Final Project Report (Group 1)

Amazon Products Review Analysis

1. Introduction

The way we buy things has been transformed by the Internet. In the retail e-commerce world of online marketplaces, it is impossible to try out things. Additionally, in today's retail marketing industry, new items are introduced on a daily basis [2]. Retailers must compete with one another's goods in order to preserve their market position. Taking client feedback into consideration is one of the most practical methods to do this [1]. Customer reviews are the most convenient approach to learn about the product's finest qualities as well as its shortcomings. Retailers want to obtain valuable evaluations as fast as possible throughout their decision-making process, thus they employ a grading system [1]. As a result, algorithms that can predict and interpret user ratings from text reviews are vital. Getting a feel of a textual evaluation as a whole might improve the consumer experience. It may also assist firms in increasing sales and improving their products by gaining a better grasp of their customers' demands [2]. As a result, client feedback may be utilized to incorporate the product's missing features into the updated version. Amazon's product website was considered for this project [2]. The name of the customer, date of the review, description and the rating were looked for and scraped.

2. Preliminary Literature Review

Since at least 1954, academic articles have studied sentiment analysis [3]. Early on, businesses had a big challenge in determining how to respond to frequently contradicting client input. Companies were able to deal with this issue because to the emergence of sentiment analysis, which allowed them to look at the overall tone of the remarks rather than the details and modify their answers appropriately [3]. A popular method involves first determining the feelings associated with each word, such as anger, contempt, fear, joy, sadness, or surprise. Natural language processing evaluates the meaning of each word to do this [3]. After that, machine learning and natural language comprehension are used to figure out what the text is truly about, such as what generated the emotions and how they are linked to one another. Since it was initially presented decades ago, sentiment analysis has gone a long way, and it's now something we utilize in our daily discussions [3]. Although this discipline has advanced significantly in recent years, there are still hurdles to be overcome if it is ever to be ideal.

3. Data description

As mentioned earlier, the data is scraped from Amazon's product website using Beautiful Soup. The links for the data source is (https://www.amazon.com/Apple-iPhone-Graphite-Carrier-Subscription/product-reviews/B08L5NHRWN/ref=cm_cr_getr_d_paging_btm_next_3?ie=UTF8&reviewerType=all_reviews&pageNumber=%d) [4]. The data is scraped and stored in a data frame. The data frame contains 121 rows and 4 columns.

The attributes that were scraped from Amazon's product website are 'Name', 'Date', 'Review' and 'Rating'. The data type of all the attributes, except 'Rating', is initially object. The attribute 'Date' will later be converted to datetime type.

4. Objectives and expected contributions

The goal is to classify positive and negative reviews and develop a model to predict user rating. We plan to address the following research questions:

- Question 1: What is the number of people who like and dislike the product? Which features of the devices are liked and disliked by the customers?
- Question 2: Which features should be included in the upgraded version of the device?
- Question 3: What will the user rating of the product be in the future?

Answering the above questions will have the following impact:

- Measure the public's interest in the product
- Predict the performance of the product
- Enhance upcoming products with features that lacked in the previous version
- Make future decisions more effectively

5. Methodology

The goal of our project is to classify positive and negative reviews and predict the user rating of the reviews which would help us get a better understanding of the product. The data is collected and analyzed using Python 3 with Amazon's product website being the data source.

We start by scraping the text data from Amazon's product website using Beautiful Soup or Selenium web driver. We proceed by analyzing the structure of the data and cleaning it if needed using various cleaning techniques like type conversion, elimination of duplicates, PCA and so on. It is then followed by visualization. Thereafter, sentiment analysis would be performed to classify reviews into positive and negative categories.

