Foody&Foodie

Consumer Sentiment Analysis

Business Understanding

Overview

Foody&Foodie are a family run restaurant in San francisco, CA. As a veteran in the food business they understand that they will need to keep track of sentiments from their customer base to maintain a competitive edge in the market.

Problem Statement

The management of Foody&Foodie understand the need to find reliable feedback to guide their decision-making aiming to improve their business and match their markets ever changing needs.

Business Understanding

Objectives

Main Objective

- To create a model that could successfully predict the sentiment of a customer's review. The model would attain a recall score and accuracy score above 80%

Specific Objective

- To identify the most common words used in the dataset using a Word cloud.
- To confirm the most common words that are positively and negatively tagged.
- To recognize the products that have been opined by the customers.
- To spot the distribution of the sentiments.

Business Understanding

Challenges

Within the Food industry, they're several measurable parameters to that determine the success of a restaurant and we must find a neutral data set where all these aspects can be fairly evaluated for an accurate result.

Proposed solution

We will need to create a model to analyze customer sentiments through reviews on restaurants within the target area using a single popular review site, Yelp, to pull the relevant data as it is the most comprehensive compilation of reviews in the target market.

Data Understanding

Our dataset is a compilation of writen reviews, ratings, review IDs, review date and business IDs from the Yelp website. The compilation of this data allows us to clearly identfy positive and negatve sentiments in relation to a rating given on a scale of 1-5 as well as reactions to the sentiment by readers categorized as cool, useful or funny. Post cleaning the data, these are the metrics we will use to isolate and model a collective non biased scale of opinions on restarunts in the area.

The dataset for this analysis is sourced from Yelp and contains detailed food reviews. It encompasses a total of 429,771 rows and 8 columns, structured in a wide format as shown below.

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	iBUJvIOkToh2ZECVNq5PDg	iAD32p6h32eKDVxsPHSRHA	YB26JvvGS2LgkxEKOObSAw	5	0	0	0	I've been eating at this restaurant for over 5...	2021-01-08 01:49:36

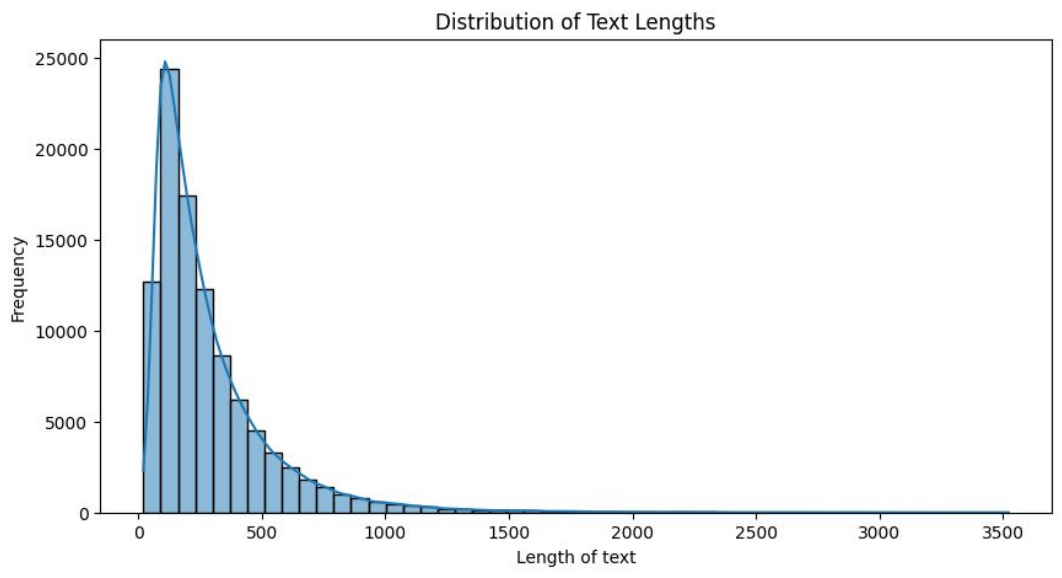
Data Processing

We loaded a dataset of 429,771 restaurant reviews and examined its structure by viewing the first and last few rows, checking the column names, data types, and memory usage. This gave us a clear understanding of the data's content and format, helping us identify any necessary preprocessing steps for further analysis.

We cleaned and preprocessed the dataset by:
Dropping unnecessary columns and duplicate rows.
Handling missing values by either removing rows or filling them appropriately.
Calculating the length of each text review.
Categorizing reviews into positive, neutral, or negative based on star ratings.
Converting all text to lowercase to ensure uniformity.
Removing punctuation, numbers, and common stopwords to reduce noise.
Lemmatizing words to reduce them to their root forms, simplifying the text.
Saving the cleaned data back to the original file for future analysis.

