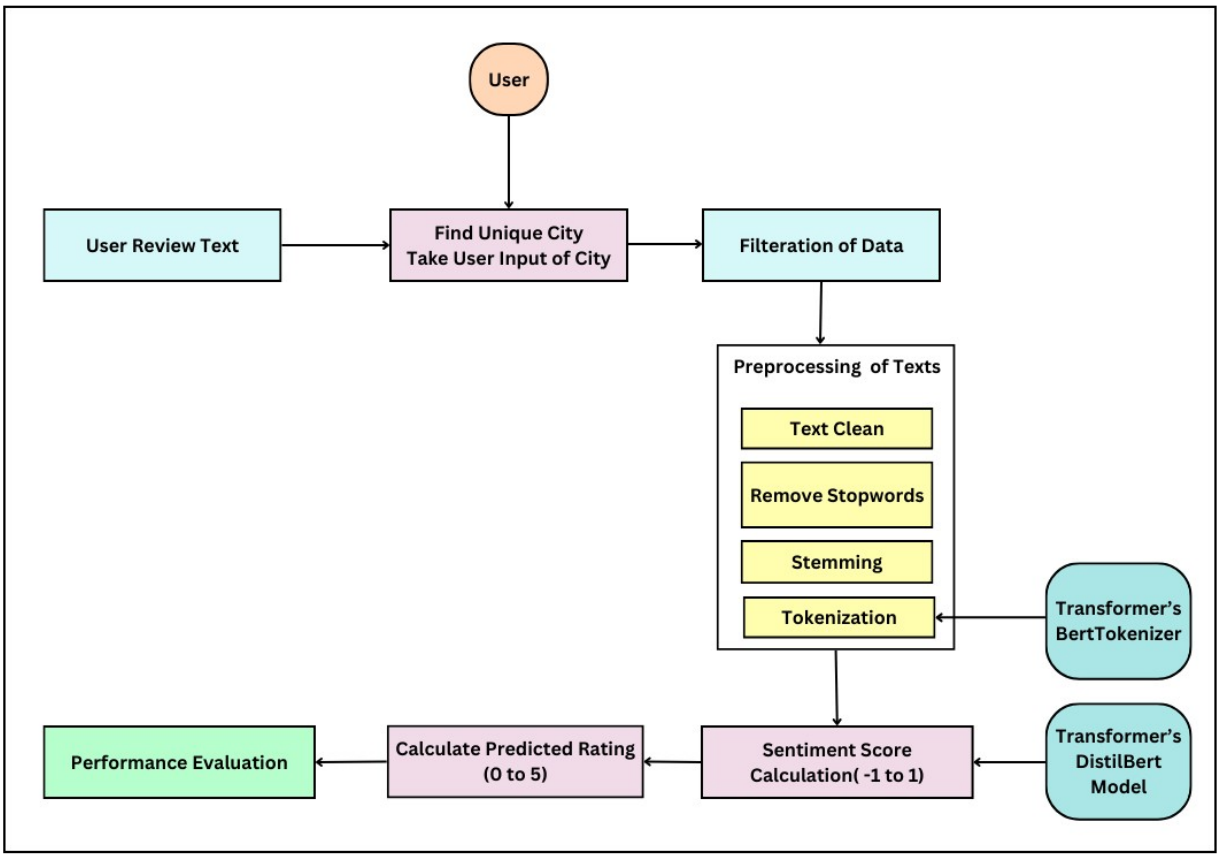

Figure 7 . Steps for sentiment analysis of Review text

1. Data Collection:

Our primary dataset is derived from a preprocessed dataset that includes extensive reviews and business information. We pay attention to the Yelp Restaurant Reviews dataset [14]. This dataset includes important information like the business ID, review date, review ID, user-assigned star ratings (from one to five), unique user IDs, and user-provided veterinarian review text. These data sets are combined to provide a strong basis for our sentiment analysis task, which enables us to comprehend user sentiments towards different businesses.

2. Filtration:

Understanding user preferences is key to tailoring this analyses.Engaging with users involves seeking their input on restaurant reviews for their chosen city of interest. This user-driven filtration process ensures that subsequent analyses are region-specific, providing a more personalized and relevant perspective on restaurant reviews.

3. Preprocessing of Text:

After user-driven filtration, text preprocessing is a vital step to refine and enhance the quality of our review texts achieved. This involves several sub-steps:

a. Text Cleaning:

Carefully reviewing the texts involves removing any odd formatting, superfluous punctuation, and unnecessary whitespace, ensuring that the text data is consistent, noise-free, and ready for analysis.

b. Removal of Stopwords:

Words that are commonly used in a language but usually have little meaning on their own are known as stopwords. Examples include conjunctions (like "and," "but"), articles (like "the," "a"), and common prepositions (like "in," "on").Eliminating stopwords[16] from reviews enables the sentiment analysis model to concentrate on the words that express sentiment and user opinions. When working with limited computational resources, this preprocessing step is especially helpful as it helps to streamline the analysis process.

c. Stemming:

Using the Porter stemming algorithm, we perform word stemming [17], a process that involves removing suffixes to extract the root or stem of each word. This aids in standardizing words, reducing dimensionality, and capturing the core meaning of words.

d. Lowercasing:

For consistency and simplicity in analysis, we convert all characters in the dataset to lowercase. This normalization ensures that there is no distinction between uppercase and lowercase letters, streamlining subsequent analyses.

e. Tokenization:

To digitise the cleaned and processed text, we load the BertTokenizer [18] from the transformers library. Tokenization allows for more in-depth analysis and comprehension of the text by breaking it down into discrete units, or tokens.

