Analyzing Amazon Product Reviews using Sentiment Analysis

Report submitted in fulfillment of the requirements for the Data Mining Project of

Third Year B. Tech.
in
Computer Science and Engineering

Submitted by

Roll No	Names of Students
18074013	Nishant Mittal
18074011	Mudit Bhardwaj
18075006	Aditya Patidar
18075014	Ayush Damele

Under the guidance of
Dr. Bhaskar Biswas

Department of Computer Science and Engineering
Indian Institute of Technology (BHU) Varanasi
Varanasi, Uttar Pradesh, India – 221005

Semester V

Dedicated to

Our beloved parents, teachers,

Our laptops and Varanasi

Declaration

We certify that

1. The work contained in this report is original and has been done by ourself and the general supervision of our supervisor.
2. The work has not been submitted for any project.
3. Whenever we have used materials (data, theoretical analysis, results) from other sources, we have given due credit to them by citing them in the text of the thesis and giving their details in the references.
4. Whenever we have quoted written materials from other sources, we have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi
Date: November 21, 2020

Nishant Mittal (18074013 - IDD)
Mudit Bhardwaj (18074011 - IDD)
Aditya Patidar (18075006 - B.Tech.)
Ayush Damele (18075014 - B.Tech.)

Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Department of Computer Science and Engineering
Indian Institute of Technology (BHU) Varanasi
Varanasi, INDIA 221005.

Certificate

This is to certify that the work contained in this report entitled **"Analyzing Amazon Product Reviews using Sentiment Analysis"** *being submitted by* **Nishant Mittal (18074013)***,* **Mudit Bhardwaj (18074011)***,* **Aditya Patidar (18075006)** *and* **Ayush Damele (18075014)***, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of my supervision.*

Dr. Bhaskar Biswas
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Place: IIT (BHU) Varanasi
Date: November 21, 2020

Acknowledgments

It is a great pleasure for us to express respect and deep sense of gratitude to our supervisor Dr. Bhaskar Biswas, Associate Prof., Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, for his wisdom, vision, expertise, guidance, enthusiastic involvement and persistent encouragement during the planning and development of this work. We are indebted and grateful to the Almighty for helping us in this endeavor.

Place: IIT (BHU) Varanasi
Date: November 21, 2020

Nishant Mittal
Mudit Bhardwaj
Aditya Patidar
Ayush Damele

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Customer Experience (CX) is the key to business success.Now, more than ever, it's key for companies to pay close attention to Voice of Customer (VoC) to improve the customer experience. By analyzing and getting insights from customer feedback, companies have better information to make strategic decisions, an accurate understanding of what the customer actually wants and, as a result, a better experience for everyone.

But, what are customers saying about the brands? How can we provide them a better experience? This can be done easily with the help of sentiment analysis.

Sentiment analysis is the automated process of understanding the sentiment or opinion of a given text. This machine learning tool can provide insights by automatically analyzing product reviews and separating them into tags: Positive, Neutral, Negative.

By using sentiment analysis to structure product reviews, we can:

1. Understand what customers like and dislike about several product.
2. Compare similar product reviews of various companies.
3. Which products should be kept, dropped from Amazon's product roster (which ones are junk?).
4. Get the latest product insights in real-time, 24/7.
5. Save hundreds of hours of manual data processing.
6. We can predict scores for reviews based on certain words.

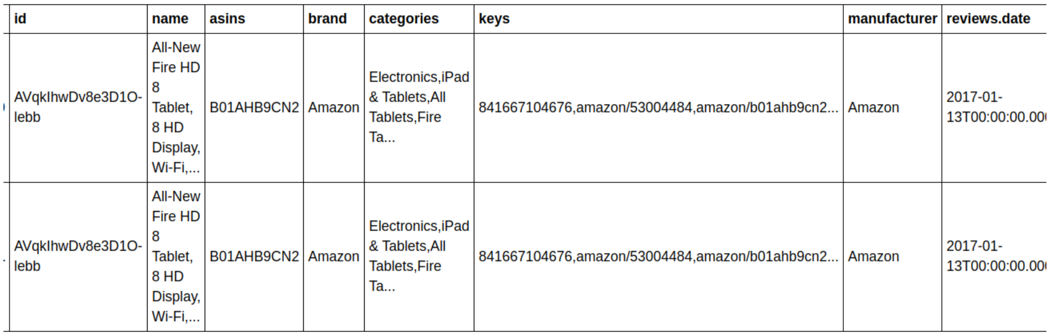
2 Dataset

2.1 Overview

The dataset which we are going to use is based on Amazon branded/Amazon manufactured products only, and Customer satisfaction with Amazon products seem to be the main focus in the given dataset.

