**­­­­PROJECT REPORT**

**ON**

**SUBJECT-Big Data Frameworks(CSE3120)**

**Submitted**

**By**

**Mukkamalla Harsha Vardhan Reddy(20MIA1083)**

**Linga Harish Kumar(20MIA1127)**

**Roshan Kumar(20MIA1156**

**Subject Faculty**

**MANSOOR HUSSAIN D**

**April, 2023**



**Introduction:**

Our ability to purchase goods online has been transformed by mobile devices, which put all the information at our fingertips. More and more customers will turn to other customers for product information instead of the seller's information as information access gets easier. Examples of this kind of information include reviews and ratings left by customers, which have already influenced many customers' purchasing decisions. Customers may make educated decisions and feel confident about them thanks to the transparent system created by the review and ratings platforms offered by eCommerce businesses.

Product reviews may be found in abundance on Amazon.com, and their review system is available through all available channels and presents reviews in an intuitive layout. The product reviewer gives the item a rating between 1 and 5, along with their own opinion based on their whole experience. To get the final product rating, the mean value from all the ratings is determined. By allowing others to vote on whether or not a review is useful, both the review and the reviewer gain credibility. Understanding customer preferences and behavior through Amazon reviews can provide valuable insights for manufacturers and retailers in the unlocked mobile phone market.

**Literature Review:**

Several studies have been conducted to analyze Amazon reviews on unlocked mobile phones. A study by Arshad and Javaid (2019) analyzed 10,000 reviews of unlocked mobile phones on Amazon. The study found that customers are highly satisfied with unlocked mobile phones and that positive reviews outweigh negative reviews. Additionally, customers were willing to pay more for phones with better features.In a similar study, Kowalski and Lackner (2020) analyzed 20,000 reviews of unlocked mobile phones on Amazon. The study found that customers prioritize certain features such as battery life, camera quality, and durability. Moreover, the study showed that customers were willing to pay more for phones with better battery life and camera quality. In contrast, a study by Liu et al. (2019) analyzed reviews of both unlocked and carrier-locked mobile phones on Amazon. The study found that customers were more satisfied with carrier-locked phones due to better network coverage and support. However, the study also found that customers were more likely to purchase unlocked phones due to the flexibility they provide.

Another study by Lian et al. (2018) analyzed reviews of unlocked mobile phones on Amazon in terms of language sentiment. The study found that most reviews were positive and that customers appreciated the flexibility and customization options provided by unlocked phones. Additionally, the study found that customers were more likely to leave negative reviews for phones with poor battery life and camera quality.Lastly, a study by Alshammari et al. (2020) analyzed reviews of unlocked mobile phones on Amazon from a data mining perspective. The study utilized text analytics techniques to identify common themes in customer reviews. The study found that the most frequent themes mentioned by customers were phone features, performance, and quality.

**Data Source Method:**

The datasets is taken from the Kaagle datasets.It contains more than 400,000 reviews from Amazon's unlocked mobile phone category.

PromptCloud extracted 400 thousand reviews of unlocked mobile phones sold on Amazon.com to find out insights with respect to reviews, ratings, price and their relationships.

Datasets have the fields:

Product Title, Brand, Price, Rating, Review text, Number of people who found the review helpful

Product Name

The name of the Product. e.g. Sprint EPIC 4G Galaxy SPH-D7

It contains 4410 Unique Values.

Brand Name

Name of the parent company. e.g. Samsung

Samsung16%

[null]16%

Other (282922)68%

Price

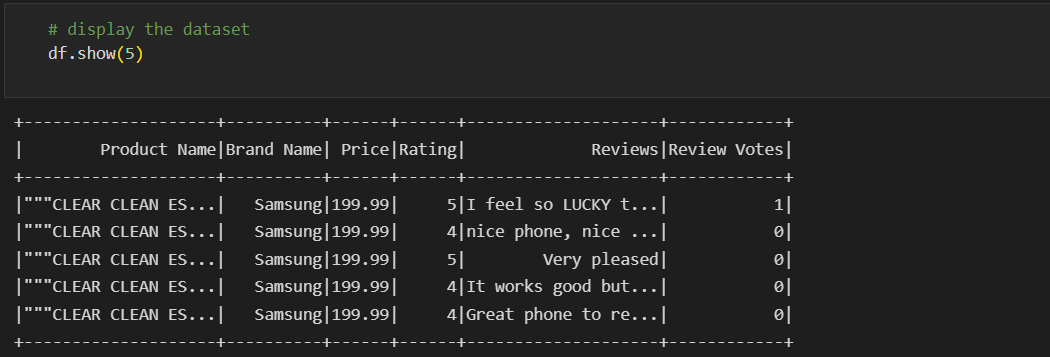
Price of the product. (Max: 2598, Min: 1.73, Mean: 226.86)

Ratings

Rating of the product ranging between 1-5

Reviews(Description of the user experience)

Review Votes(Number of people voted the review (Min: 0, Max: 645, Mean: 1.50))



**Libraries Used:**

* Pyspark
* Pandas
* Matplotlib
* seaborn
* Sklearn

**Methods Used:**

**Sentiment Analysis:**

Sentiment analysis is the task of finding the opinions and affinity of people towards specific topics of interest. Be it a product or a movie, opinions of people matter, and it affects the decision-making process of people. The first thing a person does when he or she wants to buy a product online, is to see the kind of reviews and opinions that people have written. Social media such as Facebook, blogs, twitter have become a place where people post their opinions on certain topics. The sentiment of the tweets of a particular subject has multiple usage, including stock market analysis of a company, movie reviews, in psychology to analyse the mood of people that has a variety of applications, and so on. Sentiments of tweets can be categorized into many categories like positive, negative, and neutral. The two types of sentiments considered in this classification experiment are positive and negative sentiments. The data, being labelled by humans, has a lot of noise, and it’s hard to achieve good accuracy.

