Customer Review Sentiment Analysis Report

For: Amazon
10/08/24
By: Alex Hosp

Executive Summary

This report presents the results of a predictive analysis to determine whether the overall sentiment of customer reviews in the given dataset **amazon_product_reviews** is predominantly positive or negative. The main objective of this project was to answer the question: **“Are most of the customer reviews positive or negative?”**

To answer this question, I analyzed a subset of the data containing both **star ratings and textual product reviews**. I conducted **exploratory data analysis (EDA)** to assess the distribution of star ratings and employed a **pre-trained DistilBERT transformer model**, fine-tuned for sequence classification tasks such as sentiment analysis.

The **key findings** indicate that **78.9% of the reviews have 5-star ratings**, while **68% showed positive sentiment** based on text analysis. This suggests that overall customer sentiment is largely positive, though some individual reviews may contain mixed feedback.

Due to the use of a **binary classification model**, some subtleties in the reviews may not have been fully captured. However, the overall alignment between star ratings and sentiment analysis provides a reliable indication of customer perception. These findings indicate that **customers are generally satisfied with Amazon’s products**, though **32% of reviews exhibited some negative sentiment**. A more nuanced analysis of the text could provide further insights into areas for potential improvement.

Objective

The objective of this project was to analyze Amazon's `amazon_product_reviews` dataset to determine whether customer reviews are generally positive or negative. Specifically, this analysis aims to answer the question, **“Are most of the customer reviews positive or negative?”**

To answer this question, **star ratings** and **textual product reviews** are analyzed to assess customer satisfaction comprehensively. The goal is to leverage sentiment analysis to extract meaningful insights, enabling *Amazon* to better understand customer perception of their products.

Methodology

1. Data Preparation

Imported the dataset into a Google Colab notebook and cleaned the data in the `reviewText` field by:

- **Removing Nulls and Duplicates:** Excluded entries with missing `reviewText` and duplicate reviews to ensure data integrity.
- **Handling Text Length:** Truncated reviews that exceeded the model's token limit to maintain consistency.
- **Eliminating Special Characters:** Removed special characters that could interfere with model processing.

2. Exploratory Data Analysis (EDA)

Performed EDA to get a preliminary understanding of the sentiment distribution by:

- **Analyzing Star Ratings:** Examined the distribution of 1 to 5-star ratings to get an overview of customer satisfaction levels based on ratings.
- **Visualization:** Created a histogram chart to illustrate the frequency of each star rating, highlighting key trends.

3. Data Preprocessing & Model Selection

Prepared the data for sentiment analysis using **DistilBERT**, chosen for its 91.3% accuracy on the SST-2 dataset. The data preprocessing and model selection steps included:

- **Tokenization:** Converted text reviews into tokens using DistilBERT's tokenizer.
- **Padding and Truncation:** Standardized input lengths by adding padding to shorter reviews and truncating longer ones to fit the model's requirements.
- **Model Configuration:** Selected a fine-tuned, pretrained snapshot of the DistilBERT model for its efficiency and strong performance in sequence classification tasks.

4. Inference

Performed sentiment prediction through the following steps:

- **Batch Processing:** Split the dataset into batches to optimize resource usage during inference.
- **Model Inference:** Passed preprocessed batches through DistilBERT to obtain sentiment predictions (1 for positive, 0 for negative).

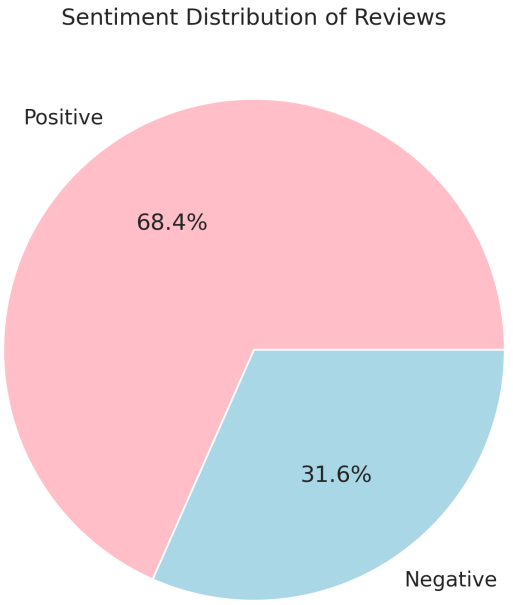
- **Result Extraction:** Applied the argmax function to determine the most likely sentiment for each review.
- **Result Storage:** Processed all batches and stored the prediction results for further analysis.
- **Adding Sentiment Labels**: Converted the predicted sentiment values to classes (1 for **Positive**, 0 for **Negative**) and added them as a new column in the dataset.