4. Sentiment score calculation:

The sentiment score of a review text functions as a numerical indicator, providing a quantitative depiction of the emotional tone contained within the content. Sentiment analysis [26], an advanced natural language processing technique, is used to analyse and interpret sentiments expressed in written or spoken language. The sentiment score, which is typically positioned between positive and negative or measured on a scale from 0 to 1, offers a precise assessment of the emotional landscape present in textual expressions.We have opted to use the powerful "DistilBert" model for generating these sentiment scores [15]. Advanced language comprehension techniques are used in this cutting-edge model to analyse the words, phrases, and contextual nuanced parts of a given text. DistilBert [15] excels at identifying sentiments and categorising them as positive, negative, or neutral after receiving extensive dataset training. The resulting sentiment score allows for a nuanced understanding of user sentiments and attitudes in addition to providing a quantitative measure of the emotional context. In real-world terms, a sentiment score of one indicates a positive and consistent tone, whereas a score closer to zero indicates a more critical or inconsistent sentiment. This sentiment analysis, which is based on the DistilBert model [15], provides businesses, researchers, and analysts with invaluable insights to evaluate customer satisfaction and make data-driven, well-informed decisions based on the emotional ambiguities present in textual data. The implementation of DistilBert [15] enhances the accuracy and scope of sentiment analysis, resulting in a more thorough understanding of user sentiments in a range of contexts, including social media interactions, restaurant reviews, and product reviews.

DistilBert model

Through the use of transfer learning [15] from expansive language models, there has been notable progress in the field of natural language processing (NLP[26]) during the last 18 months. These models, typically grounded in the Transformer architecture introduced by[15]., have consistently demonstrated advancements in various NLP[26] tasks. Notably, there is a prevailing trend of expanding the size of these pre-trained language models and augmenting the scale of

training datasets. We can see in Figure 8. The evolution of these models is evident in their increasing size. The latest model developed by Nvidia features an impressive 8.3 billion parameters, surpassing BERT-large by a factor of 24 and GPT-2 by a factor of 5. In a similar vein, Roberta, Facebook AI's recent accomplishment stands out for being trained on a substantial 160GB corpus of text. This tendency toward larger model sizes and more extensive training data characterizes a broader movement in the field, where researchers and organizations are actively pushing the limits to enhance the capabilities of language models. As these large-scale models consistently set new benchmarks, their impact on the landscape of NLP[26] is becoming increasingly significant, underscoring the transformative the potential of transfer learning in reshaping tasks related to natural language understanding and generation.

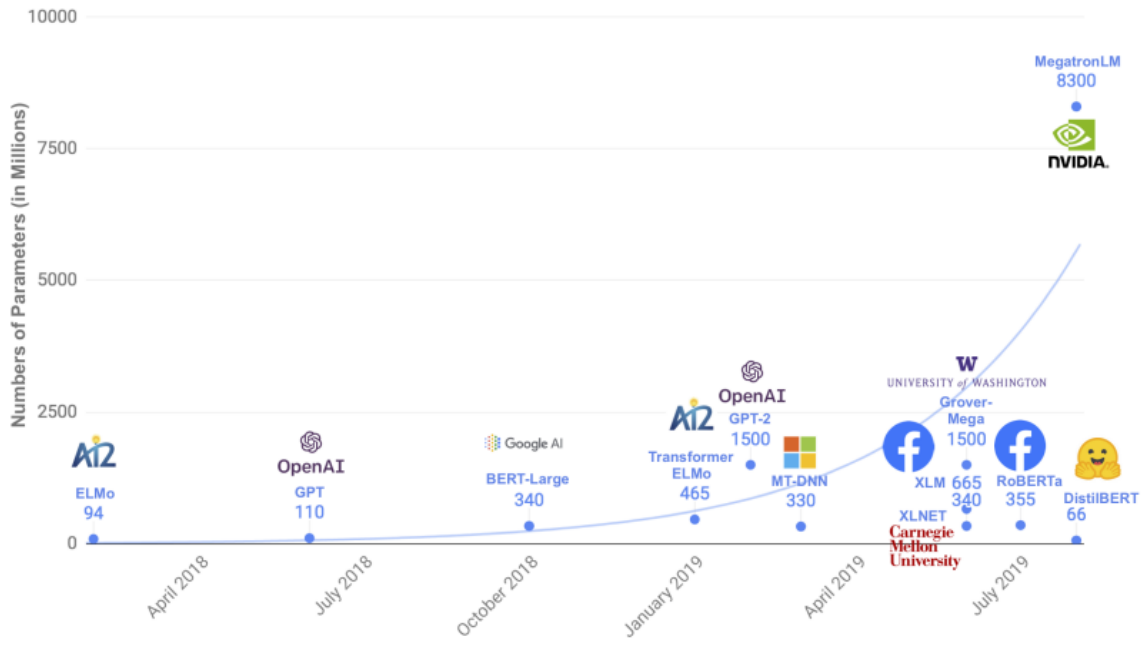

Figure 8. Latest large models and their size in millions of parameters.[15]

Architecture:

Based on the transformer architecture, a type of neural network architecture introduced by [15], Figure 9 depicts the architecture of DistilBERT [15].

