Big Data Analytics on Amazon Product Reviews

Vaibhav Joshi
vj3470@rit.edu

Satyanarayan Iyengar
si7849@rit.edu

Amritha Venkataramana
axv3602@rit.edu

ABSTRACT

Product reviews reveal customer sentiments helping the manufacturers decide what is required to make the product a success. Depending on the reviews provided by the users, the product is classified as good or bad and improvements can be made accordingly to cater to the requirements of the customers. For the purpose of this term project, we have chosen to analyse review on Amazon Products from datasets obtained via two different sources.

1. INTRODUCTION

As the world is moving more and more towards online retail, the companies involved have been collection enormous amount of data. Analysing this data can generate numerous insights that can be subsequently used to improve customer service, recommendations, etc. The goal of this project is to understand and analyze the various Amazon Reviews with the help of different visualization methods and data mining models. will help in learning about the various trends and insights present in the reviews. Additionally, analysing text present in the reviews will also aid in determining the various user groups and rating sentiments.

We will be performing the above mentioned tasks on a combined dataset consisting of about 120 K instances and 15 attributes. The dataset is combined from two different sources stated below :-

1. **Amazon Customer Reviews Dataset**[1]
   This dataset is a collection of customer reviews written in the Amazon.com marketplace. It also has metadata associated with the products being reviews. This dataset is made available to facilitate research and help understand peoples sentiments towards a product.

2. **Amazon Product Reviews Data**[2]:
   This dataset is divided into product reviews dataset and product metadata dataset. The reviews dataset includes information about a review like ratings, text and helpfulness. The product metadata dataset includes information about the products like category, descriptions, price etc.

For the purpose of our project we will be limiting our scope to Books reviews. Amazon dataset has reviews for dozens of categories but Books has one of the highest reviews and ratings among all. We felt it captures the essence of the entire dataset and hence we chose to go ahead with it.

The datasets from both the sources had overlapping columns such as Product ID, Customer ID, Votes, etc. We used these columns to merge the two datasets.

2. DATA PRE-PROCESSING

Data processing constitutes about 80% of a data mining task. It is essential because more often than not the data gathered is not in a proper format. Datasets often contain large amount of attributes and many among those are not useful for analysis. Thus, we need to filter out the attributes that do not contribute significantly to our analysis. Furthermore, the data often contains missing, null and empty values that need to be handled.

The following pre-processing tasks were performed on the two datasets:

1. **Data Loading and Formatting**
   The AWS dataset had a lot of formatting issues. There were missing tabs, extra commas, missing quotation marks, etc. This lead to import and read errors while using pandas. Thus, a fair amount of time was spent in properly formatting the file to make it error-free. For the Stanford dataset, the file was a JSON file compressed in gzip format. In order to properly load the dataset into Python, a boilerplate code provided on the Stanford dataset page was used.

2. **Data Conversion**
   For this project, the two datasets needed to be merged. One of the datasets was in a JSON format while the other was a Tab Separated file. In order to work on this dataset, a common format was to be found. Thus, both files were converted to SQL format for future use.

3. **Dropping Unnecessary Columns/Attributes**
   There was a difference in columns in both datasets. For the purpose of our analysis, we were concerned with only a specific set of attributes. Thus, the attributes that were deemed to be insignificant to our analysis were dropped. For instance, the verified_purchase vine and image attributes in the AWS dataset were dropped while the image_url was dropped from the Stanford Dataset. Also, since the focus of this analysis is only on Books, the product category attribute was deemed to be insignificant and hence it too was dropped from both datasets.

4. **Handling Missing Values**
   There were a few instances of missing and null values in both the datasets. Since the instances were mostly text-based attributes, it did not make sense to replace

them with the mode. There were a few records which had an N/A value for reviewText. Hence, those records were dropped. The numeric missing values (like rating) were replaced with the mean of the entire column.

