Amazon Customer Satisfaction Sentiment Analysis

Avinash Bondalapati
abondalapati1@student.gsu.edu
Georgia State University
Atlanta, Georgia

Krishna Mani Teja Katragadda
kkatragadda1@student.gsu.edu
Georgia State University
Atlanta, Georgia

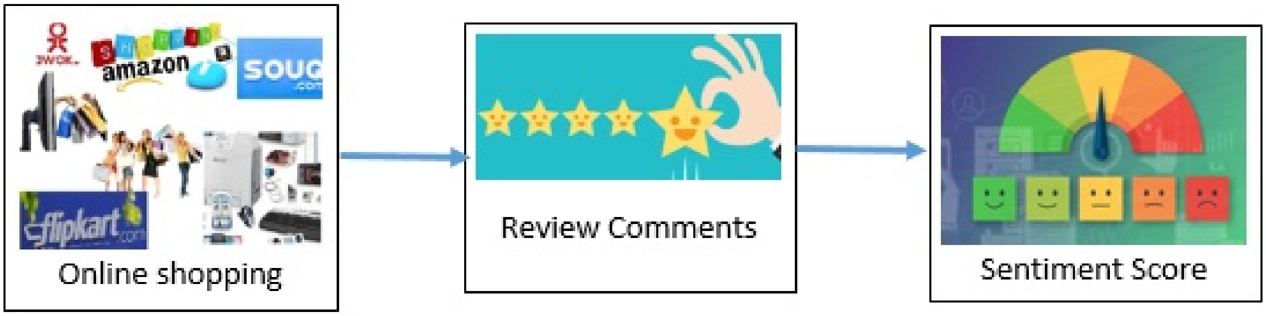


Figure 1. Entire workflow in a single image.

Abstract

This project will be helpful in identification of customer satisfaction across various product categories of amazon by analysing the individual product reviews in various product categories and then assigning a sentiment value for each of the product review. These individual product reviews are then combined to find the final sentiment value for the product category. By this customer sentiment value for each product category seller will be able to determine for which of the product categories customers are more satisfied and also in which categories additional steps needed to be taken to improve the customer satisfaction.

CCS Concepts: • **Computing methodologies** → **MapReduce algorithms**.

Keywords: datasets, spark , machine learning, sentiment analysis, amazon, customer satisfaction

1 INTRODUCTION

With the increase in the number of products sold on E-commerce increases, it becomes increasingly difficult for a single person to understand the data trends and the customer satisfaction simply based on teh ratings that a product receives. Amazon alone has number of product categories, with that the identification of customer satisfaction across the product categories is increasingly difficult when compared to a single product customer satisfaction. The data-set which we used for this project contains Amazon reviews and this cna be easily extended for other E-commerce websites. The major objectives of this project is to satisfy below criteria.

- Obtain a sentimental value for each review of the product by analyzing the text present in the review body of record using textblob.
- Combine the sentiment value for each product category by aggregating single product reviews for same category.
- Project the final sentimental value for each product in an easier way to understand.

While assigning a sentimental value to each of the product reviews requires an ML model to identify if a particular review is positive, negative or neutral. We use TextBlob to identify the review.Once the process of assigning a sentimental value is fixed then the reviews in each product category can be processed adn assigned a score. The intial plan was to have a single mapper job to do the work of assigning a sentimental value to a single product review. But after moving to spark we have used spark transformations to make the changes of finding the sentiment value for each product review and then aggregating the values.

The results which are generated by spark RDD transformations give us the total count of positive, negative and neutral reviews for a product in a specific category. This can be further used to generate graphs which are easier to make sense of.

2 BACKGROUND

As noted in the introduction, With increasing number of E-commerce websites the increase in sales of products increases the difficulty in interpreting customer satisfaction. This can be observed when the products are spanned across various product categories. This is same for any Online E-Commerce platform where a single seller involves sales in various product categories.

While the customer satisfaction is One such factor that we can obtain by analysing the data, we can also retrieve some other important factors that can be helpful in understanding the factors like average rating that a product received etc.

2.1 Product categories

Amazon has more than 50+ product categories. For this project we have selected around 12 product categories out of those 50 product categories to calculate the sentiment value and then find the customer satisfaction for each of these product categories. Following are the product categories that we have selected.

- Apparel.
- Automotive
- Books
- Camera
- Electronics
- Health Personal Care
- Home Entertainment
- Appliances
- Music
- Office Products
- Personal Care Appliances
- Watches

2.2 Data

While the number of product categories that we considered for this project is less, it can be noted that the number of product reviews in each product category is no less than 50,000 reviews, The total data that we considered is around 50 Gigabytes of memory and due to the memory restrictions on the KVM we have settled for around 4% of the data. Each product review has around 15 different fields which provide us with various information related to the system. The following are the list of fields in each product review record.