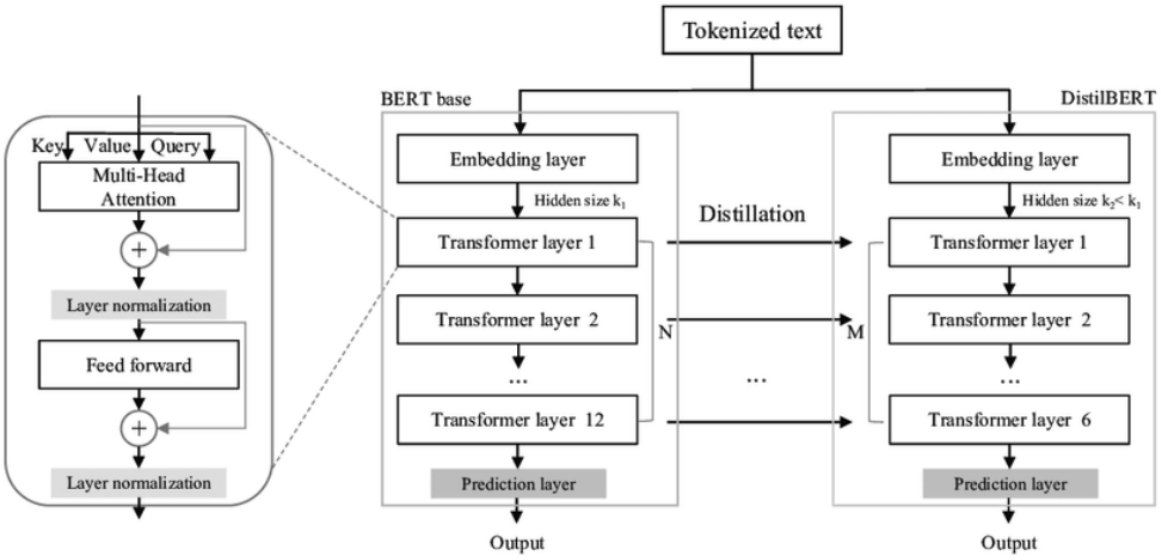

Figure 9. Architecture of DistilBert Model[15]

1. Transformer Architecture:

DistilBert [15], like BERT, uses Transformer architecture that includes encoder operation. These categories involve multiple layers of self-monitoring and neural integration.

2. Encoder Layers:

The structure consists of a group of transformer encoder layers. Each layer has a self-monitoring process that allows the model to consider a number of different sets of inputs when making predictions

3. Attention Mechanism:

DistilBert[15] achieves reduction using various distillation techniques. These include information distillation (where a model is trained to produce output from a larger model) and pruning (which involves removing some connections or weights from the network).

4. Parameter Reduction:

DistilBert[15] uses a variety of distillation techniques to achieve parameter reduction. These included pruning, which entails removing specific connections or weights from the network, and knowledge

distillation, in which a model is trained to imitate the outputs of a larger model.
5. Output Layers:
The final layers of DistilBert[15] generate output representations suitable for the specific NLP[26] task at hand. For example, in a classification task, the model might have output layers for predicting different classes..

Advantages of the DistilBert Architecture

- Reduced Model Size: DistilBert [15] boasts a 40% reduction in parameters compared to BERT, facilitating quicker training and deployment processes.
- Improved Inference Speed: With a 60% increase in speed over BERT, DistilBert [15] is well-suited for real-time applications, enhancing efficiency in processing and generating results promptly.
- Parameter Efficiency: Despite using significantly fewer parameters, DistilBert [15] achieves performance levels comparable to BERT, demonstrating its efficiency in utilizing computational resources.

Applications of DistilBert

The compact size and enhanced speed of DistilBert[15] make it applicable to various Natural Language Processing (NLP[26]) tasks, including:

- Real-time Language Translation: DistilBert[15] proves useful in real-time language translation, facilitating seamless communication across different languages.

- On-device NLP Applications: DistilBert[15] can be deployed on resource-constrained devices, such as mobile devices, for tasks like text summarization and sentiment analysis.

- Large-scale NLP Pipelines: Integration of DistilBERT[15] into extensive NLP pipelines can lead to reduced computational costs and overall improvement in efficiency, making it a practical choice for large-scale language processing tasks

-

Integrating DistilBERT in Our Proposed Model

- The tokenized input is fed into the pre-trained DistilBERT[15] model, and the resulting outputs are obtained.
- Softmax Transformation and Probability Calculation:
  The raw output logits from the model are processed through a softmax function to generate probabilities for each class. The sentiment score is extracted by considering the probability associated with the positive class.
- Mathematical Representation:
  sentiment_score=softmax(logits)[0][1]

5. Predicting Rating Based on Sentiment Score:

In this step, the sentiment score, which ranges from -1 to 1, is transformed into a more interpretable and applicable range of 0 to 5 using mathematical transformations. This conversion ensures alignment with common rating scales and facilitates the interpretation of sentiment in the context of ratings.