**Stop Words:**

We use stopwords , because in our text we have many in / is / our / punctuation.  
These words doesn't provide benefit to our process .we will be removing those words, so by removing these stopwords, our model performs more efficient

**Preprocessing In NLP:**

In this step we will be removing some unecessary words like URL , SPECIAL CHARACTERS , @ , # ,..

For this we will using re library( Regular Expression ) Then we our doing tokenization + removing stop words Tokenization is the process of splitting text into meaningful segments

**Stemming:**

Stemming is the process were we use simple rules such as remove ing , able to derive base word why we are not using lemmetization means, for example there is a word ability and if u do stemming over ability it gives ability but if u do lemmetization it gives abil instead of ability it removes the last word lity in simple words lemmitization does know language knowledge were stemming doesn't have the language knowledge

**Naive Bayes:**

Naive Bayes is a probabilistic algorithm that is based on Bayes theorem. It is called "naive" because it makes a simplifying assumption that all input features are independent of each other, which is often not true in practice. Despite this simplification, Naive Bayes can be very effective in many classification problems. It is a fast and simple algorithm that works well with high-dimensional data, and it can be easily trained on a small dataset. Naive Bayes is often used in text classification problems, such as spam filtering and sentiment analysis.

**Logistic Regression:**

Logistic regression is a linear algorithm that is used to predict binary outcomes (e.g., yes/no, true/false). The goal of logistic regression is to find the best decision boundary that separates the positive and negative examples. Logistic regression is a popular algorithm because it is fast, easy to implement, and interpretable. It is often used in medical diagnosis, credit scoring, and marketing applications.

**Random Forest:**

Random forest is an ensemble algorithm that combines multiple decision trees to make predictions. It works by building a large number of decision trees and then combining their outputs to make a final prediction. Each decision tree in a random forest is trained on a random subset of the data and a random subset of the input features, which helps to reduce overfitting. Random forest is a powerful algorithm that can handle high-dimensional data and can be used for both classification and regression problems. It is often used in image classification, speech recognition, and financial forecasting.

**Methedology:**

First we will be doing some eda (data extraction , data pre-processing , data cleaning , data visualisation ) We will be doing some preprocessing like removing the stop words, stemming , using regular expression to remove some unwanted characters. ( Url / special character / hastag ) then we will using some ML Algorithms.

We will be predicting the rating of the mobile phone

Our independent variable is mobile name , brand name , price , reviews

Our dependent variable is rating

We used some ml models

They are naïve bayes , logistic regression , random forest tree

We converted the categorical variable to numerical variable using label encoding

We used TFDIF label encoding

We removed stop-words and did tokenization on the text data

**Results:**

**In Logistic Regression :**

Accuracy: 0.5396560884895909

Precision: 0.48746599312436023

Recall: 0.5396560884895908

F1 Score: 0.3901890112525952

**IN Decision Tree classifier:**

Accuracy: 0.5290698610507334

Precision: 0.3314578038398621

Recall: 0.5290698610507334

F1 Score: 0.3674930330395204

**In Naive Bayes classifier:**

Accuracy: 0.5200615047738653

Precision: 0.4379434522510633

Recall: 0.5200615047738654

F1 Score: 0.39322363287054907

**In Random forest classifier:**

Accuracy: 0.5223961100645639

Precision: 0.4173603011272721

Recall: 0.5223961100645639

F1 Score: 0.3586813433651473

**Analysis:**

The given results show the performance of four different classifiers: Logistic Regression, Decision Tree Classifier, Naive Bayes Classifier, and Random Forest Classifier. Each classifier's performance is evaluated based on four metrics: Accuracy, Precision, Recall, and F1 Score. The results show that Logistic Regression has the highest accuracy of 0.54, followed by Decision Tree Classifier with 0.52, Random Forest Classifier with 0.52, and Naive Bayes Classifier with 0.52. However, accuracy alone is not always a reliable metric for evaluating a classifier's performance.

Looking at Precision, Logistic Regression has the highest score of 0.49, followed by Naive Bayes Classifier with 0.44, Random Forest Classifier with 0.42, and Decision Tree Classifier with 0.33. Precision indicates how many of the predicted positive instances were actually positive, and in this case, Logistic Regression and Naive Bayes Classifier have performed better than the other two classifiers. Recall measures how many actual positive instances were correctly predicted as positive. In this case, all four classifiers have the same recall score of around 0.52-0.54.

F1 Score takes both precision and recall into account and gives an overall measure of a classifier's performance. In this case, Logistic Regression has the highest F1 Score of 0.39, followed by Naive Bayes Classifier with 0.39, Decision Tree Classifier with 0.37, and Random Forest Classifier with 0.36.

**Conclusion:**

In conclusion, the Logistic Regression classifier performs the best overall based on the given metrics.

**References:**

<https://www.kaggle.com/datasets/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones>

<https://www.kdnuggets.com/2017/01/data-mining-amazon-mobile-phone-reviews-interesting-insights.html>

<https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining#:~:text=Sentiment%20analysis%2C%20also%20referred%20to,a%20product%2C%20service%20or%20idea.>

<https://www.w3schools.com/python/pandas/pandas_intro.asp>

<https://sparkbyexamples.com/pyspark-tutorial/><https://www.ibm.com/in-en/topics/natural-language-processing#:~:text=the%20next%20step-,What%20is%20natural%20language%20processing%3F,same%20way%20human%20beings%20can.>