3. DATA MANAGEMENT

```
1 •  SELECT * FROM amazon.review;
2
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

customer_id	product_id	review_body	helpful_	overall_	review_date	review_headline	star_rating
10005833	B002A48...	Well deserved to be a ...	0	0	2015-08-2...	Five Stars	5
10008659	0671733354	Perfect.	0	0	2015-08-2...	Five Stars	5
10010780	1502525496	This book kept my inter...	1	1	2015-08-3...	Love it love it lo...	5
10011040	0881030368	Enlightening....eye op...	0	0	2015-08-3...	Must read for e...	4
10012167	1508973458	JR Harding has been a ...	0	0	2015-08-3...	WOW	5
10014050	1579653510	we love to cook. but t...	2	3	2015-08-3...	we love to cook	3
10014149	151482096X	Great book. The storie...	0	0	2015-08-2...	Love in Mistleto...	5
10014701	0692406735	Received a copy of her...	0	0	2015-08-2...	In my reading s...	5
10015224	0061258474	Stupid Wars is a non-fi...	0	0	2015-08-2...	Impressive Tak...	4
10015224	0758203993	I purchased this book ...	0	0	2015-08-2...	Extremely Hard...	1
10015224	1564144844	As an avid history-buff...	0	0	2015-08-2...	Some of my fav...	4
10016045	0800721985	I received a copy of thi...	0	0	2015-08-2...	Choppy action	3
10016045	085721604X	I received a copy of thi...	0	0	2015-08-2...	Very thorough, ...	4
10016708	1608193942	In the same way that ...	3	3	2015-08-2...	All it takes is a li...	5
10017695	1477816208	*I received a free cop...	1	1	2015-08-3...	* I really enjoy...	4
10017822	0987650408	I've been using this bo...	0	0	2015-08-3...	Cautiously opto...	4
1001811	0399536213	It was everything that ...	0	0	2015-08-2...	Books	5
10018115	0887431488	Great for review.	0	0	2015-08-3...	Five Stars	5
10018115	0938256343	Great for review.	1	1	2015-08-3...	Five Stars	5
10018115	0938256467	Great for review.	0	0	2015-08-3...	Five Stars	5
10018207	0991858891	Great book.....gr...	0	0	2015-08-3...	Five Stars	5
10018207	1493010042	Great book.....great s...	1	1	2015-08-3...	Five Stars	5
10018887	0692289771	Good format... easy to...	2	2	2015-08-3...	Great guide an...	5
10020112	1514273934	Disappointed. Did not l...	0	1	2015-08-2...	Disappointed. D...	1
10020322	1451666179	Best book ever.	0	0	2015-08-2...	Five Stars	5

Figure 1: Data Management in MySQL

Data management comprises of creating and managing databases. It is achieved via a Database Management System (DBMS). This system enables end users to create, read, update and delete (CRUD) data stored in a database. Data management is also useful as it keeps the data separate from the analysis. For our task we have used MySQL which is an extremely popular open-source DBMS.

After performing data pre-processing, the unnecessary attributes are discarded and a data base schema is prepared from the remaining attributes. The database table contains as column names the attributes of the data. A Python script handles the entire process starting from database connection to performing CRUD operations. Figure 1 shows a sample fetch query performed on the system.