Mathematical Transformation:

The sentiment score (S) is mapped to the range [0, 5] using the following transformation:

$$\text{Predicted Rating} = ((S + 1) \div 2) * 5$$

This transformation ensures that a sentiment score of -1 corresponds to a predicted rating of 0, a sentiment score of 0 corresponds to a predicted rating of 2.5, and a sentiment score of 1 corresponds to a predicted rating of 5.

6.Sentiment analysis performance evaluation:

The efficacy of the sentiment analysis model is evaluated using a range of metrics to assess its performance in accurately predicting sentiment. Several often used metrics consist of:

- **MAE:**The Mean Absolute Error, or MAE, measures the precision of the model by computing the average absolute differences between predicted and actual ratings.

  $$MAE=\frac{1}{N}\Sigma\left|y_i-\hat{y}\right|$$

  where,

  $\hat{y}$= predicted rating of calculated
  y=actual rating provided by users

- **MSE:** MSE(Mean Squared Error) is similar to RMSE but without the square root operation. It penalizes larger errors more significantly, offering insights into the model's overall performance.

  $$MSE=\frac{1}{N}\Sigma\left|y_i-\hat{y}\right|^2$$

  where,

  $\hat{y}$= predicted rating of calculated
  y=actual rating provided by users

- **RMSE:** Root Mean Squared Error, or RMSE, quantifies the average magnitude of the differences between predicted and actual ratings. Better accuracy is indicated by lower RMSE values.

  $$RMSE=\sqrt{MSE}=\sqrt{\frac{1}{N}\Sigma\left|y-\hat{y}\right|^2}$$

where,
$\hat{y}$= predicted rating of calculated
y=actual rating provided by users

3.4.3 Restaurant Recommendation System

3.4.3.1 Flowchart

In Figure 10 , different steps for recommending the restaurants are shown:

- Sentiment Score Integration
- Mean Sentiment Polarity Calculation(spm)
- Stars vs. Sentiment Polarity Analysis
- Dataset Refinement
- User cuisine centric Top-N-Recommendation

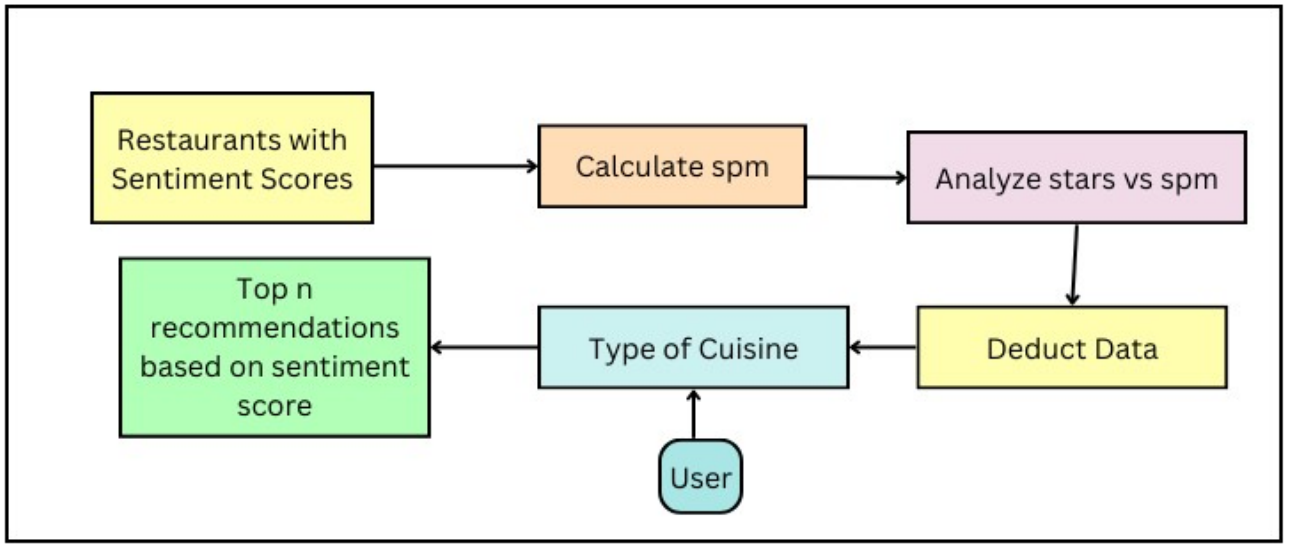

Figure 10. Steps for restaurant recommendation

1. Sentiment Score Integration:

Each restaurant in our dataset is now associated with a sentiment score, reflecting the overall sentiment derived from customer reviews.

2. Mean Sentiment Polarity Calculation(spm):

We calculate the mean sentiment polarity by grouping entries based on their unique business ID. This allows for a comprehensive analysis of sentiment trends across restaurants.

3. Stars vs. Sentiment Polarity Analysis:

We analyze the relationship between star ratings and mean sentiment polarity. Figure 12 illustrates the relationship between stars and spm for values within the range of (1, 1.5). Similarly, Figure 13 depicts the correlation between stars and spm for values falling within the range of (2, 2.5). Moving forward, Figure 14 showcases the stars versus spm dynamics for the range of (3, 3.5). Lastly, Figure 15 provides insights into the interplay between stars and spm for values encompassing the range of (4, 4.5, 5). These figures collectively offer a comprehensive visual representation of the variations in stars and spm across different specified ranges, contributing to a deeper understanding of the underlying data patterns.