- marketplace
- customer_id
- review_id
- product_id
- product_parent
- product_title
- product_category
- star_rating
- helpful_votes
- total_votes
- vine
- verified_purchase
- review_headline
- review_body
- review_date

Out of the above 15 fields we make use of the review_body field to find the sentiment value for each product review and product_category to identify to which product category this product category belongs to. We also use data from fields product_id to identify the products with best customer ratings and customer_id to identify customer who provided with highest number of reviews.

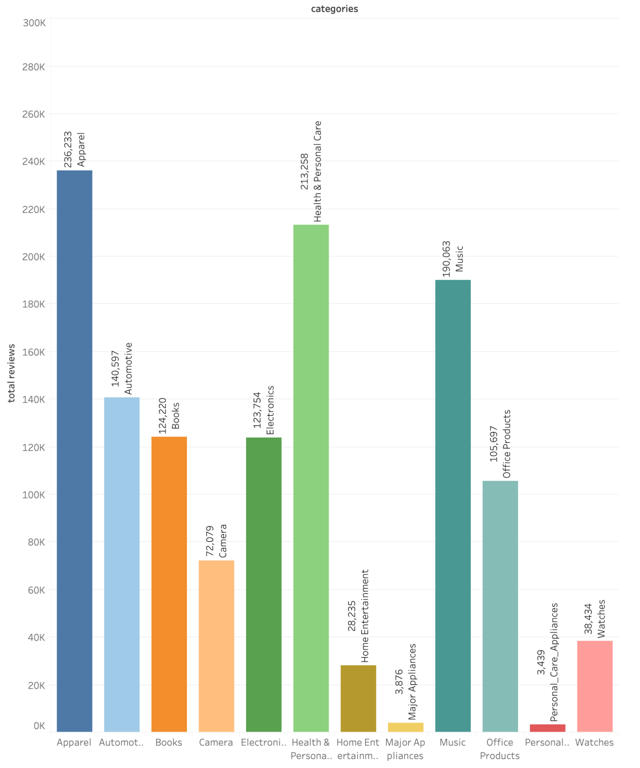


Figure 2. Reviews included in test dataset

3 Problem Statement

With the increasing sales on E-Commerce websites, it becomes difficult for an individual to understand the data and interpret if the customer satisfaction is positive or negative across the various product categories with multiple product reviews. So Analysing the product reviews will reflect only a single user perspective but by considering each user perspective(Satisfaction) we can draw conclusion on the customer satisfaction as a whole for a specific product category.

4 Design

The Initial design is to process each product category and then move to the next category in the second Iteration. A Basic view of the process is shown above where we have multiple iterations and in each iteration, we process a product category.

But with the limitation on KVM we have modified from our initial plan of having a separate file for each product category to having a single file with records merged from all product categories chosen.

We can improve or obtain closer to correct results by considering the "star_ratings" field in individual product review and assigning some extra weight to these if the customer has a verified purchase. Thus by reducing the influence of fake or made up reviews for a product. We can follow a simple rule multiplying the number of votes with the sentiment value and using the obtained result as the sentiment value for that product review.

5 Sentiment Calculation

While there are many different methods and algorithms which can be used to calculate the sentiment value. All these can be broadly divided into three types of methods.

- Rule-Based Methods that perform sentiment analysis based on a set of manually crafted rules.
- Automatic Methods that rely on machine learning techniques to learn from data.
- hybrid Methods that combine both rule-based and automatic approaches.

While the machine learning method looks promising its very difficult to setup and then train the model with the datasets. Since we want to highlight the BigData part of our project we wanted to move from the Machine learning task of creating a model, training it and then testing it. So we have explored for our options to calculate sentiment value using an existing opensource APIs and we found the following two.

- Natural Language Toolkit(nltk)
- TextBlob

5.1 Natural Language Toolkit(nltk)

nltk is a python package that is opensource and is designed mainly for the language processing and sentiment analysis. nltk is the basis for many of the available python packages which calculate sentiment analysis including TextBlob. while nltk has a lot of functionality built into it the programming interface is still not that intuitive. It requires setup of multiple additional packages and download of files for it to successfully work this makes it very difficult when you want to export or share your project with others to experiment.

5.2 TextBlob

Textblob is also a python package that is similar in a way to nltk but it has a much more refined API. TextBlob is based on nltk thus having all the functionality of nltk but in a refined way. the following are a list of things that TextBlob provides.

