Amazon Review Sentiment Analysis

Freenor's Fourth Years - Christina Barton, Rohan Kohli, Abby Goss
Group Leader: Rohan Kohli
DS 4002, 2/8/2026

Goal Statement:
The goal of this project is to determine whether sentiment from Amazon customer review text can accurately predict customer satisfaction rating (> 90% accuracy when distinguishing between high and low).

Research Question:
Can sentiment analysis from Amazon customer review text reliably predict whether a customer will give a high or low satisfaction rating?

Modeling Approach:
We will preprocess review text using standard natural language processing techniques, including tokenization, stop-word removal, and text normalization [1]. Sentiment scores will be generated using a rule-based sentiment analysis method such as VADER, which is designed for short, opinionated text and has been shown to perform well on online reviews [2]. The sentiment scores can then be used as features in a classification model, such as logistic regression, to predict whether a review corresponds to a high or low rating. Sentiment analysis performance will be determined by comparing the feature importance of the sentiment analysis variable compared to other variables in our model. To evaluate model performance, we will use metrics such as accuracy, precision, recall, F1, and correlation between predicted sentiment and actual ratings. This approach allows us to assess both the predictive power and limitations of sentiment analysis in modeling customer satisfaction.

Executive summary:
This document outlines our dataset and plan for analysis for a study of Amazon customer reviews from 2024. We provide details on the dataset, highlight findings and resolved questions from our exploratory data analysis, and provide a road map for future work to be done by our team.

Data set establishment details:

This dataset contains customer reviews for various Amazon experiences. Columns include information about the reviewer, their rating of the product or service, and detailed customer feedback in review text. We aim to utilize review text to predict customer ratings using sentiment analysis. This dataset was found in Kaggle; we downloaded it as a CSV to perform exploratory data analysis.

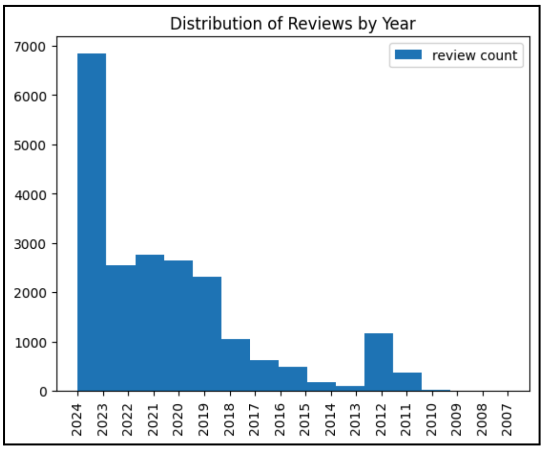
Data Dictionary:

Column	**Description**	**Example Response**
Reviewer name	The name or pseudonym of the reviewer	Anita James

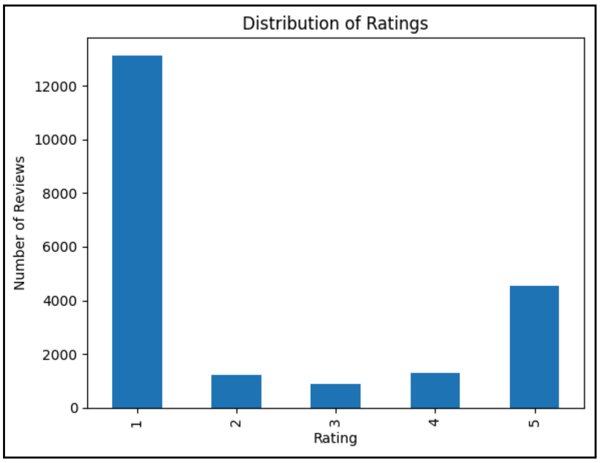
Profile Link	A link to the user's Amazon profile	/users/652e72b822182a0013706f7f
Country	The country of the reviewer's location	US
Review Count	Number of reviews by the same user	2 reviews
Review Date	Date and time of when the review was posted	2024-09-16T13:44:26.000Z
Rating	A numerical measure of satisfaction with the product or service	Rated 1 out of 5 stars
Review Title	Text that summarizes the review sentiment	I love amazon
Review Text	Detailed customer feedback of the product or service	Once again my delivery is late with a message saying "we're so sorry. I'm so tired of this company…
Date of Experience	Date of when the product or service was experienced	September 14, 2024

Questions explored and answered about the dataset:

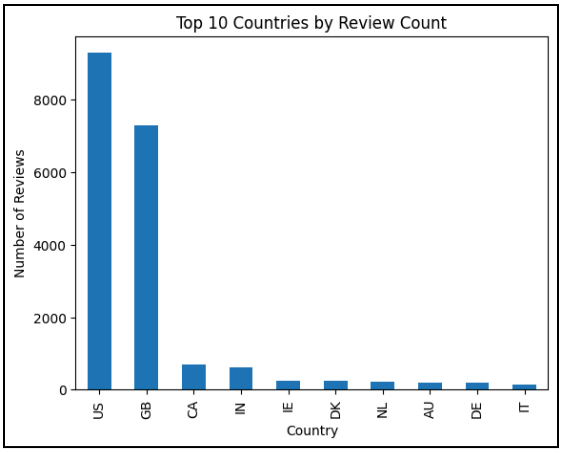
- What is the distribution of reviews by year? Is it mostly reviews from recent years?
    - Most of the reviews (~7000) come from 2024, the most recent year in the dataset. There is far more data from 2019 onwards than from 2010-2018.