4. DATA MINING

Data Management is helpful in obtaining a quick overview of the data. It is useful in web-applications and query-based environments. It can execute complex queries however it cannot yield insights and it is difficult to perform visualizations. Thus, data mining is needed for predicting, modeling and visualizing data. Data mining is used across all domains including retail. Customer reviews can be mined to generate trends as well analyse past history to improve future recommendations.

Cross Industry Standard Process for Data Mining (or CRISP-DM) is the most popular technique for Data mining tasks. It consists of the following steps:

- Business Understanding
- Data Understanding
- Data Preparation
- Data Modeling
- Data Evaluation
- Deployment

Business Understanding discusses about the development domain and the goals associated with the analysis. The Abstract and Introduction sections describe our motivation and goals behind choosing the customer reviews domain. Data preparation involves data pre-processing tasks such as data cleaning, formatting, conversion. Section 2 covers data pre-processing.

Data Understanding is a key factor in this process. It gives a good overview of the data as well as the relationship between the different attributes. Data Modeling involves tasks such as regression, classification, clustering which help in achieving the business goals and initiatives. Data Evaluation is about testing and validating the models implemented.

4.1 Data Understanding

Data understanding gives a deeper analysis of the data from an analytical perspective. For the Amazon dataset, attributes such as helpful votes, ratings, review year, etc. are useful in providing insights about trends in the reviews. The following series of visualizations explore the relationship between different attributes via pairwise comparisons, trends and timeline charts.

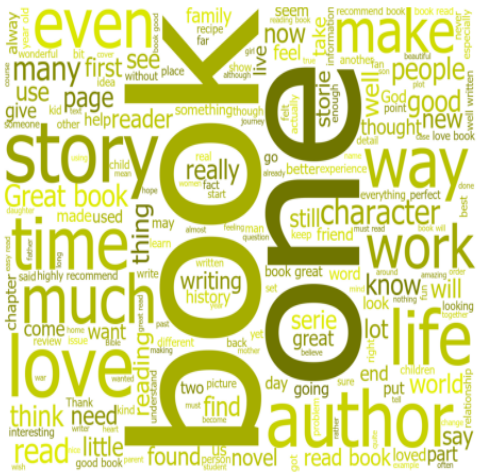

Figure 2: Word cloud for all reviews

4.1.1 Word Clouds for Reviews

Since this a review dataset, it is naturally composed of a lot of text attributes such as review headlines and contents. Here we have plotted different word clouds which are helpful in analysing the most common words present in the reviews (Figure 2).

Figure 3: Word cloud for positive reviews (star rating >=3)

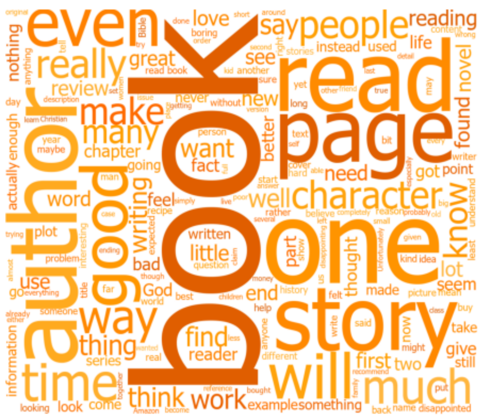

Figure 4: Word cloud for negative reviews (star rating < 3)

Furthermore, we have also plotted wordclouds for negative reviews (Figure 4) and positive reviews (Figure 3).

4.1.2 Average Book Ratings Over the Years

Figure 5 plots the average star rating for book products from 1997 - 2015. We observe that the average ratings have been fairly high for most of the years

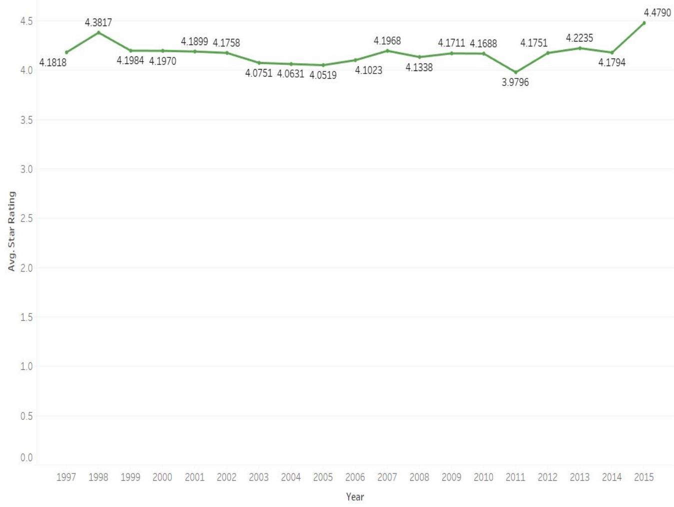

Figure 5: Average Ratings over the years

4.1.3 Total Book Reviews Over the Years

Figure 6 plots the count of total reviews for each year from 1997 - 2015.