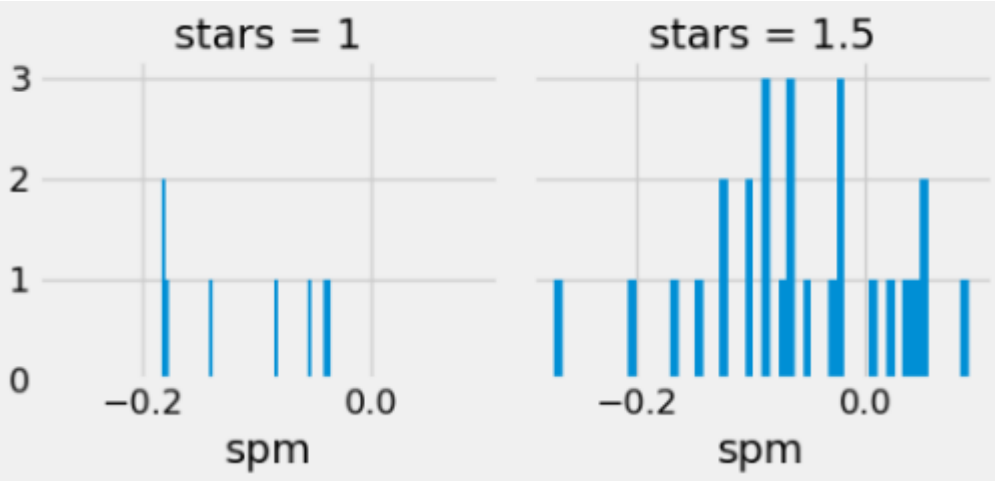

Figure 11. stars vs spm of (1,1.5)

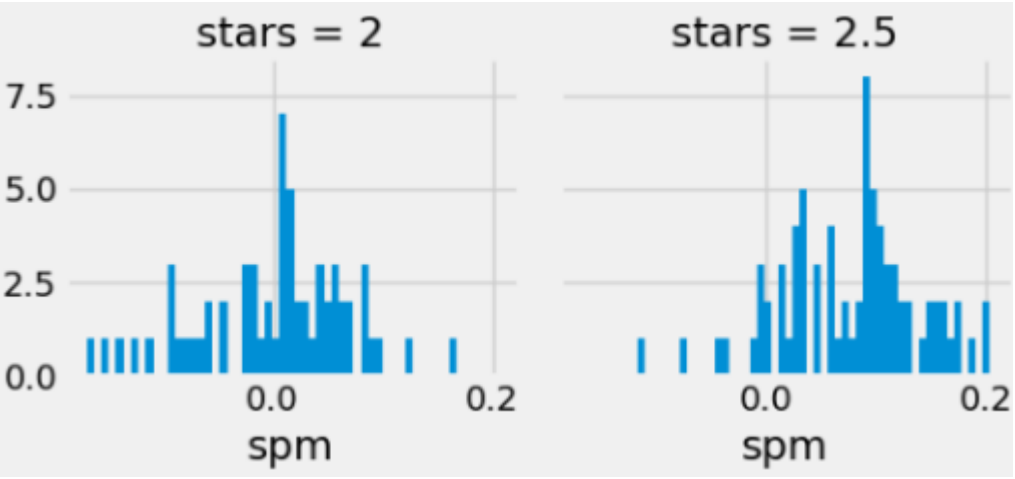

Figure 12. stars vs spm of (2,2.5)

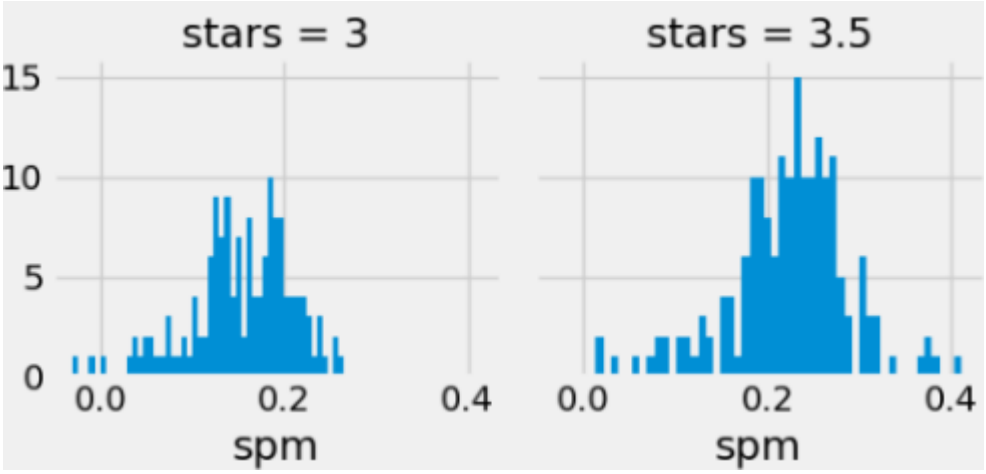

Figure 13 . stars vs spm of (3,3.5)