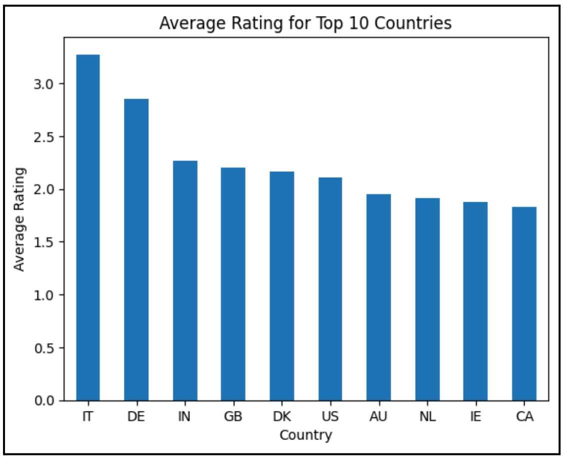
- What are the distributions of ratings like?
    - There are a larger number of 1/5 and 5/5 reviews than other scores. Thus, the dataset consists of more extreme/polarizing reviews



- Which countries are the reviews from?
    - The US and Great Britain make up the vast majority of reviews. Other countries include Canada, India, Ireland, Denmark, the Netherlands, Australia, Germany, and Italy.



- Does the average rating vary by country?
    - Within the top ten countries, the average rating only varies by a small amount, with Italy having the highest average reviews above a 3 out of 5.



Unknowns resolved in analysis plan:

- Is the Review Title a good indicator of review sentiment?
    - There are times when a review title is not a good indicator. For example, one title read “Every time there is a problem,” indicating negative sentiment.

However, the full review text stated “every time there is a problem, they fix it”. While the title might be indicative of a lower rating, the content within the review is actually positive.
    - To resolve this, we plan to focus on the review text itself rather than review titles for sentiment analysis to predict ratings. We will drop the “Review Title” column.
- The review score column has text like “1 out of 5 stars” rather than a single number. This is not useful for analysis.
    - We will create a new column that converts this text into an integer with solely the review score.
- There are not many review scores of 3 (indicating a somewhat neutral review), making it hard to predict. It is generally hard to predict neutral sentiment.
    - We will focus on predicting positive or negative reviews only. As such, we will drop all rows with a review score of 3.
- There are far more reviews from 2024 compared to other years.
    - Thus, we will focus on 2024 reviews only
    - Filter the “Review Date” column to only include data from 2024
- Columns like reviewer name and profile link are not helpful for predicting ratings
    - We will drop these columns
- Nearly every column has null values somewhere in the dataset
    - We will drop any rows with a null value in either the rating or review text column.

Refinement of goal/research plan:

- Due to the information we have found in the original exploratory analysis of the data, the goal of this project has been slightly adjusted. Originally, we had planned to create a model that can predict a review score based on sentiment, but due to the high distribution of extreme review scores, we now aim to predict high (4 or 5 out of 5) and low scores (1 or 2 out of 5).

<u>Analysis Plan</u>

- Goal: Clean data from Kaggle dataset and analyze the sentiment.

The goal of this analysis is to determine whether sentiment extracted from Amazon customer review text can accurately predict customer satisfaction ratings, with a target classification accuracy greater than 90% when distinguishing between high and low ratings. This analysis will guide whether sentiment analysis alone is sufficient for modeling customer satisfaction or whether additional features are needed.

- Preprocessing
    - Select only reviews/rows from 2024
    - Exclude columns (name, profile link, review title)
    - Split date columns

The first step focuses on preparing a clean dataset from the Kaggle Amazon reviews CSV. Only reviews from the year 2024 will be included to align with the study scope and avoid temporal inconsistencies. The date column will be split into year, month, and day components, and reviews outside of 2024 will be removed. Columns not relevant to the analysis, such as reviewer name, profile information, or other personally identifying fields, will be excluded. The retained columns will include review text, rating, review date. Rows with missing data will be removed, and ratings will be classified into “high” and “low” scores with 1-2 out of 5 being “low” and 4-5 out of 5 as “high” and removing reviews with 3 as the ranking.

- Methodology
    - Sentiment analysis
    - Logistic regression for prediction

Once the text data is cleaned, sentiment scores will be generated using a rule-based sentiment analysis method. The VADER sentiment analyzer will be applied to each review’s text. VADER produces compound, positive, neutral, and negative sentiment scores, which work well for short, opinionated online reviews. The compound sentiment score will be used as the primary feature; this step transforms unstructured text into structured numerical features usable in statistical and predictive modeling.
Logistic regression will be used as the main classification model due to its suitability for binary outcomes. The dataset will be split into training and testing sets (80% training, 20% testing), and we will use stratified sampling to preserve the proportion of high and low ratings in both sets. K-fold cross-validation (k=5) will be applied on the training data to assess model stability and reduce overfitting.

- Evaluation
    - Metrics: feature importance, accuracy, precision, recall, F1, and correlation between predicted sentiment and actual ratings

The analysis will be considered successful if the classification accuracy is >90% for the test set. Additionally, we will consider the most important feature in the model to determine if sentiment analysis is the greatest predictor.

Path to success:

<u>References</u>

[1] Mustapha Tijani, “The Complete Guide to NLP Text Preprocessing: Tokenization, Normalization, Stemming, Lemmatization, and More,” DEV Community, Nov. 14, 2025. https://dev.to/themustaphatijani/the-complete-guide-to-nlp-text-preprocessing-tokenization-normalization-stemming-lemmatization-50ap.

[2] C. Hutto and E. Gilbert, “VADER: a Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, no. 1, May 2014, doi: https://doi.org/10.1609/icwsm.v8i1.14550.