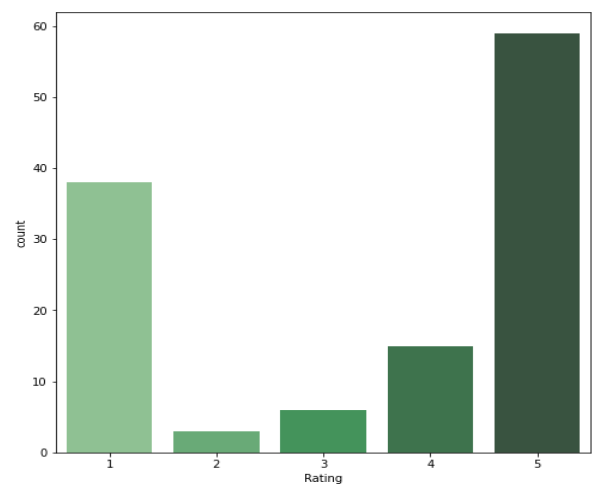
The most frequent words and specifications would be scraped and calculated which will help us point out the likes and dislikes of the product and customers feelings. This done by displaying the most frequent nouns, adjectives and adverbs from the positive and negative review list which was created and split from the original data frame. For instance, displaying the top 20 frequent nouns from the positive word list displays terms like (cameras,6), (screens,4), (pixels,3), etc. This indicates that people feel positively about features like the camera quality of the device.

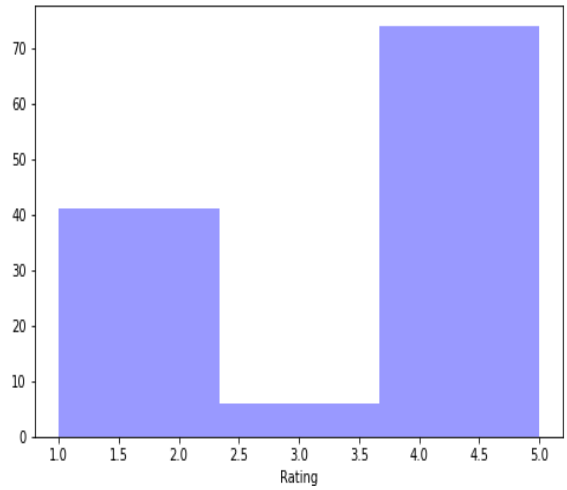
Similarly, displaying the top 20 frequent nouns from the negative word list displays terms like (headphones,2), (earpods,2), (diagnostics,2), etc. This tells us that the audio features of the device have room for improvement according to the customers. The 5-point summary and the data type of each attribute is also examined and displayed for future analysis.

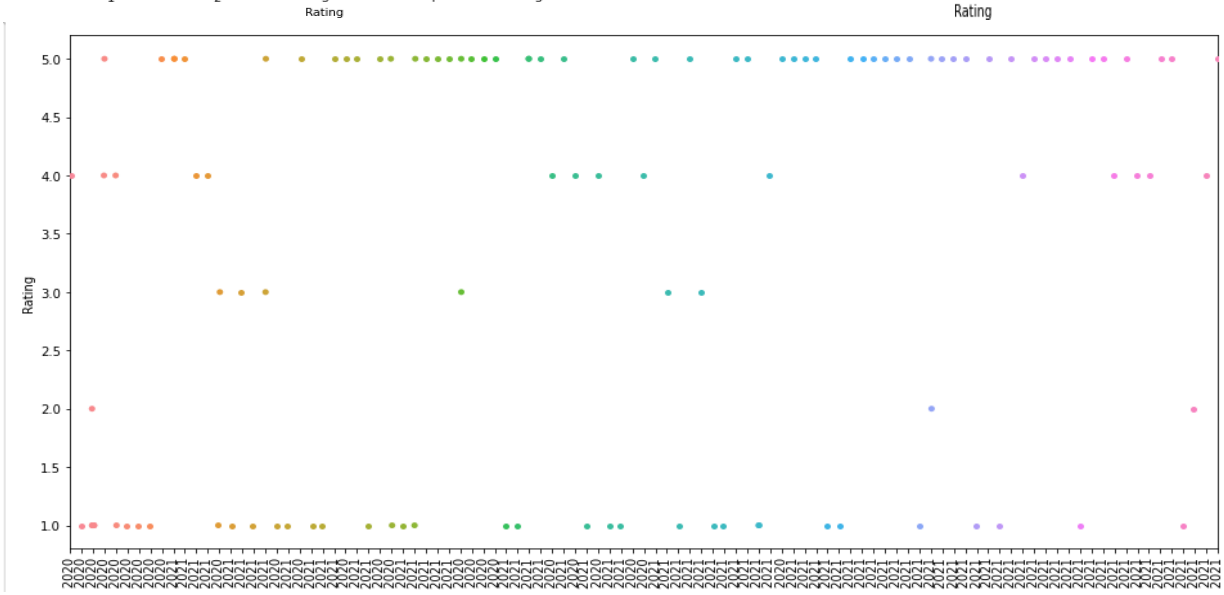
Using seaborn library for visualizations, we can see the distribution plot of the ratings. Also, a histogram of the ratings can give us an idea about the overall performance of the product. Using a

