**Project Report**

**Data Analysis on Amazon Customer Review**

**Instructor**: Dr.Yalong Wu

**By**

Shalini Dhamodharan

Priyanka Kumari

Vinuthna Ghattamaneni

Chandrika Pathakamuri

TABLE OF CONTENTS

ABSTRACT

1. Project introduction
2. Usage of This Report:
3. Prerequisite knowledge for Beginners
4. Report Structure
5. Dataset
   1. Dataset Acquisition
   2. Dataset json to CSV
   3. Data Cleaning and Tagging
6. Amazon Review – Tagging Negative Review in Amazon
   1. Dataset and pre-processing
   2. BERT Model
7. Design and Implementation

CONCLUSIONS

Future Direction

References

ABSTRACT

This report describes the details about the design and implementation of analysis on Amazon customer reviews that helps in effectively understanding the sales of products and ways to improve product quality to maintain overall ratings. Through this project, we examined Amazon's dataset provided by the University of California, San Diego (UCSD).

**Keywords:** product issue, seller issue, Amazon review, mapper, reducer, rating, BERT

1. **Project introduction**

Amazon is an internet-based enterprise mainly focused on selling products and media. Amazon has multiple sellers who sell similar products. 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.

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 product review is publicly available on the platform, and therefore reviews and overall ratings will affect the decision of new or existing customers.

In this project, we would try to analyze the Amazon customer review effectively to understand sales of products and ways to improve product ~~quality~~ sell to maintain overall ratings. Amazon sellers are 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 to understand how negative reviews impact the overall rating and find ways 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. **Usage of This Report**

This report helps in understanding the need and necessity of analyzing customer reviews. The report talks about how analyzing customer reviews helps the product seller understand the root cause of negative reviews and work on product quality for better sales and customer satisfaction. The report is about the Amazon Review Analysis project which can be used to find what type of an issue/ problem within the product which in turn affects product sales resulting in sale downsize. The study may be extended to understand how linguistic cues can be used to analyze the effect of the customer on sales of the products.

1. **Prerequisite knowledge for Beginners**

**"E-commerce**" or "electronic commerce" is the trading of goods and services on the Internet. It is your bustling city center or brick-and-mortar shop translated into zeroes and ones on the internet superhighway. An estimated 2.14 billion people worldwide buy goods and services online. An e-commerce website is your digital storefront on the internet. It facilitates the transaction between a buyer and seller. It is the virtual space where you showcase products, and online customers make selections. Your website acts as the product shelves, sales staff, and cash register of your online business channel. Amazon is one of the most popular e-commerce websites used around the world.

Since there is no one-on-one contact between customers and Amazon, 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 general, there are two types of Amazon reviews: verified purchases (those who bought the product on Amazon marketplace) and non-verified purchases (those who made purchases from other sites). These can be as an overall rating of 5 stars, or the customer can provide descriptive opinions as well.

The aim to understand the reviews is achieved here by analyzing the patterns in them. The project tries to integrate a machine learning model based on Amazon Review – Tagging Negative Review in Amazon. The goal of this thesis is to create an automation system that requires system that automatically tags the sellers' review dataset. Data labeling is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels to provide context so that a machine learning model can learn from it. Firstly, the data -preprocessed tag negative review dataset will be divided into training and test/evaluation. Machine Learning is the use and development of computer systems that can learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. Secondly, A machine learning algorithm is required to train the model, which can automatically tag the review. The model will be trained on the training dataset. The whole process will find better accuracy for the tagging, which means that the accuracy model can predict the product category and find the best model. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

Hadoop is an open-source distributed processing framework that manages data processing and storage for big data applications in scalable clusters of computer servers. It's at the center of an ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining, and machine learning. This project uses Hadoop to map-reduce the tagged dataset.

1. **Report Structure**

In the further part of the report, we will elaborate on the flow of how we come to start with the data, clean it, visualize the cleaned data to understand the patterns in the ratings, use map-reduce functions to find the percentage of negative reviews, try to predict the issue type that resulted in negative rating and finally be able to analyze the patterns and the findings.