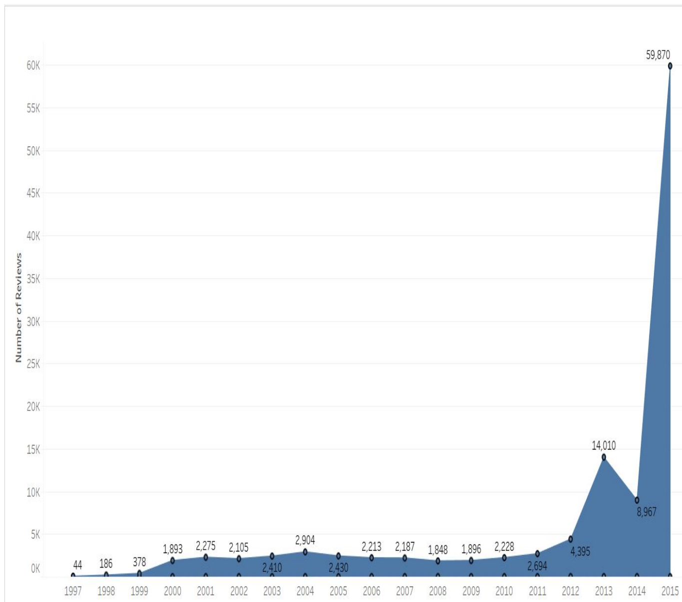

Figure 6: Average Reviews over the years

4.1.4 *Pairwise comparison of helpful votes and overall votes*

Helpful votes in Amazon are given to reviews which enable other purchasers to get a good idea about the book and influence their decision on whether to buy the book. They are a more important metric than overall votes. Figure 7 visualizes the relationship between overall and helpful votes over the years.

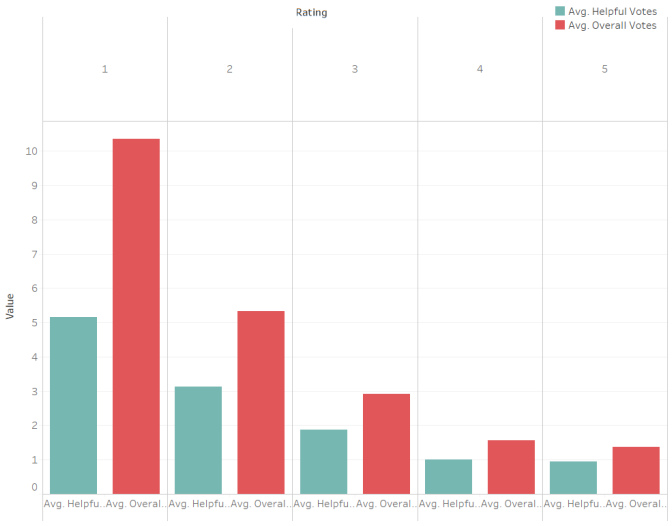

Figure 7: Helpful Votes vs Overall Votes

4.1.5 *Average number of helpful votes per rating*

This is another useful pairwise comparison between the helpful votes and ratings. Negative reviews often are considered to be more helpful in retail compared to positive ones.

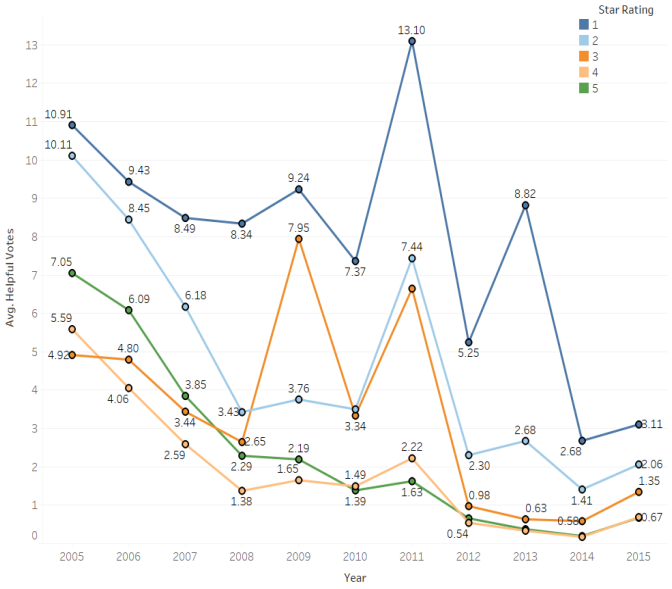

Figure 8: Average helpful votes per rating

This trend is justified by Figure 8. We observe that reviews with low ratings have high number of helpful votes.

4.1.6 *Frequency of star ratings for each year*

Figure 9 shows a bar chart where each bin describes the frequency distribution of star ratings for each year.

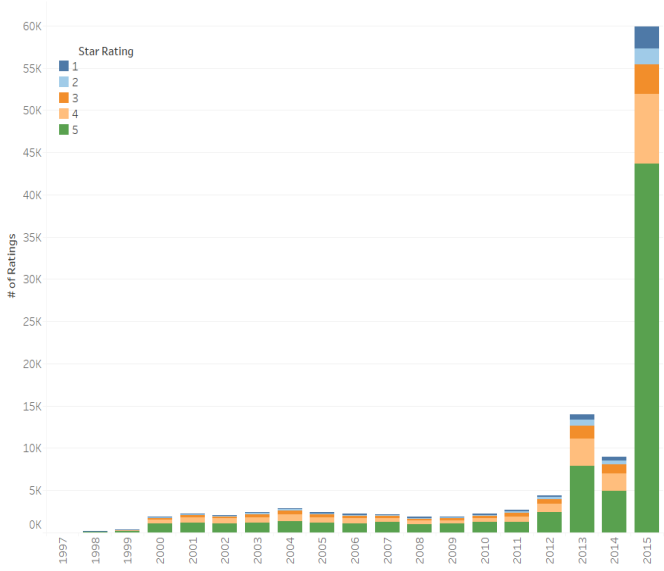

Figure 9: Frequency of each star rating for each year

4.2 Data Modeling

Data Modeling is the phase where core data mining algorithms are applied to achieve analytical results. It comprises of techniques such as classification, regression, clustering and mining. Here we have performed classification and clustering as part of our analysis

- Classification : It is a supervised learning approach in which we train a model to learn the instances of a dataset. This model is later used to classify incoming new instances. There are various types of classification algorithm ranging from rule-based algorithms such as Naive-Bayes and Decision Trees, to hyperplane based algorithms such as Support Vector Machines(SVMs). Here we have performed classification using SVM.

  – SVM :
  It is a popular classification algorithm found to run well on small datasets. SVM uses kernel functions. These functions can be linear or non-linear. The algorithm works by making a dividing hyperplane which separates the data in two classes.When using SVM, a large separation between the classes is useful.

  We have implemented the SVM classification to classify a review as helpful or unhelpful. A review is said to helpful if the ratio of helpful votes to overall votes exceeds 50%. We have created a class label for the same. We split the data into training and test set in a ratio of 80/20.

We use the review_body text attribute to classify a review. Feature vectors are created which are then fitted into the SVM. The SVM is used with a non-linear Radial Basis Function (RBF) kernel. The training and testing accuracies are determined and evaluation is performed

- Clustering :
