Classifying Sentiments in the Olist Reviews Dataset

Gabriel Yamashita

I. Dataset

The dataset utilized for this classification task includes two main files: `olist_orders_dataset.csv` and `olist_order_reviews_dataset.csv`. These files were loaded using the pandas library and subsequently merged on the `order_id` to create a comprehensive view of customer reviews and their corresponding orders. The merged dataset contains various fields, including the review score and the review comments, which are crucial for the sentiment classification process.

II. Classification Pipeline

For the classification task, Logistic Regression was implemented as the classification method. A machine learning pipeline was established, which included preprocessing steps such as tokenization, stemming, and vectorization of the review text data using TF-IDF. This approach enables the model to quantify the importance of each word in the classification process effectively.

III. Evaluation

The model's performance was evaluated using multiple metrics, including accuracy, confusion matrix, and classification report. These metrics provide insight into how well the model classifies the sentiments expressed in the reviews, allowing for the identification of any misclassifications and overall effectiveness.

IV. Dataset Size

The dataset comprises a significant number of reviews, and the distribution of star ratings was analyzed to understand how sentiments vary. The analysis included a bar chart visualizing the number of reviews corresponding to each star rating, providing a clear view of customer satisfaction levels.

V. Topic Analysis

Exploratory data analysis was performed to gain insights into the sentiments expressed in the reviews. Although specific topic analysis was not detailed in the notebook, the distribution of reviews suggests trends in customer satisfaction and highlights areas for improvement based on the reviews provided.