1. **Dataset**

**5.1 Dataset Acquisition:**

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 asin, helpful (number of helpful votes and total votes), overall rating, review text, reviewer ID, reviewer name, summary, and post time (Unix Time) for Clothing, Shoes and Jewelry. In our project we used following features which is shown in Figure 1:

• asin - Unique ID of the product being reviewed, string

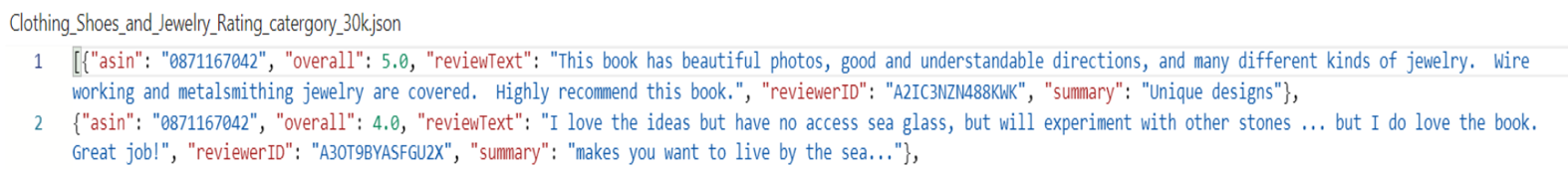
• overall - The reviewer's rating of the product, int64

• reviewText - The review text itself, string

• reviewerID - Unique ID of the reviewer, string

• summary - Headline summary of the review, string

The dataset looks like below:



*Figure 1: Clothing, Shoes and Jewelry Dataset*

*A picture containing screenshot, text, plot, number

Description automatically generated*

*Figure 2: The count of each type of item*

In Figure 2, The graph shows the number of products with the asin id. It has the asin id on the x-axis and the count of that product on the y axis. This can be used to find out the number of items that are of similar type.

Background pattern

Description automatically generated

*Figure 3: Graph mapping the reviewer with review category*

Using the Figure 3, we analysis the more detail of the asin id. The Sankey diagram above maps reviewer ids with the most frequent rating that they give. We have the reviewer id on one axis and the rating on the other axis. The graph does the mapping based on the most frequent rating given by the reviewer. This information can be useful to find the reviewers who give fake reviews based on the frequency of the review in a particular category. The graph is constructed with only few instances to reduce ambiguity. The same approach can be applied to the whole data to find out fake reviewers.

**5.2 Dataset Json to CSV**

The Clothing, Shoes and Jewelry dataset was in the json format, but for working on BERT model and Hadoop to map-reduce we require dataset to be in the CSV format so we converted dataset from the Json to CSV. The Json file shown in figure 4 and CSV file shown in figure 5.

*A black screen with white text

Description automatically generated with low confidence*

*Figure 4: Dataset Json*

*A close-up of a text

Description automatically generated with low confidence*

*Figure 5: Dataset in CSV*

*5.3* **Data Cleaning and Tagging:**

We visualize the dataset according to the overall ranking count of that product

A picture containing diagram, rectangle, line, plot

Description automatically generated

Figure 6: x-axis is overall rating and y-axis is count of that product

In Figure 6, The graph shows the number of products with the asin id. It has the asin id on the x-axis and the count of that product on the y axis. This can be used to find out the number of items that are of similar type.

After having dataset in the csv file, we required to tag the dataset into positive and negative so we can process the mapping and reducing the labeled data. We used some python code to segregate the positive and negative dataset.

#Print out of positive and negative

As this report emphases on finding the reason why the amazon seller getting low ranking we will use the negative review to pass through the on Amazon Review – Tagging Negative Review in Amazon, BERT Model and get the each review tag result which is product and seller issue.

A picture containing text, screenshot, font, line

Description automatically generated

*Figure 7: Tagging dataset using BERT Model*

1. **Amazon Review – Tagging Negative Review in Amazon**

This is thesis project which examines the impact of negative reviews on product ratings. Using Natural Language Processing (NLP) techniques, I have explored ways to enhance product sales. From these we use used the trained BERT Model and predicted our negative Clothing, Shoes and Jewelry Dataset.

**6.1 Dataset and pre-processing:**  
This thesis and our project dataset is same which help our project to help to analysis our negative dataset easily.

