SENTIMENT ANALYSIS OF AMAZON REVIEWS USING MACHINE-LEARNING ALGORITHMS

Sheikh MD. Nafis Noor Nabil
Department of Computer Science and Engineering
BRAC University
sheikh.md.nafis@g.bracu.ac.bd

Farhana Eyesmeen Maria
Department of Computer Science and Engineering
BRAC University
farhana.eyesmeen.maria@g.bracu.ac.bd

Naimul Haque Chowdhury
Department of Computer Science and Engineering
BRAC University
naimul.haque.chowdhury@g.bracu.ac.bd

Ariful Hassan
Department of Computer Science and Engineering
BRAC University
ariful.hassan@g.bracu.ac.bd

Sabir Ahmed
Department of Computer Science and Engineering
BRAC University
sabir.ahmed@g.bracu.ac.bd

ABSTRACT

Artificial intelligence and machine learning have given businesses an invaluable advantage with Amazon review analysis. It can find out the accuracy rate by analyzing the sentiment of provided reviews. In this study, various machine learning algorithms, including Naive Bayes classification, Logistic regression, Linear SVC, Decision tree classifier, Random forest classifier were used to check the accuracy rate of the reviews of the products of amazon. The results showed that the Logistic Regression algorithm had the highest accuracy of around 94

I INTRODUCTION

Nowadays, every product has a very competitive environment in the market. And a sentiment analysis system comes handy that integrates reviews from social media sites (Facebook, Twitter, YouTube, Blog and Amazon) to extract the meanings of product reviews and provide a positive or negative rating based on analyzed data. Filters can be added for the rating for better use. Based on the given ratings, top management can make decisions on product/ company reputation and create more productive marketing strategies. General people also can use this system to get to know about available products or brands and how positive or negative they are among people.

II RESEARCH METHODOLOGY

To explain the entire process, here is a workflow diagram given below

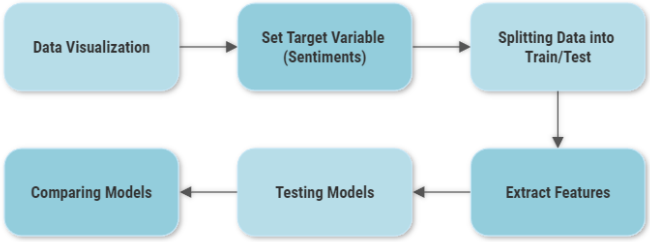

Figure 1: Flow Chart of our work flow

A Data Collection

In our project, we used a dataset collected from Kaggle. The dataset consists of 34600 rows and 21 columns including id, name, asins, brand, categories, keys, manufacturer, reviews.date, reviews.dateAdded, reviews.dateSeen, reviews.purchase, reviews.doRecommend, reviews.id, reviews.numHelpful, reviews.Rating, reviews.URLs, reviews.text, reviews.title, reviews.usercity, reviews.userprovince and reviews.username. The features include discrete, and categorical data types.

B Data Visualization

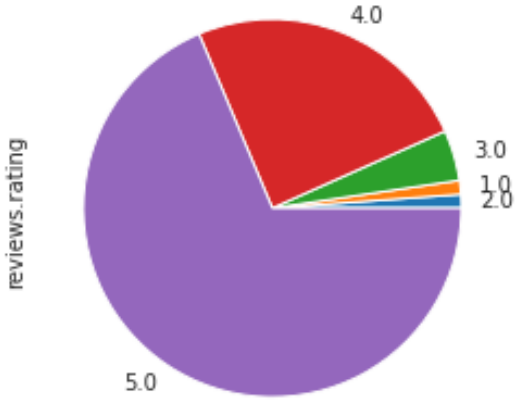

Figure 2: Review Rating Stats

This pie chart represents the statistics of provided review ratings. The rating section has been divided into five portions. Here in this chart, most of the products have five ratings which are also the highest.