Data Analysis

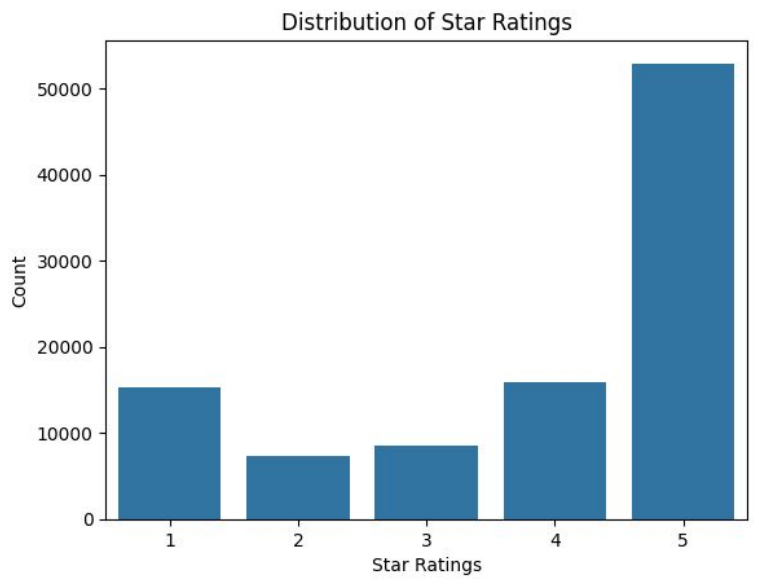
Distribution of Review Lengths



The majority of short reviews suggest users prefer brief comments, suggesting Food&Foodie should focus on concise messaging. The content may include quick impressions or detailed experiences, offering insights into customer satisfaction.

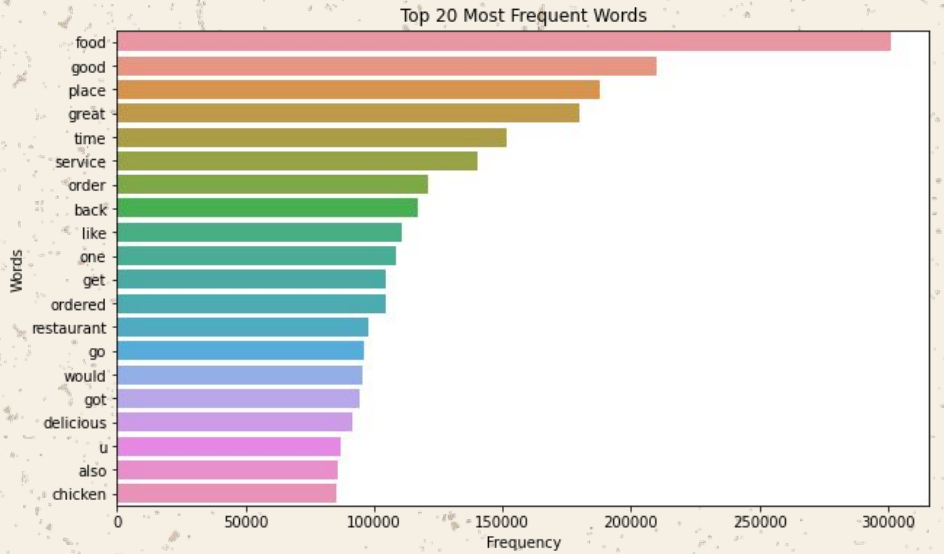
Data Analysis

Distribution of Star Ratings



The high number of 5-star reviews on Yelp may make it challenging to distinguish outstanding businesses from average ones due to the skewed nature of ratings. Additionally, Yelp reviewers may be more motivated to leave a review after a positive experience.

Data Analysis



The majority of reviews express positive sentiments, indicating customer satisfaction at restaurants. Key positive keywords focus on food quality and customer service, indicating their importance.

Data Analysis

Negative Reviews: Customer service is the most frequently mentioned area for improvement in negative reviews. This suggests a need to address potential issues in customer service training or procedures.

Neutral Reviews: Food quality appears to be the primary focus in neutral reviews. While this isn't necessarily bad, it indicates a lack of strong opinions about other aspects like service or atmosphere.

Positive Reviews: Customers express satisfaction with both service and food in positive reviews, often recommending the establishment to others. This highlights the positive aspects of the dining experience.

Data Modeling Evaluation

Our evaluation compares various machine learning models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), Multinomial Naive Bayes, and Decision Trees, before and after hyperparameter tuning. The models were evaluated based on accuracy and recall scores.

Logistic Regression and Random Forest performed the best, especially after tuning, with accuracy and recall scores around 90%.
Support Vector Machine (SVM) had a significant accuracy improvement after tuning, but its recall score dropped, indicating that the model became less balanced.
Multinomial Naive Bayes and Decision Trees showed no significant improvement after tuning, with both models maintaining an accuracy and recall of around 78%.

Model Deployment

In this deployment process, we:

Vectorized Text Data: Converted text data into numerical form using TF-IDF vectorization.
Trained a Model: Used a RandomForestClassifier to train a model on the vectorized data.
Saved the Model and Vectorizer: Stored the trained model and vectorizer as pickle files for easy reuse in production.
For predictions, the text is preprocessed, vectorized using the saved vectorizer, and passed to the saved model to predict the sentiment (positive, neutral, or negative).

This ensures that the model and vectorizer are ready for deployment and can be used to classify new text inputs efficiently.

ANY QUESTIONS?

THANK YOU!