**Data Analysis on Amazon Customer Review**

|  |  |  |  |
| --- | --- | --- | --- |
| Shalini Dhamodharan | Priyanka Kumari | Vinuthna Ghattamaneni | Chandrika Pathakamuri |

1. **Introduction:**

Amazon is an internet-based enterprise mainly focused on selling products and media. Amazon sells a variety of products about 12 million different products in 2022, and Amazon sells $1.6 million worth of products a day. Amazon has business in over 22 countries, and therefore Amazon focuses on a wide range of customers.

Customers are the individuals who generate revenue for the business. Amazon has multiple sellers who sell similar products. Almost all sellers focus on maintaining good customer relationships and customer satisfaction. The seller’s motive is to ensure they are not losing their customers to their competitors. Understanding the customer's opinion and thoughts is needed to have a healthy relationship. Happy customers tend to make frequent and more business.

In the case of Amazon, there is no one-on-one contact between customers and amazon. Amazon is an e-commerce platform, and customer review is the way of communication between the seller and the customer. Analyzing customer reviews can provide valuable insights into their perceptions and opinions. In addition, examining patterns and trends in product reviews can help identify areas that require quality improvements. Amazon customer review is a public platform, and therefore reviews and overall ratings will affect the decision of new customers.

In this project, we would try to analyze the Amazon customer effectively to understand sales of products and ways to improve product quality to maintain overall ratings. Amazon is handling huge amounts of data; therefore, data analysis must be performed in a distributed file system. We would like to incorporate the “**Amazon Review- Tagging Negative Review in Amazon Product Review”** thesis project in understanding how negative reviews impact the overall rating and find means to improve product sales.

The “**Amazon Review- Tagging Negative Review in Amazon Product Review”** thesis helps to identify specific issues that customers complain about and categorize them into a topic, such as seller problems or product defects. This process tags customer feedback to categories as Product or Seller issues.

1. **Problem Statement:**

A product listed on the Amazon website may have zero to any number of reviews, each with a star rating ranging from 1 to 5. People tend to go over each product review before buying a product since reviews are feedback from other customers who bought the same product earlier. To be more specific people do not read all the reviews, they compare the number of ratings and overall ratings, hence it is necessary to have a good overall rating for a better product sale. The following are the problem statements we would like to address in this project with respect to product rating:

1. **Overall Rating**: Understand how overall rating is affected by the ratio of positive and negative reviews.
2. **Mapping Negative Reviews:** Incorporating the NLP model from “**Amazon Review- Tagging Negative Review in Amazon Product Review”** to categorize the negative reviews to study which category ( product issue or seller issue) of the issue is impacting the product's overall rating.
3. **Objective:**

The project aims to improve product quality and customer satisfaction. We are using Amazon Review Data (2018) from the University of California San Diego (UCSD). The project mainly focuses on identifying how and which labeled negative category affects the products’ overall rating and sales. The project deals with the following processes:

1. Map and reduce entire data based on each product and aggregate the number of reviews on the product and their average overall rating,
2. Label the Negative Review using **Amazon Review- Tagging Negative Review in Amazon Product Review”** machine learning model and further map and reduce the labeled output into each product and each category of labeled output.
3. Identify the at what percentage the product or seller negatively affects the overall rating of a product.
4. **Preliminary Literature Review for amazon:**

Large-scale information sharing in digital networks has been demonstrated in a number of academic publications to potentially aid buyers and sellers in online markets communicate product/seller quality and foster trust (Ba and Pavlou 2002; Resnick and Zeckhauser 2002; Dellarocas 2003; Chen and Wu 2005; and Chevalier and Mayzlin 2006). Resnick (2002), for instance, demonstrates how seller evaluations on eBay affect the likelihood of a sale, and Chevalier and Mayzlin (2006) discover how product reviews on Amazon.com affect book sales. According to Dellarocas (2005), deliberate manipulation of consumer evaluations can influence how much information buyers can learn from a particular review.

In order to examine how the content of reviews and the reputation of reviewers affect consumers' reactions, An Study of the Differential Effect of Reviews and Reviewers at Amazon.com uses customer reviews for books that are sold on the website.1 The percentage of persons who found a review useful was used to evaluate its quality. Examining the effects of spotlight reviews that appear on product pages . These "spotlight" reviews are distinguished from the other reviews and are displayed before other reviews, so they might have a more significant impact on book sales.