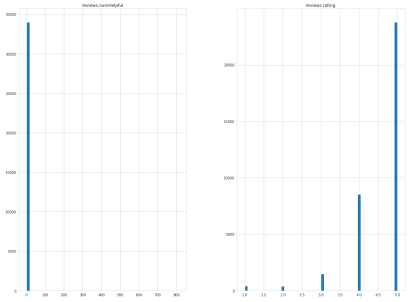

Figure 3: Distributions of numerical variables

From these graphs above , the visualization of the distributions of numerical variables has been shown. For instance, the left graph portrays the weight of the reviews that came from a higher number of people. And then the second graph represents the products being rated and there are twice the amount of 5 star ratings than the others ratings combined.

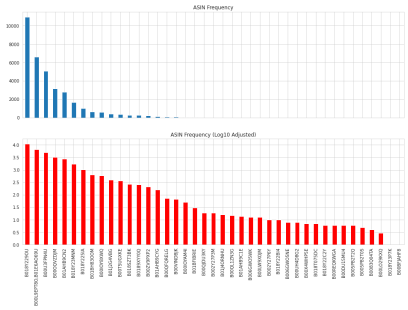

Figure 4: Review numbers

This graph shows the number of reviews of each product.

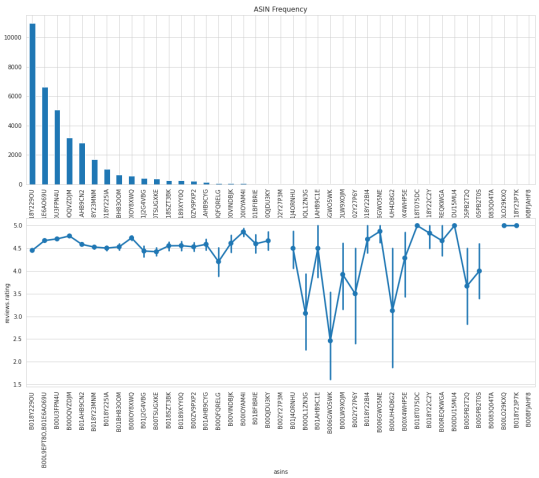

Figure 5: Review ratings

These graphs portray the Review rating of the products. Here,The most frequently reviewed products have their average review ratings in the 4.5 - 4.8 range, with little variance.

Figure 6: Correlation between review rating and helpful review

Here is a graph of the correlation to demonstrate the correlation analysis between ASINs and reviews. rating. And there is almost no correlation that is consistent with the findings.

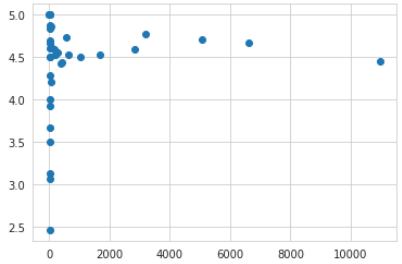

Figure 7: Scatter diagram of ASINs and reviews.rating

This chart represents our analysis in data exploration above between ASINs and reviews.rating, we discovered that there are many ASINs with low occurrence that have high variances, as a result we concluded that these low occurrence ASINs are not significant in our analysis given the low sample size.

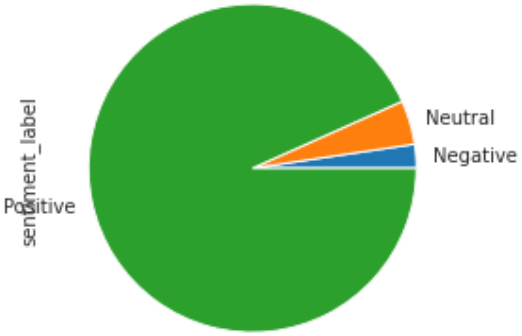

Figure 8: Pie chart of Sentiment labels

Here in this pie chart, after labeling the sentiments we can see most of the reviews are positive.