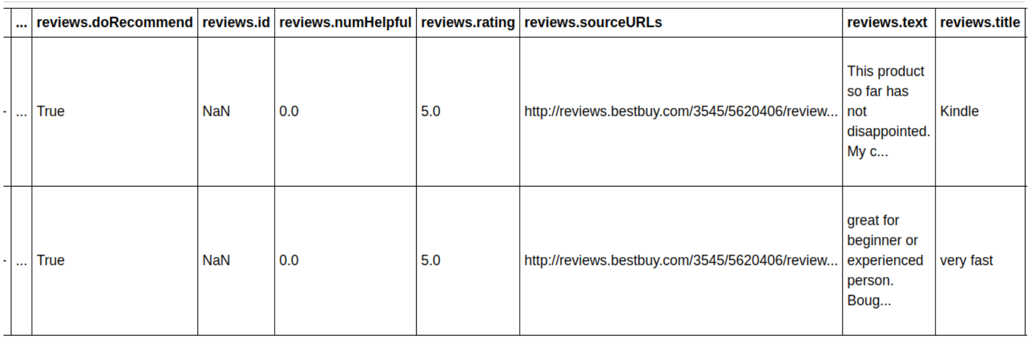
Assumptions taken:

1. We're assuming that sample size of 30K examples are sufficient to represent the entire population of sales/reviews
2. We're assuming that the information we will find in the text reviews of each product will be rich enough to train a sentiment analysis classifier and achieve great accuracy ($hopefully > 70\%$).



id	name	asins	brand	categories	keys	manufacturer	reviews.date
AVqkIhwDv8e3D1O-lebb	All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi,...	B01AHB9CN2	Amazon	Electronics,iPad & Tablets,All Tablets,Fire Ta...	841667104676,amazon/53004484,amazon/b01ahb9cn2...	Amazon	2017-01-13T00:00:00.00
AVqkIhwDv8e3D1O-lebb	All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi,...	B01AHB9CN2	Amazon	Electronics,iPad & Tablets,All Tablets,Fire Ta...	841667104676,amazon/53004484,amazon/b01ahb9cn2...	Amazon	2017-01-13T00:00:00.00

Figure 1: Dataset columns I



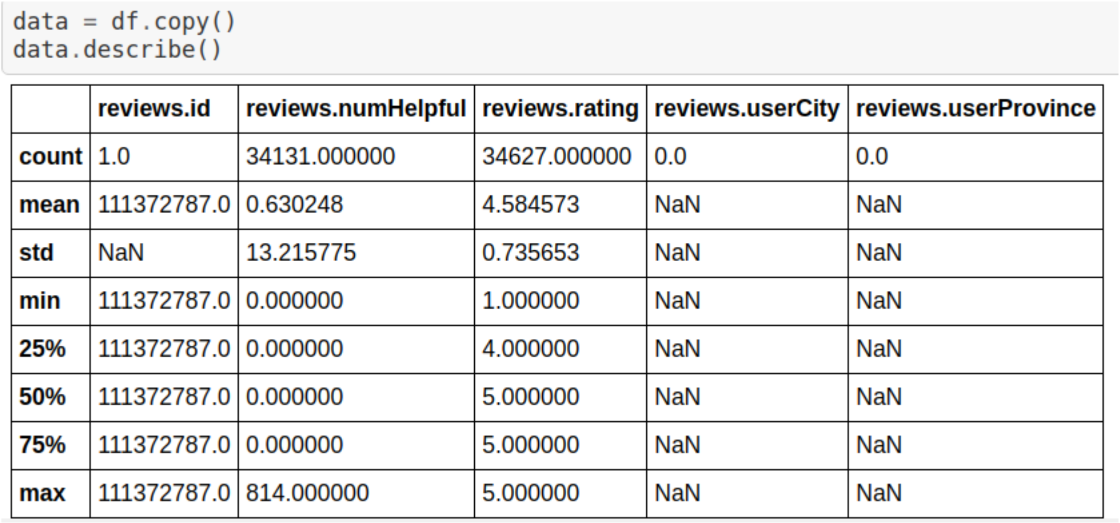
	...	reviews.doRecommend	reviews.id	reviews.numHelpful	reviews.rating	reviews.sourceURLs	reviews.text	reviews.title
-	...	True	NaN	0.0	5.0	http://reviews.bestbuy.com/3545/5620406/review...	This product so far has not disappointed. My c...	Kindle
-	...	True	NaN	0.0	5.0	http://reviews.bestbuy.com/3545/5620406/review...	great for beginner or experienced person. Boug...	very fast