strip plot, we can see the trend of the ratings over a period. It can be inferred from the plot that as time progresses, there is an increase in the 5-star rating and decrease in the 1-star rating which was dense earlier.

The plots for the histogram, distribution and strip plot are inserted below in the respective order.







A word cloud visualizes word frequency that give greater prominence to words that appear more frequently in a source text. This style of visualization can help assessors with exploratory textual analysis by highlighting terms that appear often in a collection of documents. It can also be utilized at the reporting stage to communicate the most important ideas or topics. The below displayed word cloud represent the positive reviews list and the negative reviews list.

A set of appropriate algorithms like linear regression, Support Vector Machine (SVM), Naïve Bayes, which suit the problem statement will be selected. The purpose of this is to employ ensemble method i.e., creating multiple models and vote for the one with the best performance metrics. The data is validated using techniques like train/test split, k-fold cross-validation, nested cross validation, etc. Thereafter sampling techniques like stratified sampling and cluster sampling are used to draw valid conclusions.

Once the models are ready, we use various performance metrics like F1 score, recall, precision, and accuracy to evaluate the models. Once a vote is taken based on the performance of the models, we can answer the questions posed to us by visualizing the model and data.

6. Brief Project Schedule

The project is divided into the following parts with each group member responsible for the completion of it.

Task	Assignee
Data collection	Sanju Pandiath
Data preprocessing	Karthik Shankar
Account classification	Shivani Mogili
Topic Analysis	Soumya Jain
Hypothesis testing	Sanju Pandiath, Karthik Shankar
Research report writing	All

7. Conclusion

Leveraging the visualizations, the research questions can be answered. The histogram tells us that above 60% customers liked the product and gave a positive review. The word cloud indicates the features liked and disliked by the customers. The device camera, screen, pixels are some of the features liked by the customers and charger provision, audio quality are few areas where the

product could improve. The SVM model helps us infer that the future ratings and reviews would be positive and in favor of the product.

References

[1] https://www.commerce.ai/blog/amazon-product-review-analysis-the-ultimate-guide
[2]https://towardsdatascience.com/sentiment-analysis-and-product-recommendation-on-amazons-electronics-dataset-reviews-part-1-6b340de660c2
[3] Zellig S. Harris (1954) Distributional Structure, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520
[4]https://www.amazon.com/Apple-iPhone-Graphite-Carrier-Subscription/product-reviews/B08L5NHRWN/ref=cm_cr_getr_d_paging_btm_next_3?ie=UTF8&reviewerType=all_reviews&pageNumber=%d
[5]https://medium.com/@annabiancajones/sentiment-analysis-of-reviews-text-pre-processing-6359343784fb
[6]https://towardsdatascience.com/understanding-nlp-word-embeddings-text-vectorization-1a23744f7223
[7]https://www.betterevaluation.org/en/evaluationoptions/wordcloud#:~:text=Word%20clouds%20or%20tag%20clouds,frequently%20in%20a%20source%20text.&text=Most%20word%20cloud%20generators%20have,exclude%20common%20or%20similar%20words.
[8]https://scikitlearn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,than%20the%20number%20of%20samples.