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification
- Tokenization and much more.

Of all the above mentioned usecases,we will mainly focus on sentiment analysis. TextBlob has two different analyzer built by default they are

- Pattern Analyzer
- Naive Bayes Analyzer

The default analyzer which is used by TextBlob is pattern analyzer. It provides a value within a range of [-1.0 to 1.0]. -1.0 being a completely negative review and 1.0 being a positive review. This provides us a way to easily calculate the sentiment value of each review.

6 Algorithm Design

6.1 Architecture

The dataset explained earlier is saved to local file system in KVM. Data from local system is loaded to spark engine using pyspark API (python). Spark sends data and tasks i.e map and reduce to worker nodes to do the parallel processing. Spark sends output back to the local system and then output is fed to matplotlib and tableau for visualizations and the results are used for decision making.

6.2 Data Pipeline

Data collected from different sources are saved locally and fed to spark. Spark transmits data into RDDs to its worker nodes.All the worker nodes execute their tasks and send the sorted output back to the local FS. Outputs are visualized in tableau for decision making.

for each of the review record we try to get the sentiment value using textblob explained earlier and then we try to find the count of positive, negative and neutral reviews for each of the product category which will give us the total counts which can be used to find the customer satisfaction.

7 Results

The output of our spark project is a text file which contains the total count of positive, negative and neutral reviews for each product category. we also try to find the top 20 products with positive reviews and top 20 products with negative reviews.

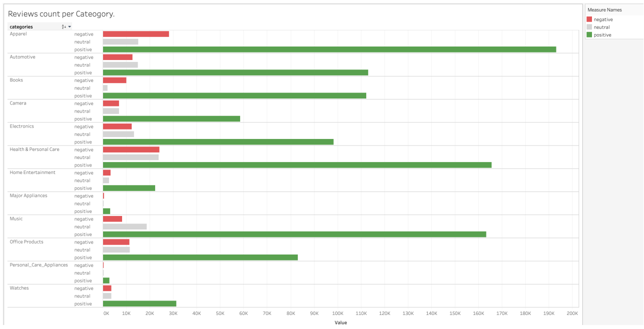


Figure 3. positive, negative and neutral reviews distribution in each category.

7.1 Reviews distribution in different product categories

From the below figure we can observe that some categories has more positive reviews and then there are some product categories with more negative reviews. This provides us an easier way to quickly glance at the test results and find or draw a conclusion on the product category customer satisfaction. If there are higher than usual negative reviews count than other product categories then we can conclude that the satisfaction is lower in that particular category and if the negative count is lower then we can say that the customer satisfaction is higher in this category.

7.2 Top 10 products with positive feedback

On utilizing the product_id we can identify which products has the highest number of positive reviews and lowest number of negative reviews. This helps us to identify the products that are doing good and also products with highest negative reviews.

On observing the top 10 products graphs we can observe that the product with top positive feedback is also in the top 10 negative feedback products this is because of the total reviews this product received. The higher the number of the negative reviews the higher the product is listed. In order to normalize this we can try to divide the total number of comments that particular product received by positive or negative reviews count so that we can get a better picture considering total reviews.

from the data set considered Google chrome cast is considered as the product with highest positive feedback, but due to its high total reviews it got higher negative reviews as well when compared with other products. This has placed it in the second place of products with high negative reviews.

7.3 negative to positive reviews count ratio

The ratio of negative to positive reviews will provide us with a more in-depth and meaningful insight where we can understand how the counts of positive reviews and negative reviews are across various product categories. This gives us a measure how how many positive reviews are for a product for each of its negative product. The higher the number the better the product is. From the dataset which we have chosen we can see that the music and books are the top two categories with highest ratio.