Figure 2: Dataset columns II

2.2 Analysis

Based on the descriptive statistics above, we can deduce the following about the dataset:

1. Average review score of 4.58, with low standard deviation.
2. Most review are positive from 2nd quartile onwards.
3. The average for number of reviews helpful (reviews.numHelpful) is 0.6 but high standard deviation.
4. The data are pretty spread out around the mean, and since can't have negative people finding something helpful, then this is only on the right tail side.
5. The range of most reviews will be between 0-13 people finding helpful (reviews.numHelpful)
6. The most helpful review was helpful to 814 people



```
data = df.copy()
data.describe()
```

	reviews.id	reviews.numHelpful	reviews.rating	reviews.userCity	reviews.userProvince
count	1.0	34131.000000	34627.000000	0.0	0.0
mean	111372787.0	0.630248	4.584573	NaN	NaN
std	NaN	13.215775	0.735653	NaN	NaN
min	111372787.0	0.000000	1.000000	NaN	NaN
25%	111372787.0	0.000000	4.000000	NaN	NaN
50%	111372787.0	0.000000	5.000000	NaN	NaN
75%	111372787.0	0.000000	5.000000	NaN	NaN
max	111372787.0	814.000000	5.000000	NaN	NaN

Figure 3: Dataset analysis

3 Preparing the data

3.1 Split into Train/Test

We split the Data into training set and test sets.Our goal is to eventually train a sentiment analysis classifier. Since the majority of reviews are positive (5 stars), we did a **stratified split** on the reviews score to ensure that we don't train the classifier on imbalanced data.

3.2 Data exploration

Data exploration helps create a more straightforward view of database rather than pouring over thousands of figures in unstructured data it helps in visual exploration to understand what is in a database and the characteristics of the data, rather than through traditional data management systems.

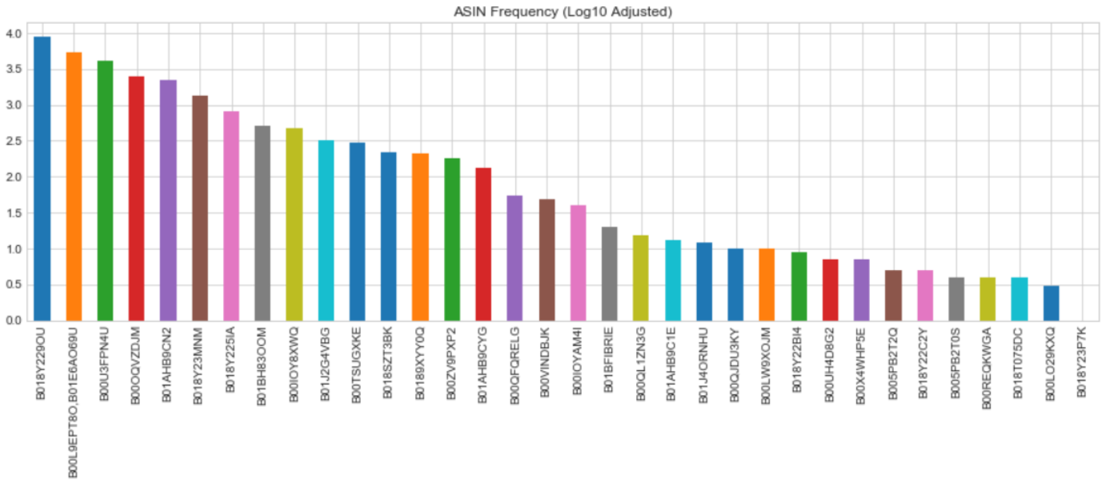


Figure 4: ASINs Frequency

It was clear that for each ASIN there can be more than one names it might be because a single ASIN can have many names due to different vendor listings or There could also a lot of missing names/more unique names with slight variations in title (ie. 4gb vs 4 gb, NAN for product names).

Based on the bar graph for ASINs, we see that certain products have significantly more reviews than other products, which may indicate a higher sale in those specific products. Also,we also see that the ASINs have a ”right tailed” distribution which can also suggest that certain products have higher sales which can correlate to the higher ASINs frequencies in the reviews. So we can say that certain ASINs (products) have better sales, while other ASINs have lower sale, and in turn dictates which products should be kept or dropped.