Also, earlier research has demonstrated that consumers may accept the supplier's reputation as a sign of the product's quality when it is difficult to verify the quality of the product (Resnick et al. 2000). Hence, customer purchasing decisions and subsequent product sales will be influenced by the perception of the product supplier or seller. This is particularly true for auction sites, where providing information on specific products is less useful because each "product" is actually a mix of a product and a seller, making it fundamentally unique.

We identified several significant distinctions with our technique based on the substantial investigation that was described in numerous scholarly journals. The objectives and methods used by Patel (2017), Woolf (2017), and Saumya, Singh, Baabdullah, Rana, and Dwivedi (2019) to analyze the Amazon product review dataset were very different from ours. 2

Patel (2017) worked to categorize positive and negative Amazon product reviews. One of the infant goods was the subject of a sentimental study he conducted. His method employed classical ML, which is not scalable and should not be utilized for datasets larger than Gigabytes, using Python, GraphLab, and S Frame. To forecast the binary rating and the 0–5 rating scale, Patel uses logistic regression and categorization.

Another study of a similar nature was carried out by Woolf (2017) using the Sparklyr platform, another scalable Big Data technique. 4.53 GB worth of data were used in his research.

In constrast to the above approaches , the approach we will be using would both big data and machine learning concepts. Tagging is a methodology commonly used in natural language processing (NLP) tasks like text classification using BERT models, which assists in understanding the structure and meaning of the text. By tagging text, it becomes possible to extract valuable insights and information from a large number of unstructured Amazon customer reviews. We start by tagging or labeling customer feedback to simplify data analysis and interpretation. An example of a product issue identified through tagging is "Dress quality was not good," which falls under the Quality Issue sub-topic of the Product Issue.

After analyzing the topics and sub-topics, we can gain a deeper understanding of the issues that the product or seller is facing. Product issues may include Quality Issues, Design Issues, and Product Description Issues, while Seller issues may include Product Authenticity Issues, Delivery and Return Issues, and Customer Service Issues. Similarly, "I like the dress, but packing is not done properly" is categorized under the Delivery and Return Issue sub-topic of the Seller Issue.

We then use Map-reduce concepts. Mapping would be done by an approach similar to Patel(2017) where the reviews will be mapped based on the ranges. This will involve separating the reviews to positive and negative. Our approach mainly focuses on the mapper -reducer pairs to associate the count of negative reviews that can be labelled to a specify product issue. This would help the manufacturers know what actually effects their consumer review.

1. **Methodology:**

The first task in the project is to map and reduce the entire data based on each product and identify the overall rating of the product which is the average of the star rating. In parallel filter the data into positive and negative reviews where star ratings 5, 4, and 3 are considered positive ratings and star ratings 1 & 2 are considered negative ratings. Integrate the machine learning model with negative ratings as input data and label the data as product or seller issue.

The next task is mapping and reducing the labeled data into each product and into each labeled category (Labelled category: Product, Seller). By calculating, the percentage of product and seller issue reviews and joining them with the overall rating we could compare and tell what percentage of the product's overall rating is affected by negative reviews.

Diagram

Description automatically generated

* 1. **Data Acquisition:**

Data Acquisition for this project involves obtaining a dataset with an overall rating and text review to create an automated system for sellers to analyze Amazon's products like Books, electronics, Clothing, Shoes, Jewelry, and more. The process of finding an ideal dataset with an overall rating and text review for negative reviews can be challenging, but Amazon's dataset provided by the University of California, San Diego (UCSD) offers additional features such as reviewerID, reviewerName, asin, and others, making it a suitable option. The data from UCSD has been collected from May 1996 to Oct 2018 and is preprocessed to include the necessary features for this thesis. The data dictionary of the dataset includes the following features UCSD from May 1996 - Oct 2018, the data dictionary is as follows:

* asin - Unique ID of the product being reviewed, string
* helpful - A list with two elements: the number of users that voted helpfully, and the total number of users that voted on the review (including the not helpful votes), list
* overall - The reviewer's rating of the product, int64
* reviewText - The review text itself, string
* reviewerID - Unique ID of the reviewer, string
* reviewerName - The specified name of the reviewer, string
* summary - Headline summary of the review, string
* unixReviewTime - Unix Time of when the review was posted, string

References:

[1]Samita dhanasobhon, Pei-Yu chen , Michael Smith,”An Analysis of the Differential Impact of Reviews and Reviewers at Amazon.com”, December, 2007.

[2] Jongwook Woo, Monika Mishra,”Predicting the ratings of Amazon products using Big Data”, November,202.