C Data Pre-Processing

C.1 ***Removing unnecessary columns.*** At first we analyzed the dataset and dropped unnecessary columns like 'keys', 'manufacturer', 'brand','reviews.id','reviews.date','reviews.dateAdded', 'reviews.dateSeen', 'reviews.userProvince', 'reviews.didPurchase', 'reviews.userCity', and 'reviews.sourceURL.', r6

C.2 ***Handling Null Values.*** First, we separate necessary reviews.rating, reviews.text and reviews.title columns in a variable called 'senti-data'. Applying '.isnull().sum()' on the imported dataset returns the amount of null values. There were missing values in all the columns. So, we dropped the rows having null values by using .dropna., r6

C.3 ***Sentiment Labeling*** . We label sentiments based on the product review ratings. Rating 5 and 4 are set as Positive, 3 as Neutral and below that as Negative., r6

C.4 ***Split into Train/Test*** . We split our data where the training dataset had 80% and the testing dataset 20% of the total data. For this task, we used sklearn's Stratified ShuffleSplit class., r6

C.5 ***Preparing Data*** . We used NLTK in this process. For the bag of words strategy, we used SciKit-Learn's CountVectorizer for Text preprocessing. TfidfTransformer was used to reduce redundancy and downscale the weights of each word., r6

III MODEL BUILDING

We have compared 5 different classifiers to identify which of the classifier is best for the model. The classifier we used is Naive Bayes Classification, Logistic Regression, Linear Support Vector Classification, Decision Tree Classifier, and Random Forest Classification.

Naive Bayes Classification. Naive Bayes classification is a machine learning algorithm that is based on the idea of making predictions using Bayes' theorem. The algorithm in Naive Bayes classification predicts the likelihood of a specific class based on the presence or absence of certain features.Naive Bayes makes the assumption that all features are independent of one another, which can help to reduce the complexity of the model and improve its performance.

Logistic Regression. When predicting the likelihood of a class based on some dependent variables, the machine learning classification approach known as logistic regression is applied. The logistic regression model, in essence, adds up the input features. Using the Sigmoid function, which transforms numerical data into an expression of probability between 0 and 1.0, it assigns probabilities to discrete outcomes. Two categories are created. After data points have been classified into a class using the Sigmoid function, a hyperplane is utilized as a decision line to divide two groups. The decision boundary can then be used to forecast the kind of upcoming data points. The logistic function is of the form:

$$p(x) = 1/L + e^{-k(x-x0)}$$

Linear Support Vector Classification. Support vector classification (SVC) is a type of supervised machine learning algorithm that can be used for classification tasks, including analyzing sentiment of amazon reviews. It works by finding the hyperplane in a high-dimensional feature space that maximally separates the different classes.It can handle high-dimensional data and is able to find the most important features so that it can use them to analyze which can help to improve the performance.

Decision Tree Classifier. Classification and regression issues can be resolved using a decision tree, a supervised learning approach. It is a tree-structured classifier in which each leaf node represents the classification outcome, each internal node represents a feature and each branch is for decision-making.

Random Forest Classification . Random forest classification is a machine learning method that can be used to predict the class of a sample based on the class of a group of decision trees, where each tree is trained on a randomly selected subset of the data.

IV RESULTS AND ANALYSIS

Model	Accuracy
Logistic Regression	94%
Linear SVC	93.53%
Random forest	93.34%
Naive Bayes	93%
Decision tree	89%

Here the accuracy rate of the models Naive Bayes, Logistic regression, Linear SVC, Decision tree classifier, Random forest classifier consecutively are 93%, 94%, 93%, 90% and 93%.As per demonstrations of the four models, it is evident that the Logistic regression produces the highest accuracy score (94%) whereas the SVM model produces the lowest accuracy score (90%).

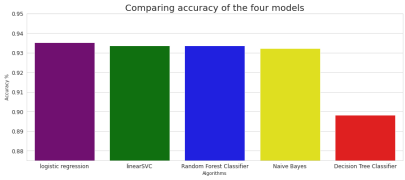

Figure 9: Accuracies of all Machine Learning Algorithms Tested

The information in the table above shows that Logistic Regression predicts sentiment for amazon product reviews with the highest accuracy. Thus, we can conclude that using Logistic Regression, for our model, is the best option to get highly accurate results.

V CONCLUSION

From the analysis of the classification reports shown above, 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 perform well and should continue to sell at a high level. Despite the skewed data set, we were still able to build a robust Sentiment Analysis machine learning system to determine if the reviews are positive or negative.To conclude, although we need more data to balance out the lower rated products to consider their significance, we were still able to successfully associate positive,

neutral and negative sentiments for each product in Amazon's Catalog.