NLP ANALYSIS OF RESTAURANT REVIEWS

INTRODUCTION

In today's digital world, online reviews play a crucial role in shaping consumer opinions and influencing business success. Customers frequently share their dining experiences on various platforms such as Zomato, Yelp, TripAdvisor, or Google Reviews. These reviews contain valuable insights that, if analyzed properly, can help restaurants improve their services and gain a competitive edge. This project focuses on analyzing restaurant reviews using Natural Language Processing (NLP) and Machine Learning techniques to determine whether a review expresses a positive or negative sentiment. By converting unstructured text data into meaningful patterns, we aim to build a predictive model that can automatically classify new reviews based on sentiment. The project demonstrates how textual data can be preprocessed, transformed into numerical features, and used to train a classifier (Random Forest) for sentiment prediction. It highlights the power of NLP in extracting insights from customer feedback and the potential of machine learning in automating decision-making processes.

OBJECTIVE

The primary objective of this project is to develop a sentiment analysis system that can classify restaurant reviews as positive or negative based on the textual content of the review. This project combines Natural Language Processing (NLP) and Machine Learning techniques to automatically interpret and evaluate customer feedback in the form of natural language.

Specific goals include:

- **Data Preprocessing:**

  Clean the raw textual data by removing noise such as punctuation, numbers, and stopwords. Apply lowercasing, tokenization, and stemming to standardize the text.

- **Text Representation:**

  Convert the cleaned text data into a structured numerical format using the Bag of Words model (CountVectorizer), which captures the frequency of words while reducing dimensionality.

- **Model Building:**

  Train a supervised machine learning model, specifically a Random Forest Classifier, using the vectorized text features to learn the patterns associated with positive and negative sentiments.

- **Model Evaluation:**

  Evaluate the performance of the model using metrics such as accuracy, confusion matrix, precision, recall, F1-score, and ROC-AUC score to ensure reliability and robustness.

- **Prediction:**

  Use the trained model to predict the sentiment (positive/negative) of unseen restaurant reviews, effectively automating the sentiment classification process.

- **Insight Generation:**

  Provide insights into customer satisfaction that can be valuable for restaurants to improve their services, menus, or overall experience based on sentiment trends.

- **Visualize Trends and Insights**

Using data visualization libraries like **Matplotlib and Seaborn**, the project transforms raw analysis into **easy-to-understand visuals**. Charts show how sentiment changes over time, which topics dominate discussions, or how ratings vary across locations.

Tech Stack

- **Languages & Libraries**: Python, Pandas, NLTK, Scikit-learn, Matplotlib, Seaborn, Numpy, re
- **Techniques**: Text Preprocessing (stopwords removal, stemming), Sentiment Analysis,Bag of Words Model (CountVectorizer),Logistic Regression, Naive Bayes, Model Evaluation (Accuracy, Confusion matrix, ROC-AUC)

METHODOLOGY

➢ Text Cleaning & Preprocessing

- Raw review texts were cleaned by removing special characters, stopwords, and punctuation.
- Applied **tokenization, stemming, and lemmatization** to standardize the text for analysis.
- Ensured uniform structure for better model performance and reliable results.

➢ Feature Extraction:

- Convert text into numerical vectors using Bag of Words.

➢ Text Vectorization

- Converted text data into numerical format.
- Enabled effective input for machine learning models.

➢ Train-Test Split:

Split data into training and test sets (75:25).

➢ Model Training:

Apply Naïve Bayes to train the model.

- **Evaluation:**

Evaluate using accuracy, confusion matrix, classification report, and ROC-AUC score.

- **Data Visualization**

- Created charts and graphs using **Matplotlib and Seaborn** to display:
    - Sentiment distribution
    - Most frequent words
    - Topic prevalence
    - Time-based sentiment trends
- Made insights easy to understand for non-technical stakeholders.

OBSERVATION

Through this project, several important observations were made at different stages of data processing, model training, and evaluation:

➢ **Data Quality & Structure**

- The dataset consists of 1000 restaurant reviews labeled as either positive (1) or negative (0).
- The reviews are in free-text format, requiring thorough cleaning before being used for analysis.
- There were no missing values, but the text contained noise such as punctuation, numbers, and capitalized letters, which had to be normalized.

➢ Text Preprocessing

- Techniques such as punctuation removal, lowercasing, stopword removal, and stemming significantly improved the quality of textual data for analysis.
- After preprocessing, the reviews were converted into a consistent and compact format, making them suitable for machine learning.

➢ Feature Extraction

- The Bag of Words (CountVectorizer) approach was used to transform textual data into numerical vectors.
- A limit of 1500 features was set to balance between performance and computational efficiency.
- Most common and meaningful words were captured well, helping the model learn effectively.

➢ Model Performance:

- The Naïve Bayes Classifier performed well on the test data, showing good accuracy and balanced classification for both positive and negative classes.
- Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score indicated that the model was not only accurate but also robust in handling both classes.

➢ Evaluation Insights:

- The confusion matrix and heatmap helped visualize the correct and incorrect predictions, highlighting that the model had more success predicting positive reviews than negative ones, which is common in imbalanced sentiment datasets.
- ROC-AUC Score confirmed that the model could distinguish between positive and negative reviews effectively.

➢ Consistency:

- It was observed that using a fixed random_state helped in maintaining consistent results during model training and test splits across multiple runs.
- Without fixing the random_state, the results (accuracy and matrix) varied due to randomness in splitting data.

CONCLUSION

This project successfully demonstrates the application of Natural Language Processing (NLP) and Machine Learning techniques in analyzing and classifying customer restaurant reviews. By preprocessing the text data, transforming it into numerical features using the Bag of Words model, and training a Naïve Bayes Classifier, we were able to build a model that effectively predicts the sentiment (positive or negative) of a given review.

The evaluation metrics, including accuracy, confusion matrix, classification report, and ROC-AUC score, indicate that the model performs well in identifying customer sentiments. Such sentiment analysis can be a powerful tool for restaurant businesses to monitor customer satisfaction, understand areas for improvement, and make data-driven decisions.

In conclusion, this project highlights the real-world value of combining NLP and machine learning to extract insights from unstructured text, and it sets a foundation for more advanced sentiment analysis applications in the future.