Sentiment Analysis Project Report

Introduction

The objective of this project is to build a Restaurant Review Sentiment Analysis system using a Machine Learning Model. The project utilizes restaurant reviews from both TripAdvisor and Google to develop a sentiment analysis model. The goal of this project is to predict the sentiment of a review as either Negative, Neutral, or Positive, based on the text of the review.

Libraries Used

The following libraries are used in this project:

- Pandas: For reading files and handling data.
- NumPy: For data manipulation and related operations.
- Matplotlib: For plotting graphs and data visualization.
- Seaborn: For enhanced data visualization.

Purpose of the Project

The project's main purpose is to complement the traditional rating system used for restaurant reviews with sentiment analysis. Here are the key reasons for integrating sentiment analysis alongside a rating system:

- Deeper Insights: Sentiment analysis provides a more detailed understanding of customer feedback by categorizing reviews as positive, negative, or neutral. This allows for deeper insights into the reasons behind the ratings.

- Filtering Ambiguity: Rating systems can be misleading when a customer gives a low score due to disliking one aspect of the service but liking other aspects. Sentiment analysis helps filter such cases more effectively.

Data Loading and Preprocessing

Data Sources

Two datasets are used in this project:

1. TripAdvisor Reviews
2. Google Reviews

These datasets are loaded using Pandas for further analysis.

Data Preprocessing

Data preprocessing includes handling missing values. Any rows with missing data are dropped from both datasets to ensure the quality of the data used for analysis.

Summary Statistics

Summary statistics are generated for both the TripAdvisor and Google Review datasets. This helps provide a quick overview of key metrics such as count, mean, standard deviation, minimum, and maximum for numerical columns in the datasets.

Data Visualization

Data visualization is performed to gain a better understanding of the datasets and extract meaningful insights.

Distribution of Ratings

Distribution of Ratings for TripAdvisor and Google Reviews is visualized using count plots, showing the frequency of each rating.

Distribution of Ratings by Location

Box plots are used to visualize the distribution of ratings by location for both TripAdvisor and Google Reviews. This helps identify variations in ratings across different locations.

Word Clouds

Word clouds are created for both TripAdvisor and Google Reviews to visualize the most frequent words in the review texts. This provides a visual representation of the most commonly mentioned words.

Distribution of Review Lengths

Histograms with KDE (Kernel Density Estimation) plots are used to visualize the distribution of review lengths for both datasets. This helps understand the distribution of review text lengths.

Sentiment Analysis

Data Merging

The TripAdvisor and Google Review datasets are merged into a single dataset to create a comprehensive dataset for sentiment analysis.

Data Splitting

The combined dataset is split into training and testing sets. The training set contains 80% of the data, while the testing set contains 20%.

Model Training

Sentiment analysis is performed using a Multinomial Logistic Regression model. The following steps are taken:

- TF-IDF Vectorization: Text data is transformed into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. A TF-IDF vectorizer is initialized and fitted on the training data.

- Model Training: A Multinomial Logistic Regression model is trained using the TF-IDF vectors of the training data and the corresponding sentiment labels (Negative, Neutral, Positive).

- Model Evaluation: The model is evaluated on the testing data, and accuracy metrics are computed to assess its performance.

Model Evaluation

The model's performance is evaluated using accuracy metrics and a classification report. The classification report includes precision, recall, F1-score, and support for each sentiment category (Negative, Neutral, Positive).

The accuracy of the model is found to be approximately 86%.

Model Deployment

The trained TF-IDF vectorizer and Multinomial Logistic Regression model are saved to files using the pickle library for future use.

Prediction Function

A prediction function is defined to predict the sentiment category of a user's input restaurant review. This function uses the saved model and TF-IDF vectorizer to preprocess the input text

and make predictions. It returns the predicted sentiment category (Negative, Neutral, Positive) along with probabilities for each category.

Sample Predictions

Sample predictions are provided to demonstrate the functionality of the prediction function. The predicted sentiment category and probabilities are shown for three different user input restaurant reviews.

Review 1:

Text: The food was terrible, and the service was slow.
Predicted Category: Negative
Probabilities:
Negative: 0.9807
Neutral: 0.0186
Positive: 0.0008

Review 2:

Text: The food was very good, and the service was excellent. Worth going to that restaurant.
Predicted Category: Positive
Probabilities:
Negative: 0.0019
Neutral: 0.0049
Positive: 0.9932

Review 3:

Text: The food was ok.
Predicted Category: Neutral
Probabilities:
Negative: 0.0713
Neutral: 0.7035
Positive: 0.2252

Conclusion

The Restaurant Review Sentiment Analysis project successfully demonstrates the implementation of a machine learning model to predict the sentiment of restaurant reviews. By combining data from TripAdvisor and Google Reviews, preprocessing, and training a Multinomial Logistic Regression model, the project provides a valuable tool for restaurant owners and stakeholders to gain deeper insights into customer feedback. The model can help identify areas for improvement and enhance the overall customer experience. Additional features such as sentiment trends over time can be incorporated by using more advanced natural language processing techniques.