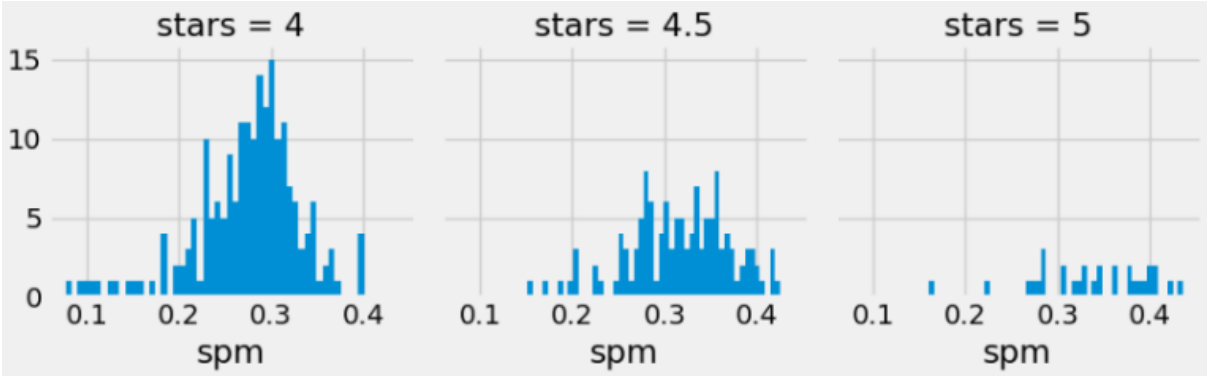

Figure 14 . stars vs spm of (4,4.5,5)

Analysis from the Graph:

- **Star Distribution Above 3.5:**
  The graph analysis reveals that restaurants with star ratings surpassing 3.5 are predominantly found within the range of 0 to 0.5 on the sentiment polarity scale.

- **Refined Dataset Criteria:**
  To narrow our analysis and concentrate on particularly positive sentiments and highly rated establishments, we refine our focus. Specifically, we isolate rows where the star rating exceeds 3.5, indicating notable customer satisfaction, and the sentiment polarity value is greater than 0.

Key Considerations:

- **Positive Sentiments and High Ratings:**
  By implementing these refined criteria, our attention is directed towards establishments that not only boast commendable star ratings but also exhibit positive sentiment in customer reviews.

- **Optimized Recommendations:**
  Through the refinement of our dataset based on the specified criteria, our recommendation system can now provide optimal suggestions. The emphasis is on recommending restaurants that have not only received high star ratings but have also garnered positive sentiments.

- **Elevated User Experience:**
  Users interacting with our recommendation system can anticipate
  a more tailored and positive dining experience. The system's focus on both high star ratings and positive sentiment ensures that
  the recommended establishments align closely with the expectations of a satisfied customer base.

4. Dataset Refinement:

To focus on positive sentiments and higher-rated establishments, we narrow down the dataset. This involves considering only rows where the star rating exceeds 3.5, and the sentiment polarity is greater than 0.

5. User-Centric Top N Recommendation System:

Upon receiving user input specifying a cuisine , our recommendatio system employs a two-tiered approach: city and cuisine. This system recommends top establishments based on cuisine preferences and the specified city.Leveraging city-based information, our recommendation system ensures a personalized experience. Recommendations prioritize establishments with the highest sentiment polarity, contributing to apositive and enjoyable dining experience.

CHAPTER 4
EXPERIMENTAL RESULTS

4.1 System configuration

To assess the DistilBERT [15] model's performance on a range of natural language processing (NLP[26]) tasks, such as summarization, question answering (QA), and natural language understanding (NLU).

1. Transformers (hugging face): A well-known open-source library called Transformers[19] offers a variety of pre-trained language models, such as BERT and BART, which are based on the Transformer architecture. These models perform extremely well in a variety of NLP[26] tasks, such as abstractive text summarization.
Version: 4.34.0

2. TensorFlow: TensorFlow[21] is an excellent open-source end-to-end machine learning platform that makes it easier to train and develop deep learning models. In this project, the neural network architectures for the NLP[26] models were implemented using TensorFlow.
Version: 2.13.0

3. NLTK: A comprehensive Python library for natural language processing (NLP[26]), the Natural Language Toolkit (NLTK)[20]. It provides an extensive array of tools for text processing tasks, such as segmentation by sentence, tokenization, and part-of-speech tagging. These features were employed during the prototyping phase of the NLP[26] pipeline.
Version: 3.8.1

4. Matplotlib: A versatile Python library for creating static, animated, and interactive visualisations is Matplotlib[22]. It offers a variety of plot types and data visualisation tools, making it possible to visualise model performance metrics and insights obtained from the NLP[26] models.
Version: 3.7.3

5. NumPy:
NumPy is a fundamental Python library for scientific computing, developed by NumPy[23]. Effective numerical operations on arrays, matrices, and other multidimensional data structures are offered by it. was employed in a number of NLP[26] pipeline functions, such as performance evaluation, model training, and data processing.
Version: 1.23.5

6. Pandas: A robust Python library for data analysis and manipulation is called Pandas[24]. It offers tabular data processing tools and structures, such as DataFrames, Series, and Index objects. To load, clean, and arrange the datasets used in the NLP[26] experiments, pandas was employed.
Version: 1.5.2

4.2 Experimental setup

The experimental setup begins with cleaning and organizing the data, merging information from business, and reviewing datasets to create a comprehensive dataset. The DistilBERT[15] model is fine-tuned to understand sentiments in the reviews, and for efficiency, a portion of the dataset is strategically reduced while maintaining its representativity. Sentiment scores are then calculated for each review,and these scores are used to predict ratings. The accuracy of these predictions is evaluated by comparing them to actual user ratings. In the final step, personalized restaurant recommendations are generated based on user-provided city information, leveraging sentiment analysis to suggest the top 10 positively reviewed establishments in the specified city. This comprehensive approach aims to offer tailored and positively rated restaurant suggestions aligned with user preferences and location choices.