Also, we saw that the first 19 ASINs show that consumers recommend the product, which is consistent with the ”reviews.rating / ASINs” analysis above, where the first 19 ASINs have good ratings between 4.0 to 5.0 The remaining ASINs have fluctuating results due to lower sample size, which should not be considered.

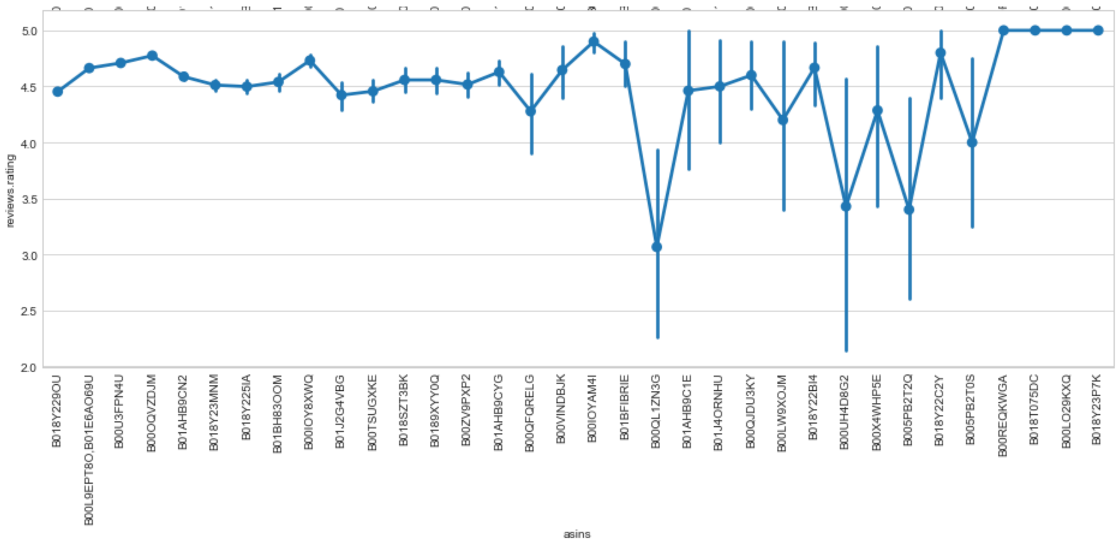
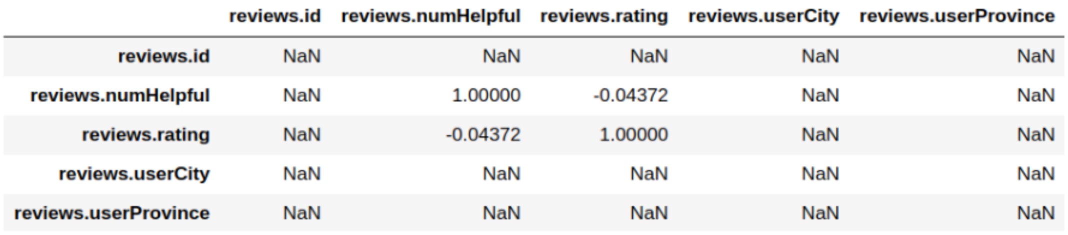


Figure 5: ASIN Reviews



	reviews.id	reviews.numHelpful	reviews.rating	reviews.userCity	reviews.userProvince
reviews.id	NaN	NaN	NaN	NaN	NaN
reviews.numHelpful	NaN	1.00000	-0.04372	NaN	NaN
reviews.rating	NaN	-0.04372	1.00000	NaN	NaN
reviews.userCity	NaN	NaN	NaN	NaN	NaN
reviews.userProvince	NaN	NaN	NaN	NaN	NaN

Figure 6: Correlation

3.3 Correlation

We used Correlation to discover relationships between variables. More precisely, the correlation is a measure of the linear relationship between two variables. Here in our analysis in data exploration we focused on correlation between ASINs and reviews.rating.

By above table it is clear that there are many ASINs with low occurrence that have high variances, as a result we concluded that theses low occurrence ASINs are not significant in our analysis given the low sample size .Similarly in our correlation analysis between ASINs and reviews.rating, we see that there is almost no correlation which is consistent with our findings.

4 Sentiment Analysis

4.1 Approach

The basic approach was to first segregate ratings from 1-5 into positive, neutral, and negative. And then, turning content i.e. reviews into numerical feature vectors using the Bag of Words strategy. This was implemented with the help of SciKit-Learn's CountVectorizer.

4.2 Bag of Words strategy

A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms. This strategy consists of three main parts i.e. Text preprocessing, Occurrence counting and creating a feature vector.

