JSS MAHAVIDYAPEETHA
JSS SCIENCE AND TECHNOLOGY UNIVERSITY
SRI JAYACHAMARAJENDRA COLLEGE OF ENGINEERING
JSS SCIENCE AND TECHNOLOGY UNIVERSITY MYSURU
- Constituent College of JSS Science and Technology University
- Approved by A.I.C.T.E
- Governed by the Grant-in-Aid Rules of Government of Karnataka
- Identified as lead institution for World Bank Assistance under TEQIP Scheme
- Diamond Jubilee year 1963-2023

YELP REVIEWS CLASSIFICATION USING NLP

Mini project report submitted in partial fulfilment of curriculum prescribed for the Machine Learning (20CS610) course for the award of the degree of

Bachelor of Engineering

In

Computer Science and Engineering

By

Ananya U	01JST21CS017
Lakshmi K M	01JST21CS064
Pruthvi K P	01JCE21CS077
Sadhana K S	01JCE21CS081

Under the Guidance of

Rakshitha R,
Assistant Professor,
Dept. of CSE,
JSSSTU, Mysuru

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

April 2024

JSS MAHAVIDYAPEETHA

JSS SCIENCE AND TECHNOLOGY UNIVERSITY

JSS Technical Institutions Campus, Mysuru – 570006

This is to certify that the work entitled **'Yelp Reviews Classification using NLP'** is a bona fide work carried out by **Ananya U, Lakshmi K M, Pruthvi K P, Sadhana K S** in partial fulfilment of the award of the degree of Bachelor of Engineering in Computer Science and Engineering of JSS Science and Technology, Mysore during the year 2024. It is certified that all corrections/suggestions indicated during CIE have been incorporated in the report. The mini project report has been approved as it satisfies the academic requirements in respect of mini project work prescribed for the **Machine Learning (20CS610)** course.

Course in charge and guide

Rakshitha R,
Assistant Professor,
Dept. of CS&E,
JSSSTU, Mysuru

CONTENTS

INTRODUCTION

In the competitive landscape of the restaurant industry, understanding customer sentiment and feedback is paramount to success. Yelp, as a popular platform for user-generated restaurant reviews, serves as a treasure trove of valuable insights for eateries seeking to improve their offerings and customer experience. Leveraging the power of natural language processing (NLP) and machine learning, this project endeavours to unlock the hidden wealth of information within Yelp reviews to aid restaurants in optimizing their operations and enhancing customer satisfaction.

The primary objective of this project is twofold: to dissect Yelp restaurant reviews to identify key strengths and areas for improvement, and to develop a predictive model capable of forecasting star ratings based on review text. By achieving these objectives, this project aims to provide actionable insights and recommendations to restaurants, empowering them to refine their services, address customer concerns, and ultimately elevate their reputation and competitiveness in the market.

PROBLEM STATEMENT

The aim of this project is to develop a comprehensive analysis framework for Yelp restaurant reviews. This framework leverages natural language processing (NLP) techniques and machine learning methods to empower restaurants to extract valuable insights from customer reviews. By combining these methodologies, the framework offers actionable recommendations and predicts star ratings for restaurants, ultimately helping them improve their customer experience.

OBJECTIVES

1. **Sentiment Analysis**: The project aims to analyze the sentiment of Yelp restaurant reviews. By processing and understanding the text content of reviews, the project seeks to identify whether a review is positive, negative, or neutral. This sentiment analysis helps restaurant owners understand how customers perceive their services and offerings.
2. **Identifying Advantages and Disadvantages**: By analyzing the content of reviews, the project aims to identify the strengths and weaknesses of restaurants. This information can be valuable for restaurant owners to understand what aspects of their business are working well and what areas need improvement.
3. **Providing Recommendations for Improvement**: Based on the analysis of reviews, the project intends to provide recommendations to restaurants on how they can improve their services, customer experience, and overall ratings. These recommendations may include suggestions for enhancing food quality, improving customer service, or addressing specific complaints raised by customers.
4. **Predicting Star Ratings**: By training machine learning models on review data, the project aims to predict star ratings for restaurants. This predictive capability can help restaurant owners anticipate how customers are likely to rate their establishments based on the content of their reviews. It enables proactive measures to maintain or enhance customer satisfaction and overall ratings.
5. **Insights for Decision-Making**: Overall, the project aims to provide actionable insights for restaurant owners and managers. By leveraging natural language processing and machine learning techniques, it seeks to extract meaningful information from large volumes of review data. These insights can inform strategic decisions aimed at improving business performance, enhancing customer satisfaction, and ultimately driving success in the highly competitive restaurant industry.

Overall, the project aims to leverage NLP and machine learning techniques to extract insights from Yelp restaurant reviews, provide actionable recommendations for improvement, and develop a predictive model for rating future reviews.