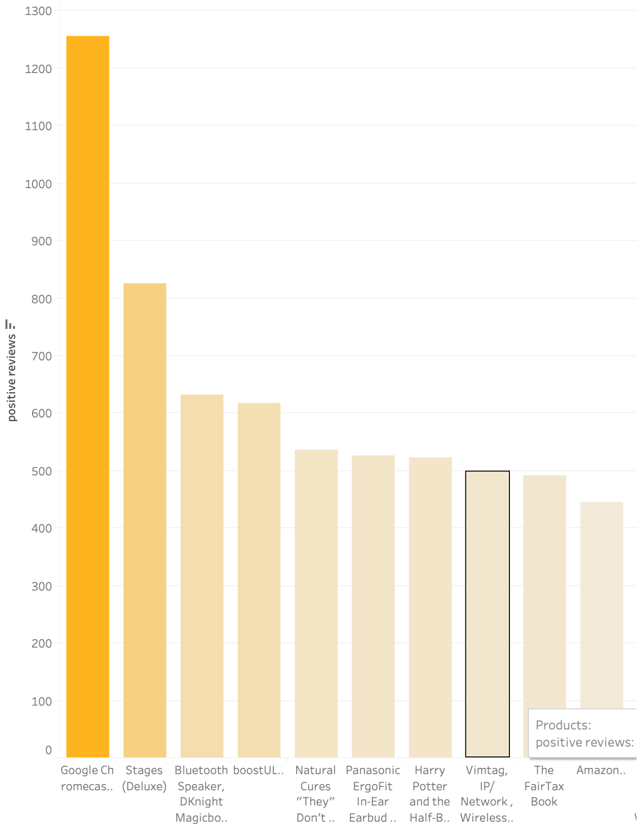


Figure 4. Top 10 products with positive feedback

7.4 Top 20 reviewers

With the dataset which we have chosen it is easier to identify or tag a particular review to a customer. Thus by making it possible to get a total number of reviews a customer has given. This makes it easier to identify which reviewer is actively providing reviews for the products he purchased. The following is a figure which shows you the top 20 customers who have provided the highest number of reviews.

from the table listed below we can see that customer with customer_id as 15536614 has given more reviews than any other customer in our dataset. we can also see the number of negative reviews he has given, all these values will provide information on how a customer is reacting to products across various product categories. This is helpful in understanding a customer and then improving the suggestions.

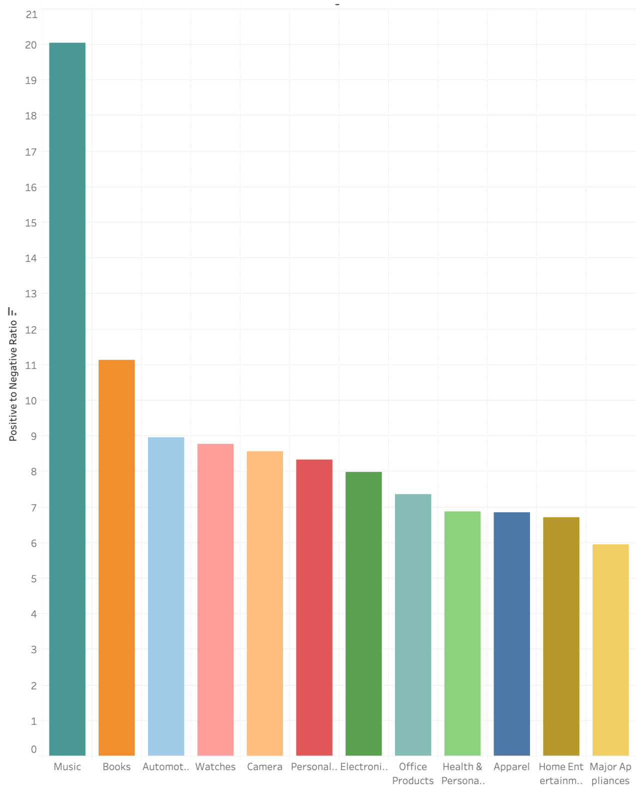
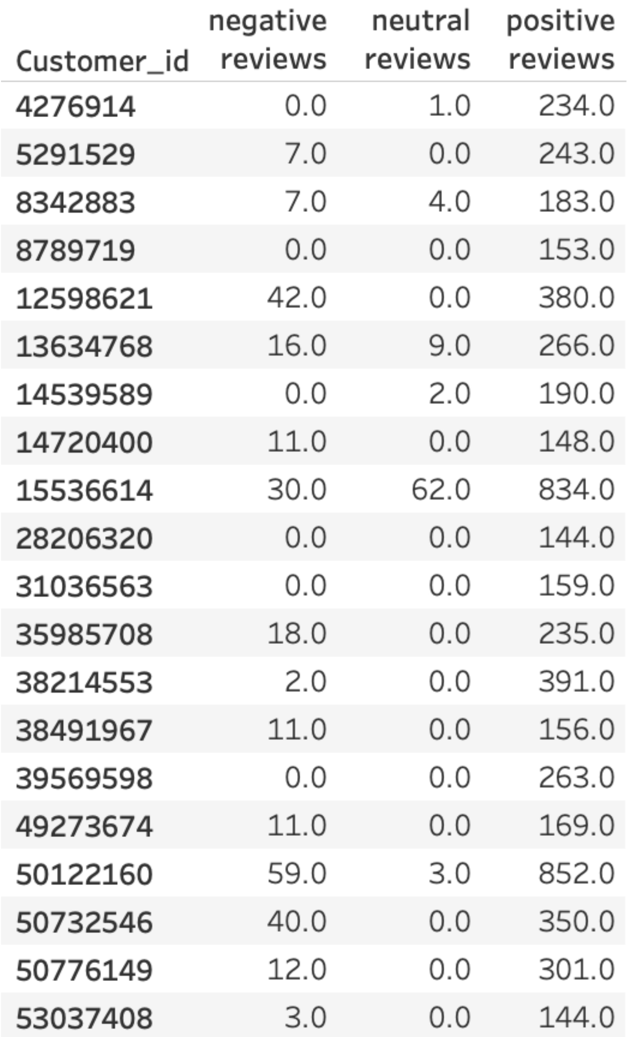