6.2 **BERT Model:**

Google developed BERT (Bidirectional Encoder Representations from Transformers) in 2018 as a powerful natural language processing model. It is a deep neural network that uses self-supervised learning to pre-train on a large corpus of text data, allowing it to learn contextual relations between words in a text[3, 4]. The transformer architecture is the foundation of BERT, which is designed to handle sequential data, such as text. It has multiple layers of self-attention mechanisms that enable it to capture long-range dependencies between words in a text. BERT is a bidirectional model that can take into account both left and right context when making predictions. This helps BERT understand the context and meaning of words in a sentence. BERT can be fine-tuned on specific downstream NLP tasks, such as sentiment analysis or named entity recognition, by training on a smaller, labeled dataset. Fine-tuning allows BERT to achieve state-of-the-art performance on a wide range of NLP tasks, even when training data is limited.

A picture containing diagram, line, origami

Description automatically generated

*Figure 8: BERT Implantation*

In Figure 8, BERT model is pre-trained on a large corpus of Amazon Negative Product Review data and can be fine-tuned for various natural language processing (NLP) tasks.

The BERT Model consists of several layers:

* **Input Layer**: The BERT Model takes inputs: ids, and mask. These inputs are encoded representations of the input text obtained using a tokenizer.
* **pre-trained Layer:** The pre-trained layer refers to a neural network pre-trained on large amounts of text data. When input text is fed into the Roberta model, the pre-trained layer processes it and produces a sequence of hidden states for each input token. The RoBERTa model has a deep architecture comprising multiple self-attention layers and feed-forward neural networks.
* **Fully connected Layer:** The fully connected layer is a linear layer that takes the output of the **pre-trained Layer** and maps it to the desired output dimensionality. In this case, the output dimensionality is two, corresponding to a topic tagging.

During the forward pass, the input ids, and mask are passed through the pre-trained layer, generating a hidden state sequence. Finally, the output of the dropout layer is passed through the linear output layer to produce the final output of the model, which is a vector of size two representing the probabilities of the two classes in the topic tagging task.

1. **Design and Implementation**

The data for the project is collected from Amazon’s dataset collected by the University of California, San Diego (UCSD) offers additional features such as reviewerID, reviewerName, asin, and others, making it a suitable option. As part of data cleaning, the columns such as unixReviewTime and review text are dropped as they are metadata for the project, and the data available do not have missing or null values. So, additional cleaning techniques were used but the dataset chosen is clean. Hence no values are dropped.

The data is split into positive and negative based on the overall star rating where star ratings 5, 4, and 3 are considered positive ratings whereas star ratings 1 & 2 are considered negative ratings. In data transformation, every instance is added with a new feature ‘ReviewTag’ which indicates if the review is positive or negative.

The Map and reduce concepts are used to calculate the count of positive and negative reviews for each product (“asin”). It helps us understand how each product has positive and negative reviews. Here is a sample output of Map - Reduce program:

A picture containing text

Description automatically generated

*Figure 9:* Map - Reduce program

The main goal of the project is to understand the negative reviews. The negative reviews are passed on to the BERT machine learning model which in turn classifies the negative reviews as product or seller issues.

The final task is mapping and reducing the labeled data based on each product and 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. The following flow diagram describes the project data flow:

Diagram

Description automatically generated

*Figure 10:* Map - Reduce program

**CONCLUSIONS**

Therefore, the project helps in understanding the impact of negative reviews on each product available in the dataset. The project gives an overall idea of how a negative review of a product affects product sales and the usage of linguistic analysis to categorize the root cause of negative reviews into Product issues and Seller issues, and how it could be used to improvising product quality. In the future, more linguistic analysis can use implemented in understanding the specific root cause of negative reviews.

**Future Direction:**

* Model: Input data formats can be expandable.
* Research for varied categorical problems under each of the issues.
* **The analysis of ratings**: This will help identify top-selling products and top reviewers who do frequent reviews. (Identifying the top-selling products and frequent reviewers can be accomplished by conducting an analysis of the ratings.)
* **Filtering Fake/False review:** we believe every customer might share only legit reviews, there might be certain people who might make false comments. By finding frequent reviewers, we could categorize the ones who make false negative reviews on all the products. (We believe that each customer provides only genuine reviews; there is a possibility that some individuals may submit fake comments. Identifying frequent reviewers makes it possible to categorize those who post false negative reviews across all products.)

**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.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.