

Savoring Sentiments: Predicting Restaurant Recipe Ratings through Sentiment Analysis

Soham Bhattacharya [21f3001729]

Indian Institute of Technology, Madras

Jan - Apr 2024

1 Introduction

This project focuses on predicting restaurant recipe ratings using sentiment analysis techniques. The dataset consists of user reviews and ratings collected from restaurant recipes. The goal is to build a machine learning model that accurately predicts ratings based on textual and numerical features.

2 Methodology

2.1 Data Collection and Preprocessing

- Raw data was obtained from CSV files and split into training and test datasets.
- Missing values were identified and handled accordingly.
- Feature engineering was performed to extract meaningful numerical and textual features.

2.2 Feature Selection

- Important numerical features: RecipeNumber, ReplyCount, ThumbsDownCount, ThumbsUpCount, BestScore.
- Categorical features: RecipeName.
- Text feature: Recipe Review.
- Feature importance was determined using a **Random Forest Classifier**. The classifier was trained on the dataset and the importance of each feature was measured using Gini importance.
- Features with an importance score greater than 0.05 were selected for the final model.
- The cumulative importance of selected features was calculated to ensure optimal feature selection.
- The most significant features included: RecipeNumber, ReplyCount, ThumbsUpCount, and BestScore.

3 Model Development

A pipeline was built using a column transformer to preprocess numerical, categorical, and text features. The LightGBM (LGBM) classifier was chosen for its efficiency and accuracy.

- Numerical features were scaled using MinMaxScaler.
- Categorical features were encoded using OrdinalEncoder.
- Text features were vectorized using TF-IDF with a maximum of 5000 features.
- The model was trained using LightGBM with default hyperparameters.

4 Results and Evaluation

- The model achieved an accuracy of **93.85%** on the training set.
- Classification report:

Precision	Recall	F1-score	Support	
Class 0	0.67	0.98	0.79	865
Class 1	1.00	1.00	1.00	210
Class 2	1.00	1.00	1.00	174
Class 3	0.98	1.00	0.99	363
Class 4	0.69	0.99	0.81	857
Class 5	1.00	0.93	0.96	11167
Accuracy			0.94	13636

- The predicted ratings were stored and exported as a CSV file for submission.

5 Conclusion

This study successfully applied machine learning techniques to predict restaurant recipe ratings based on sentiment analysis. Future work includes optimizing hyperparameters, exploring deep learning models, and incorporating additional textual sentiment analysis methods.