PROJECT OVERVIEW

Yelp, a prominent platform for business reviews, provides invaluable insights into customer experiences across various industries, with restaurants being one of the most reviewed categories. This project aims to harness the power of Natural Language Processing (NLP) and machine learning to analyse Yelp restaurant reviews comprehensively. By leveraging NLP techniques, we can extract meaningful information from textual data, such as sentiment, topics, and key insights, to understand customer preferences and sentiments towards restaurants. The goal is to develop machine learning models capable of predicting star ratings based on review text, providing valuable feedback for both restaurants and Yelp analysts.

Overview of the Dataset:

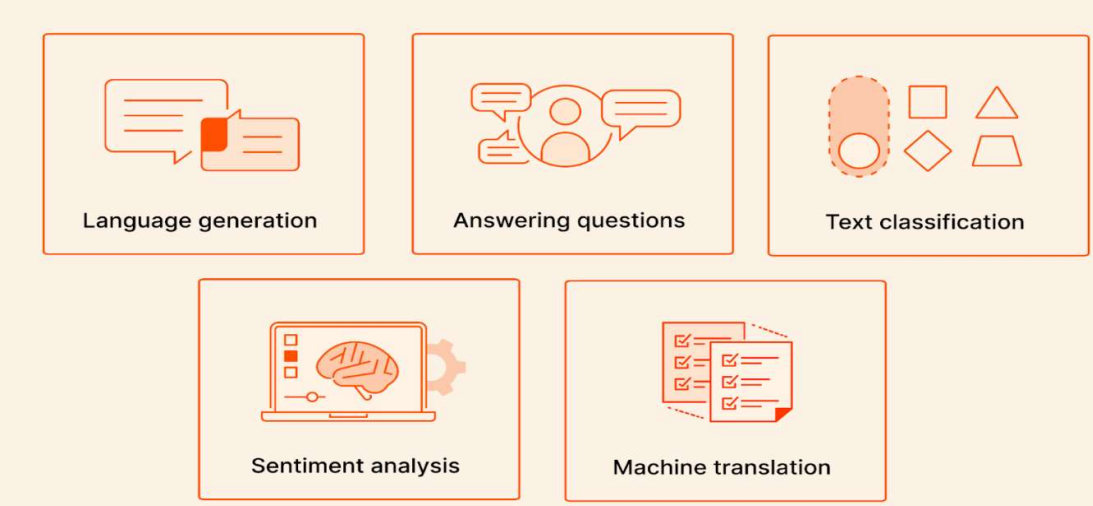
The dataset used in this project will comprise Yelp restaurant reviews, containing information such as review text, star ratings, and additional attributes like 'Cool', 'Useful', and 'Funny'. The reviews are anonymized and cover a wide range of restaurants, allowing for a diverse analysis of customer sentiments and preferences. Data preprocessing steps involve removing irrelevant columns, renaming variables, and computing text length to prepare the dataset for exploratory data analysis (EDA) and machine learning model development.

Overview of the Concepts Involved:

1. Natural Language Processing (NLP):

Natural Language Processing (NLP) is a field of artificial intelligence and computational linguistics that focuses on enabling computers to understand, interpret, and generate human language in a meaningful way. In the context of this project analyzing Yelp restaurant reviews, NLP plays a crucial role in extracting insights from textual data to understand customer sentiments, preferences, and feedback.

Key Components of NLP in this Project



- **Text Preprocessing:**
  - Text preprocessing involves cleaning and preparing textual data for analysis. This includes tasks such as removing punctuation, converting text to lowercase, and eliminating stopwords (common words that do not carry significant meaning).
  - In this project, text preprocessing is essential to ensure that the textual data from Yelp reviews is standardized and ready for further analysis.
- **Word Frequency Analysis:**
  - Word frequency analysis involves counting the occurrences of each word in the text corpus. This analysis provides insights into the most common words used in Yelp reviews, which can indicate recurring themes, topics, or sentiments.
  - By performing word frequency analysis, we can identify key terms and topics that are frequently mentioned in reviews, such as "food", "service", "delicious", "friendly", etc.
- **Sentiment Analysis:**
  - Sentiment analysis aims to determine the sentiment or emotion expressed in textual data. It involves classifying text as positive, negative, or neutral based on the language used.
  - In this project, sentiment analysis can help identify the overall sentiment of Yelp reviews. Positive sentiments may indicate

satisfaction with quality or service, while negative sentiments may highlight areas for improvement.