5. Results Calculation & Visualization

Calculated the **positive and negative sentiment percentage** and visualized it using a **pie chart**. I then compared this sentiment distribution to the **star rating distribution** from the EDA to assess the consistency between ratings and textual sentiment. I used these analysis results to answer the client's question.

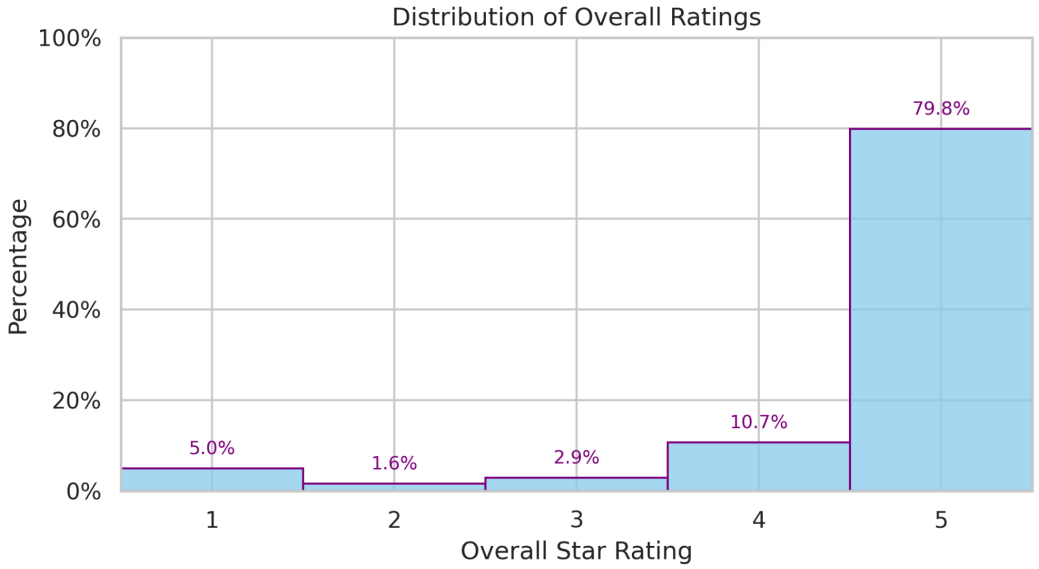
Results

The **sentiment analysis** and **star rating distribution** provide a comprehensive view of customer sentiment based on customer reviews and ratings. Based on the sentiment predictions, **68% of the reviews showed positive sentiment**, while **32% were classified as negative**. This distribution is visually represented in the **pie chart** below:



Similarly, the analysis of **star ratings** showed that **80% of the reviews** had a **5-star rating**, making it the most common rating in the dataset. **11% of reviews** received **4 stars**, while **1-star**

ratings were the next most frequent. **2-star and 3-star ratings** were almost nonexistent. This distribution is illustrated in the **histogram** below:



Comparing these results shows that most reviews have **positive sentiment** (68%), which aligns with the **80% 5-star ratings** in the dataset. However, there were instances where highly rated reviews contained **negative sentiment**, indicating mixed feedback that may suggest areas for improvement.

To answer the main question—**"Are most of the customer reviews positive or negative?"**—the analysis indicates that **most reviews are positive**. The dominance of 5-star ratings and the **68% positive sentiment** in text reviews both demonstrate that customers are generally satisfied. However, the presence of some negative sentiment, also in high-rated reviews indicates opportunities for further improvement. These opportunities could be explored by employing more nuanced text-analysis methods to determine key themes in customer reviews.