Figure 6. negative to positive reviews count ratio



Customer_id	negative reviews	neutral reviews	positive reviews
4276914	0.0	1.0	234.0
5291529	7.0	0.0	243.0
8342883	7.0	4.0	183.0
8789719	0.0	0.0	153.0
12598621	42.0	0.0	380.0
13634768	16.0	9.0	266.0
14539589	0.0	2.0	190.0
14720400	11.0	0.0	148.0
15536614	30.0	62.0	834.0
28206320	0.0	0.0	144.0
31036563	0.0	0.0	159.0
35985708	18.0	0.0	235.0
38214553	2.0	0.0	391.0
38491967	11.0	0.0	156.0
39569598	0.0	0.0	263.0
49273674	11.0	0.0	169.0
50122160	59.0	3.0	852.0
50732546	40.0	0.0	350.0
50776149	12.0	0.0	301.0
53037408	3.0	0.0	144.0

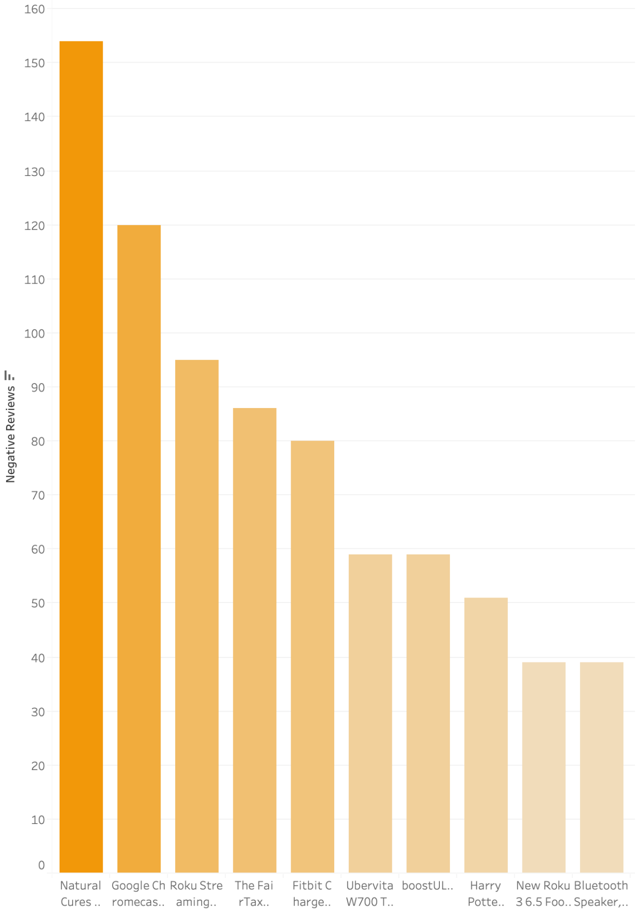


Figure 5. top 10 products with negative feedback

8 Conclusion

the project scope was to successfully understand and project customer satisfaction across various amazon product categories with the help of above graphs we can easily achieve the project goal. Even though the data set selected for this project is around 4% of the data procured. The insights we can get form these dataset is beneficial in understanding how different product categories progress and how customers feel across various product categories.

References

[1] Hota and S. Pathak. Knn classifier based approach for multi-class sentiment analysis of twitter data. In International Journal of Engineering Technology, pages 1372–1375. SPC, 2018

[2] Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. Swarthmore College, 2013

[3] Dataset - https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

[4] Tableau Integration - https://docs.databricks.com/integrations/bi/tableau.html#tableau-20193

[5] Journal paper - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

[6] TextBlob - https://textblob.readthedocs.io/en/dev/