4.3 Comparison with existing models

In our relentless pursuit of refining sentiment analysis, we embark on a meticulous comparison of distinct models to discern the most effective approach. Commencing with the adoption of the TF-IDF algorithm,enriched by the incorporation of n-grams and tokenization, we endeavor to vectorize our text data. Allocating 15% for testing (X_test and y_test), the system undergoes training on the remaining 85% (X_train and y_train). The employment of a Linear Support Vector Machine (SVM), acknowledged for its proficiency in text classification, yields an accuracy of 59.6%. The implementation of a sentiment analysis model using an LSTM network for predicting star ratings from Yelp reviews involves loading, preprocessing, and splitting the dataset for training and testing. The deep learning model, comprising an embedding layer,SpatialDropout1D, and an LSTM layer, are compiled and trained with early stopping to prevent overfitting. Evaluation using mean absolute error (MAE) demonstrates an accuracy of 70%.. Concurrently, we explore the word2vec tokenizer coupled with Logistic Regression, achieving a respectable accuracy of 66.3%. However, the zenith of our exploration unfolds with the application of DistilBERT[15], where our model excels with a remarkable accuracy of 82%. This pinnacle underscores the supremacy of DistilBERT[15] in discerning sentiments from textual data, showcasing its superior performance when compared to traditional approaches such as TF-IDF, LSTM, and word2vec-based methodologies. This exhaustive head-to-head analysis illuminates the unparalleled effectiveness of DistilBERT[15] in the intricate task of sentiment analysis.

Table 3. Comparison of accuracy and errors

Model	**Tokenizer**	**Accuracy**	**MSE**	**RMSE**	**MAE**
distilbert-base-uncased	DistilBert Tokenizer	0.83	0.34	0.58	0.45
LSTM	keras Tokenizer	.69	.41	0.64	.51
Random forest	word2vec	.66	.69	.83	.45

CHAPTER 5
CONCLUSION & FUTURE WORK

5.1 Conclusion

In the past, methods like TF-IDF and Word2Vec had trouble understanding emotions in text. They struggled with seeing how words related to each other and understanding the real meaning behind sentences. Word2Vec had issues with uncommon words and didn't pay attention to the order of words in a sentence. These problems made them not very good at figuring out how people felt in what they wrote. . In a comparison, models like random forest and LSTM showed decent accuracies, but the game-changer was DistilBERT, achieving an impressive 83% accuracy. User input on city and cuisine preferences serves as a crucial filtration step, tailoring our analyses to specific geographic locations and culinary tastes, amplifying the relevance of our recommendations and catering to diverse user preferences.Strategic result filtering based on user-specified criteria, combined with top polarity values, elevates our recommendation system's effectiveness. This approach ensures recommendations aren't just highly rated but also align with positive sentiments expressed by users, refining the selection process for establishments that consistently evoke positive experiences.The integration of DistilBert, with its streamlined architecture and impressive performance, aligns with our commitment to efficiency and user satisfaction. Its role in extracting sentiment from reviews, paired with user-focused filtering and recommendation strategies, emphasizes our dedication to a tailored and enjoyable dining experience.

In our comprehensive examination of diverse sentiment analysis models, we have observed distinct performances that shed light on the the evolving landscape of natural language processing. Methods such a as the random forest algorithm and LSTM demonstrated respectably accuracies of 66% and 70%, respectively, showcasing their effectiveness in capturing sentiment nuances. However, The implementation of the DistilBERT model has yielded promising results, with an accuracy rate of 83%. Additionally, the Mean Squared Error (MSE) stands at 0.34, the Root Mean Squared Error (RMSE) at 0.58, and the Mean Absolute Error (MAE) at 0.45.

5.2 Future Work

As we think about the future, there are several areas worth exploring in sentiment analysis. Taking a closer look at larger and more diverse datasets could reveal additional insights into how well models can adapt to different languages and subjects. Using multiple criteria to filter information holds promise for improving accuracy and strength. Exploring domain-specific training for models like DistilBERT opens up exciting possibilities for tailoring sentiment analysis to specific industries and applications. Additionally, continuously refining and adapting sentiment analysis models to changes in language patterns and user behaviors is crucial for ensuring their effectiveness in dynamic real-world situations. The field of sentiment analysis calls for further research and innovation, promising ongoing progress in the realm of natural language processing.

