Classifying Sentiments in the Olist Reviews Dataset

Gabriel Yamashita

I. Dataset

The dataset used in this analysis includes customer reviews and order data from the Olist e-commerce platform. We utilized two main files: `olist_orders_dataset.csv` and `olist_order_reviews_dataset.csv`. After loading the data, the datasets were merged on the `order_id` field to combine review data with associated orders. This dataset contains review scores, review messages, and other order-related fields.

II. Classification Pipeline

We implemented a Logistic Regression classifier to predict the sentiment (positive or negative) of customer reviews. The text data was preprocessed using stemming techniques (RSLP Stemmer), and vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) to quantify the importance of each word.

A pipeline was constructed to automate these steps:

- Text Preprocessing (Tokenization, Stemming)
- Vectorization (TF-IDF)
- Classification (Logistic Regression)

III. Evaluation

To evaluate the model's performance, we used several metrics:

- Accuracy: Proportion of correctly classified reviews.
- Confusion Matrix: Visualization of predicted vs. actual classifications.
- Classification Report: Precision, Recall, and F1-Score for both classes.

IV. Dataset Size

The dataset contains thousands of reviews, with the following distribution of review scores (5 stars being the highest, 1 star being the lowest). Figure 1 shows the distribution of star ratings.

V. Topic Analysis

A preliminary analysis of the reviews suggests common themes around customer satisfaction and product quality. Further topic analysis would allow the identification of frequent topics mentioned in positive and negative reviews.

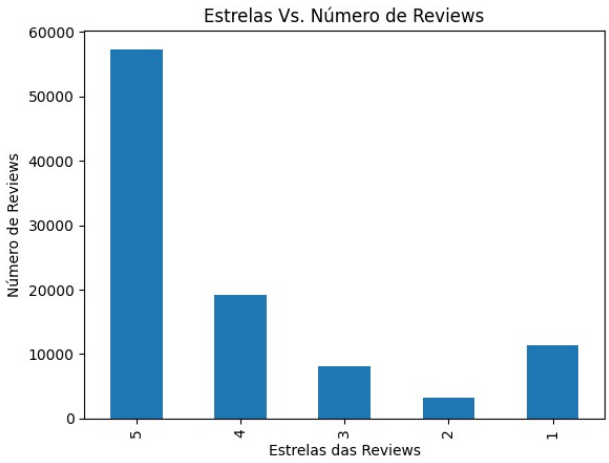

Fig. 1. Distribution of Review Scores. The majority of reviews are concentrated on 5 and 4-star ratings, indicating positive feedback.