- **Topic Modeling:**
  - Topic modeling is a technique used to discover abstract topics or themes present in a collection of documents. It identifies clusters of words that frequently co-occur in the text corpus, representing underlying topics.
  - By applying topic modeling to Yelp reviews, we can uncover common themes or topics discussed by customers, such as "food quality", "service experience", "ambiance", etc. This can provide valuable insights into areas of strength or weakness for restaurants.

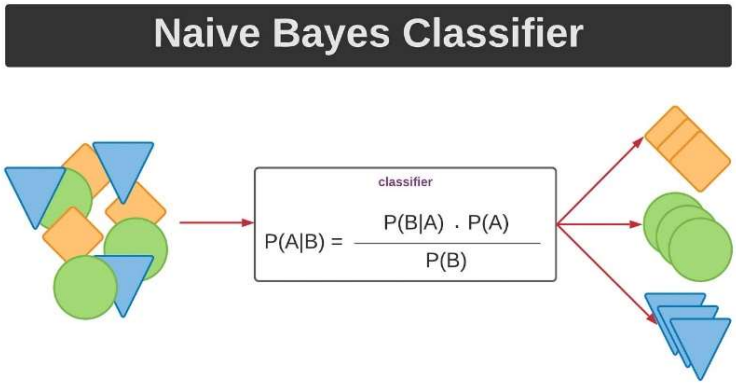
Application of NLP in Predictive Modelling:

- In addition to extracting insights from textual data, NLP techniques can also be used to generate numerical representations of text, which can then be fed into machine learning models for predictive analysis.
- Text vectorization methods such as CountVectorizer and TF-IDF transform textual data into feature vectors that capture the presence and importance of words in the text corpus.
- These feature vectors serve as input to machine learning models, allowing them to learn patterns and relationships between textual features and target variables, such as star ratings in this project.

2. Machine Learning:

Machine learning algorithms are utilized to develop predictive models that can classify or regress based on input data. In this project, machine learning models such as Multinomial Naive Bayes are employed to predict star ratings using review text as input features.

Key Concepts of Naive Bayes:



- **Naive Bayes Classifier:**
  - Naive Bayes is a probabilistic classifier based on Bayes' theorem, which calculates the probability of a hypothesis given evidence. In the context of text classification, the hypothesis is the predicted class label (e.g., star rating), and the evidence is the presence of words in the document.
  - Naive Bayes assumes that features (words) are conditionally independent given the class label, which means that the presence of one word in a document does not affect the presence of another word. This is known as the "naive" assumption.
- **Multinomial Naive Bayes:**
  - Multinomial Naive Bayes is specifically designed for classification tasks with discrete features, such as word counts or term frequencies in text data.
  - In this project, each feature represents the frequency of a word in the review text, and the class labels are the star ratings (e.g., 1, 2, 3, 4, 5 stars).
  - The algorithm calculates the probability of each class label given the feature values (word frequencies) using Bayes' theorem and the naive assumption of feature independence.

- **Probability Estimation:**
  - To classify a new document (review), MNB calculates the probability of each class label given the observed feature values (word frequencies) using the above mentioned formula.
  - Here, P(A|B) is the probability of class label A given the feature values B (word frequencies), P(A) is the prior probability of class label A, P(B|A) is the likelihood of observing the feature values given class label A, and P(B) is the marginal likelihood of observing the feature values.
  - The class label with the highest probability is assigned to the new document.

3. Text Vectorization:

Text vectorization is the process of converting textual data into numerical representations that can be used as input for machine learning algorithms. In the context of this project analyzing Yelp restaurant reviews, text vectorization is a crucial step in preparing the textual data for analysis and predictive modeling.

Key Concepts of Text Vectorization

- **Bag-of-Words (BoW) Model:**
  - The Bag-of-Words model represents each document (review) as a numerical vector, where each element corresponds to the frequency of a word in the document.
  - In this model, the order of words in the document is disregarded, and only the presence or absence of words is considered.
  - Text vectorization techniques such as CountVectorizer and TF-IDF transform the textual data into BoW representations.
- **CountVectorizer:**
  - CountVectorizer is a text vectorization technique that converts a collection of text documents into a matrix of token counts.
  - Each row in the matrix represents a document, and each column represents a unique word (token) in the corpus.

- The value in each cell of the matrix represents the frequency of the corresponding word in the document.
- CountVectorizer is suitable for tasks where word frequencies are important features, such as document classification and topic modeling.

- **Term Frequency-Inverse Document Frequency (TF-IDF):**
  - TF-IDF is a text vectorization technique that calculates the importance of a word in a document relative to its frequency in the entire corpus.
  - It consists of two components: Term Frequency (TF) and Inverse Document Frequency (IDF).
  - Term Frequency (TF) measures the frequency of a word in a document and is calculated as the ratio of the number of times the word appears to the total number of words in the document.
  - Inverse Document Frequency (IDF) measures the importance of a word across the entire corpus and is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the word.
  - TF-IDF is suitable for tasks where common words across documents should be penalized and rare words that are discriminative should be emphasized, such as document classification and information retrieval.

4. Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is a crucial step in data analysis that involves exploring and understanding the characteristics of the dataset before performing further analysis or modeling. In the context of this project analyzing Yelp restaurant reviews, EDA plays a vital role in uncovering insights, patterns, and relationships within the dataset.

Key Aspects of Exploratory Data Analysis:

- **Data Overview:**
  - EDA begins with gaining a comprehensive understanding of the dataset, including its size, structure, and attributes.
  - For the Yelp restaurant review dataset, this involves examining the number of rows and columns, data types, and any missing or duplicate values.

- **Summary Statistics:**
  - Calculating summary statistics such as mean, median, standard deviation, minimum, and maximum values provides insights into the distribution and variability of numerical attributes.
  - For example, in this project, summary statistics can help understand the distribution of star ratings, text lengths, and other numerical attributes like 'Cool', 'Useful', and 'Funny'.
- **Data Visualization:**
  - Visualizing the data using plots and charts helps uncover patterns, trends, and relationships that may not be evident from summary statistics alone.
  - Common types of visualizations include histograms, scatter plots, box plots, and heatmaps.
  - Visualization techniques are applied to explore the distribution of star ratings, relationships between numerical attributes, and patterns in review text length.
- **Feature Relationships:**
  - Exploring relationships between features (attributes) in the dataset helps identify correlations and dependencies.
  - For example, examining the relationship between star ratings and other numerical attributes like 'Cool', 'Useful', and 'Funny' can reveal insights into customer behavior and review engagement.
- **Text Analysis:**
  - In addition to numerical attributes, EDA involves exploring and analyzing textual data, such as review text.
  - Text analysis techniques include word frequency analysis, sentiment analysis, and topic modeling to uncover patterns and themes within the text corpus.
  - Visualizations such as word clouds and bar charts can help visualize the most frequent words and sentiments expressed in the reviews.

TECHNOLOGIES REQUIRED

The project requires a variety of tools and libraries to perform data analysis, NLP, and machine learning tasks. The tech stack includes:

- Python programming language for coding and data manipulation.
- Pandas and NumPy libraries for data preprocessing and manipulation.
- Scikit-learn library for implementing machine learning models and text vectorization techniques.
- Matplotlib and Seaborn for data visualization and exploratory data analysis.
- NLTK (Natural Language Toolkit) and WordCloud for NLP tasks such as text preprocessing and word cloud generation.

By leveraging this tech stack, the project aims to provide a comprehensive analysis of Yelp restaurant reviews, uncovering valuable insights and enabling accurate prediction of star ratings based on review text.

CONCLUSION

By the end of this project, restaurants will have useful information from looking closely at Yelp reviews. They'll know what they're good at, what they can do better, and even get an idea of how many stars they might get based on what people write. This is possible because we'll be using special computer tools called machine learning and natural language processing. These tools help us analyze the reviews in detail, finding patterns and understanding what customers are saying. Armed with this knowledge, restaurants can make smart decisions to make their food and service better, keep customers happy, and stand out in the crowd of other restaurants.

Understanding what they're good at will help restaurants highlight those things and attract more customers. Fixing things that need improvement, like slow service or limited menu options, will make customers happier and more likely to come back. Plus, knowing ahead of time what kind of reviews they might get helps restaurants prepare and avoid any problems before they happen. Overall, this project is about helping restaurants improve and do well in a competitive restaurant world, using advanced computer tools to understand and learn from customer feedback.

REFERENCES

- Yelp official website: https://www.yelp.com/\
- Dataset: https://www.kaggle.com/code/zhenyufan/nlp-for-yelp-reviews/input?select=yelp.csv
- https://www.signpost.com/blog/yelp-business-reviews/
- https://dopweb.com/add-google-and-yelp-reviews-onto-your-website/
- https://docs.developer.yelp.com/reference/v3_business_review_highlights

Research papers referred:

- Natural Language Processing Using Yelp Reviews: https://medium.com/@testandlearn/what-are-your-customers-saying-natural-language-processing-nlp-with-yelp-review-data-65c4ce57287b
- A Qualitative Assessment of Yelp.Com Users' Motivations to Submit and Read Restaurant Reviews: https://www.researchgate.net/publication/267641450_A_Qualitative_Assessment_of_YelpCom_Users'_Motivations_to_Submit_and_Read_Restaurant_Reviews
- CS 229 Project: Improving on Yelp Reviews Using NLP and Bayesian Scoring: https://cs229.stanford.edu/proj2011/BechonGrimaldiMerouchi-ImprovingYelpReviews.pdf