Text pre-processing further consists of two parts. The first one is tokenization i.e. breaking sentences into words and the other is filtering the stop-words like ”the”, ”and”, etc. The count occurrences algorithm is used to iterate over an array to count the number of times that a particular word occurs. And then finally, we create the feature vector which we will use to train our models.

4.3 Classifiers Used

These are the classifiers that we used initially to see their performance on this data.

Table 1: Classifiers vs Accuracy

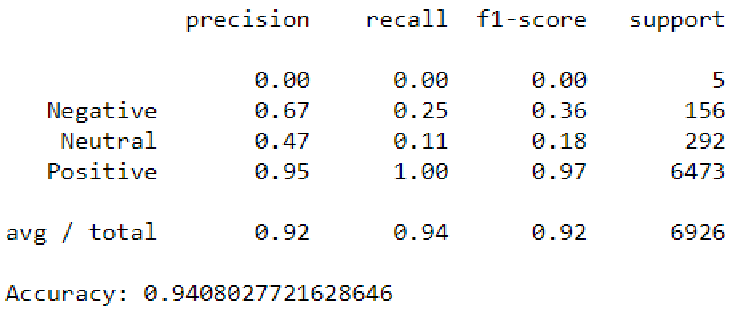
Classifier	Accuracy
Multinominal Naive Bayes	93.45%
Logistic Regression	93.70%
Support Vector Machine	93.94%
Decision Tree	90.18%
Random Forest	93.46%

4.4 Fine tuning in SVM classifier

As we observed above that Support Vector Machine classifier performed better than rest of the classifiers. So, we decided to proceed with SVM and decided to fine tune the parameters of this classifier to maximize it's accuracy. First, we ran a grid search of the best parameters on a grid of possible values. Then, we fit the grid search to our training data set. Then, we tried to test this final classifier (after fine tuning) on some arbitrary reviews. Finally, we tested the accuracy of this improved classifier on our final dataset and observed that the accuracy has increased to **94.08%**.

5 Analysis

5.1 Performance analysis



```
             precision    recall  f1-score   support

                  0.00      0.00      0.00         5
   Negative       0.67      0.25      0.36       156
    Neutral       0.47      0.11      0.18       292
   Positive       0.95      1.00      0.97      6473

avg / total       0.92      0.94      0.92      6926

Accuracy: 0.9408027721628646
```

Figure 7: Performance analysis

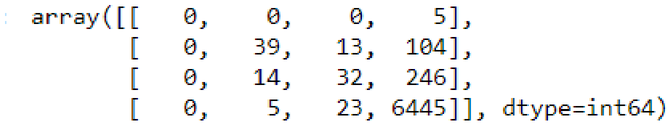
The results in this analysis confirms our previous data exploration analysis, where the data are very skewed to the positive reviews as shown by the lower support counts in the classification report. Also, both neutral and negative reviews has large standard deviation with small frequencies, which we would not consider significant as shown by the lower precision, recall and F1 scores in the classification report.

However, despite that Neutral and Negative results are not very strong predictors in this data set, it still shows a 94.08 percent accuracy level in predicting the sentiment analysis, which we tested and worked very well when inputting arbitrary text (new_text). Therefore, we are comfortable here with the skewed data set. Also, as we continue to input new dataset in the future that is more balanced, this model will then re-adjust to a more balanced classifier which will increase the accuracy level.

5.2 Confusion matrix

We see that positive sentiment can sometimes be confused for one another with neutral and negative ratings, with scores of 246 and 104 respectively. However, based on the overall number of significant positive sentiment at a score 6445, then confusion score of 246 and 104 for neutral and negative ratings respectively are considered insignificant.

Therefore, we conclude that the products in this dataset are generally positively rated, and should be kept from Amazon's product roster.



```
: array([[   0,    0,    0,    5],
         [   0,   39,   13,  104],
         [   0,   14,   32,  246],
         [   0,    5,   23, 6445]], dtype=int64)
```

Figure 8: Confusion matrix

Conclusions and Discussion

From the analysis above in the classification report, we can see that products with lower reviews are not significant enough to predict these lower rated products are inferior. On the other hand, products that are highly rated are considered superior products, which also performs well and should continue to sell at a high level.

The good news is that despite the skewed dataset, we were still able to build a robust Sentiment Analysis machine learning system to determine if the reviews are positive or negative. This is possible as the machine learning system was able to learn from all the positive, neutral and negative reviews, and fine tune the algorithm in order to avoid bias sentiments. Also, we were able to successfully associate positive, neutral and negative sentiments for each product in Amazon's Catalog.