Name: Mostafa Murshad Ali

Email: mostafamurshad.ali@baruchmail.cuny.edu, mostafamurshadali@gmail.com

Proposal: Sentiment Analysis on Amazon US Customer Reviews

Milestone 1

Dataset Description

The Amazon Customer Reviews dataset, colloquially known as Product Reviews, is an assemblage of insights drawn from over two decades of customer feedback on Amazon.com. Since the advent of the first review in 1995, millions of customers have shared their perspectives, culminating in over a hundred million reviews. These reviews not only capture the nuances of customer experiences but also provide a glimpse into their perceptions across geographical diversities, potential promotional biases, and an array of product categories. This dataset serves as a rich reservoir for researchers venturing into the realms of Natural Language Processing, Information Retrieval, and Machine Learning.

Dataset Location

The dataset can be accessed and downloaded from: https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset?select=amazon_reviews_multilingual_US_v1_00.tsv.

Attributes of the Dataset

marketplace: Letter country code of the review's origin.

customer_id: Customer identifier for individual reviewers.

review_id: Unique ID for each review.

product_id: Unique identifier for each product.

product_parent: Identifier to group reviews of the same product.

product_title: Name of the product.

product_category: Broad categorization of products.

star_rating: Rating (1-5) given by the customer.

helpful_votes: Count of votes deeming the review helpful.

total_votes: Total votes received by the review.

vine: Indicator if the review was part of the Vine program.

verified_purchase: Flag for reviews of verified purchases.

review_headline: Review title.

review_body: Detailed text of the review.

review_date: Date of the review's creation.

Objective

My primary intent is to delve into Sentiment Analysis of the reviews. Given the importance of user feedback in shaping purchase decisions, understanding the sentiment behind reviews can empower businesses to respond better to their consumer base.

Potential Possible Predictions/Model from the Dataset

Binary Classification: By leveraging the review_body, we aim to predict whether a given review has a positive (star_rating >= 4) or negative sentiment (star_rating <= 2).

Keywords indicators: To find if there are certain keywords that indicates customer giving higher ratings or lower ratings from star_rating and review_body

Trends over Time: With review_date and star_rating, ots possible to uncover any temporal trends in product feedback. This can elucidate if certain products or categories face varying sentiment over different times of the year.

Product Recommendation Prediction: Based on customer_id, product_id, and star_rating, recommend other similar products that the customer might like and predict whether the user follows a trend based on ratings.

Duplicate Review or Anomaly Detection: Identify potentially duplicate reviews by the same customer_id or for the same product_id. Also, detect anomalous reviews which might be spam or fake based on patterns in review_body and other associated features.