  Clustering is a data mining process of grouping a set of data points in a way that points in the same cluster are more similar to each other than to those in other clusters. Clustering is useful in textual data and hence we have used it for clustering customer reviews. Divisive Clustering (k-means), Agglomerative Clustering, and density based clustering are among the popular clustering techniques. Here we have employed Birch which is a hierarchical clustering technique.
  - Birch:
    Birch is an unsupervised clustering algorithm suited for clustering large-scale datasets. We have implemented Birch clustering to see if the reviews can be grouped into meaningful clusters.

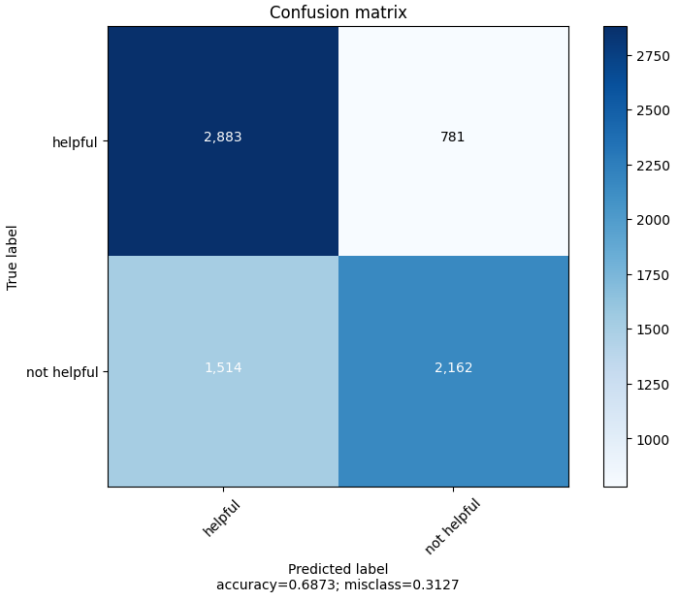

Figure 10: SVM Accuracy

4.3 Data Evaluation

In this section we evaluate the models that were implemented in the previous section.

- SVM Evaluation :
  The training and testing accuracies for SVM were found and a confusion matrix was plotted as below (Fig. 10) showing the results. We see that SVM achieved around 68% accuracy for the testing data with a misclassification rate of about 31%.
- Birch Evaluation :
  Birch grouped the reviews into only 3 clusters. Majority of the reviews only talk about the rating given by the reviewer. They do not convey any information about the book. These reviews are grouped into a single large cluster. The other 2 clusters are about a particular topic such as food, crime, etc. One of the clusters found in the analysis was a cluster of reviews on recipe books. This contained food based words such as vinegar, soda and baking. (Fig. 11).

Figure 11: Cluster containing food related reviews

5. CONCLUSION AND FUTURE WORK

Data analysis on customer reviews is extremely useful for companies as it helps to improve customer experience and increase sales. We have performed classification and clustering on the product reviews via SVM and Birch respectively. We achieved a reasonable accuracy of 68% with SVM. Birch clustering shows that most of the reviews are very generic and convey little information about the product. We found a few topic relate clusters such as food. We have also performed visualizations on the data and explored the relationship between different attributes. Overall, via this project, we have captured the essence of an industry level data mining project.

Our analysis can be further extended by incorporating different product categories in addition to Books. Also, there is an additional Metadata dataset in the Stanford corpus which can be leveraged to perform a much deeper analysis such as user identification. A decision tree can be designed to classify a particular user based on his reviews. This can be achieved using the attributes of product id and customer id along with review text.

6. DATA SOURCES / LINKS

1. **Amazon Reviews Dataset - AWS**
   `https://s3.amazonaws.com/amazon-reviews-pds/readme.html`
2. **Amazon Product Data - Stanford**
   `http://jmcauley.ucsd.edu/data/amazon/`