REFERENCES

[1] R. Rajasekaran, U. Kanumuri, M. Siddhardha Kumar, S. Ramasubbareddy, and S. Ashok, "sentiment analysis of Restaurant Reviews," in *Smart Intelligent Computing and Applications*, Singapore: Springer Singapore, 2018, pp. 383–390 [Online]. Available: http://dx.doi.org/10.1007/978-981-13-1927-3_41. [Accessed: Nov. 30, 2023]

[2] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *Journal of Computational Science*, vol. 36, p. 101003, Sep. 2019, doi: 10.1016/j.jocs.2019.05.009.

[3] R. K. Mishra, S. Urolagin, and A. A. Jothi J, "A sentiment analysis-based hotel recommendation using TF-IDF Approach," in 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dec. 2019 [Online]. Available: http://dx.doi.org/10.1109/iccike47802.2019.9004385. [Accessed: Nov. 30, 2023]

[4] X. Wang, I. Ounis, and C. Macdonald, "Comparison of sentiment analysis and User Ratings in Venue Recommendation," in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2019, pp. 215–228 [Online]. Available: http://dx.doi.org/10.1007/978-3-030-15712-8_14. [Accessed: Nov. 30, 2023]

[5] R. M. Gomathi, P. Ajitha, G. H. S. Krishna, and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Feb. 2019 [Online]. Available: http://dx.doi.org/10.1109/iccids.2019.8862048. [Accessed: Nov. 30, 2023]

[6] A. Kulkarni, R. M., P. Barve, and A. Phade, "A Machine Learning Approach to Building a Tourism Recommendation System using sentiment analysis," *International Journal of Computer Applications*, vol. 178, no. 19, pp. 48–51, Jun. 2019, doi: 10.5120/ijca2019919031.

[7] R. M. Gomathi, P. Ajitha, G. H. S. Krishna, and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Feb. 2019 [Online]. Available: http://dx.doi.org/10.1109/iccids.2019.8862048. [Accessed: Nov. 30, 2023]

[8] I. Maks and P. Vossen, “A lexicon model for deep sentiment analysis and opinion mining applications,” *Decision Support Systems*, vol. 53, no. 4, pp. 680–688, Nov. 2012, doi: 10.1016/j.dss.2012.05.025.

[9] W. Medhat, A. Hassan, and H. Korashy, “sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[10] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha, and S. Yenduri, “Application of Deep Learning to sentiment analysis for recommender system on cloud,” in *2017 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Jul. 2017 [Online]. Available: http://dx.doi.org/10.1109/cits.2017.8035341. [Accessed: Nov. 30, 2023]

[11] R.-C. Chen and Hendry, “User Rating Classification via Deep Belief Network Learning and sentiment analysis,” *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 535–546, Jun. 2019, doi: 10.1109/tcss.2019.2915543.

[12] A. Mounika and S. Saraswathi, “Design of Book Recommendation System Using sentiment analysis,” in *Evolutionary Computing and Mobile Sustainable Networks*, Singapore: Springer Singapore, 2020, pp. 95–101 [Online]. Available: http://dx.doi.org/10.1007/978-981-15-5258-8_11. [Accessed: Nov. 30, 2023]

[13] J. Mawane, A. Naji, and M. Ramdani, “Recommender E-Learning platform using sentiment analysis aggregation,” in Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, Sep. 2020 [Online]. Available: http://dx.doi.org/10.1145/3419604.3419784. [Accessed: Nov. 30, 2023]

[14] Inc. Yelp, “Yelp Dataset,” *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset.

[15] V. Sanh, “ Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT,” *HuggingFace*, Aug. 31, 2020 [Online]. Available: https://medium.com/huggingface/distilbert-8cf3380435b5. [Accessed: Nov. 30, 2023]

[16] *NLTK :: Search*. (n.d.). Retrieved December 1, 2023, from https://www.nltk.org/search.html?q=stopwords

[17] Contributors to Wikimedia projects. (2023, November 29). *Stemming*. Wikipedia. https://en.wikipedia.org/wiki/Stemming

[18] *text.BertTokenizer*. (n.d.). TensorFlow. Retrieved December 1, 2023, from https://www.tensorflow.org/text/api_docs/python/text/BertTokenizer

[19] Contributors to Wikimedia projects. (2023a, November 26). *Transformer (machine-learning model)*. Wikipedia. https://en.wikipedia.org/wiki/Transformer_(machine-learning_model)

[20] *NLTK :: Search*. (n.d.). Retrieved December 1, 2023, from https://www.nltk.org/search.html?q=stopwords

[21] *TensorFlow*. (n.d.). TensorFlow. Retrieved December 1, 2023, from https://www.tensorflow.org/

[22] *Matplotlib — visualization with python*. (n.d.). Retrieved December 1, 2023, from https://matplotlib.org/

[23] *NumPy*. (n.d.-a). Retrieved December 1, 2023, from https://numpy.org/

[24] *pandas*. (n.d.). Python Data Analysis Library. Retrieved December 1, 2023, from https://pandas.pydata.org/

[25]code link: https://colab.research.google.com/drive/1LhUCMLGMK7ruZ5_02uY9kyjT4wDnEEOV?usp=sharing

[26] Contributors to Wikimedia projects. (2023, November 28). *Natural language processing*. Wikipedia. https://en.wikipedia.org/wiki/Natural_language_processing

[27]Contributors to Wikimedia projects. (2023b, December 5). *sentiment analysis*. Wikipedia. https://en.wikipedia.org/wiki/Sentiment_analysis