iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34623

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Consumer Product Recommendation by Sentiment Analysis of Online Reviews

Dr. Manish Goswami[1], Ms. Trishna Chakraborty[2], Sagar Gabhane[3], Neha Dahare[4], Aman Chatur[5]
[1, 2, 3, 4, 5]*Department of Information Technology, Rajiv Gandhi College of Engineering & Research, Nagpur- 441110, India,*
(Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University)

***Abstract:** Sentiment analysis, also refers as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given data. Using machine learning techniques and natural language processing we can extract the subjective information of a data and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling an object, buying an object and so on. Sentiment analysis is actually far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary, grammar,) but it is also why it is very interesting to work on.*
In this research work, five different algorithms were chosen for sentiment analysis that is Decision Tree Classifier, Random Forest, Naive Bayes, and SVM & Logistic Regression. We compare all five algorithms, out of which we found that SVM gives the best accuracy that is 94.08%.

I. INTRODUCTION

Sentiment is thought, judgment prompted by feeling. From a customer's point of view, customer are able to post their own content through various social media, such as online social networking sites, etc. From a researcher's point of view, many social media sites release their APIs and analysis by researchers or developers. However, those types of online data have several faults. The first fault is that users can freely post their own content, the quality of their opinions cannot be guaranteed. The second fault is that the data checking of such online data is not always available.

A data checking is more like a tag of a certain opinion, indicating their positive, negative or neutral opinion. When purchasing the most recent items on Amazon, perusing reviews is an essential piece of the acquiring procedure. Client surveys/evaluations from clients who have really obtained and utilized the item being referred to can give you more setting to the item itself. Every commentator rates the item from 1 to 5 star-rating, and then gives a content outline of the encounters and feelings about the item. And this helps to generalize whether the product is doing good or bad. Ratings done on Electronics items from Amazon often rate the product 4 or 5 star, and such reviews are found to be quite often viewed as supportive. 1 and 2 stars given are utilized to imply objection, and 3 stars by any means have no noteworthy effect. With a 5-star framework, one can enable the forthcoming user to make more educated examination between two items a: a user might be bound to buy an item that evaluates 4.2 star than an item that is appraised 3.8 star, this helps the customer to buy the product.

II. LITERATURE SURVEY

There is substantial amount of literature work available regarding sentiment analysis of online reviews on consumer product recommendation.[1] focuses on a typical sentiment analysis model consisting of three core steps, namely data preparation, sentiment classification and review analysis, and describe representative techniques involved in these steps.[2] gives stress on a number of useful insights have been derived from the visualizations and analysis which may help in improving the existing review system of Amazon to make it better for the customers as well as the sellers.[3]shows how to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis.[4] try to have correlation between the amazon product reviews & the rating of the products given by customers.[5] gives on trend to be used supervised gaining information of technique on an oversized scale amazon dataset to polarize it and obtain fine accuracy.[6] presents an empirical study of efficacy of classifying product review by tagging the keyword.[7] uses opinion Mining and Sentiment Analysis for Amazon Product Review using Lexicon , Rule-Based Approach and Testing on Machine Learning Algorithm.

III. METHODOLOGY

A. *Approaches Of Sentiment Analysis*

There are many algorithms for performing sentiment analysis out of which we choose five different algorithms to perform sentiment analysis in this project.

Simply said that machine learning permit computers to learn new things without being expressly programmed to perform them. Sentiment analysis models can be trained to read beyond mere definitions, to understand things like, context, irony, etc.

For example*: "Super user-friendly interface. Yeah right. An engineering degree would be helpful."*

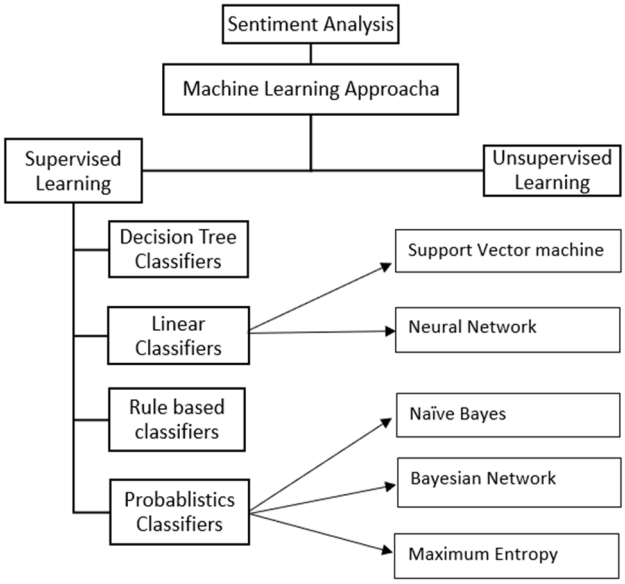


Fig. Sentiment Analysis Using Machine Learning

The words 'helpful' and 'super user-friendly' could be read as positive but this are negative comment. Using sentiment analysis, computers could automatically process text data and also understand it just as a human would, saving hundred of employee hours.

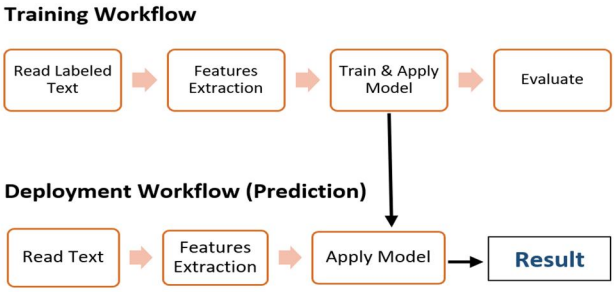


Fig. How Sentiment Analysis with Machine Learning work

B. *Algorithms Used*

1) *Naive Bayes:* Naive Bayes is a fairly simple group of probabilistic algorithms that, for sentiment analysis classification, assigns a probability that a given word or phrase should be considered as a positive or negative. A Naive Bayes classifier is a probabilistic ml model that's is used for classification task.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

2) *Logistic Regression:* Logistic regression is the type of Supervised Learning method. It is also used for predicting the categorical dependent variable(target) using a given set of independent variable. Logistic regression predicts the output of a categorical dependent variable(target).

$$0 \leq h_\theta(x) \leq 1$$

a) Type of Logistic Regression: Logistic Regression can be classify into three types:

- Binomial Logistic regression.
- Multinomial Logistic regression.
- Ordinal Logistic regression.

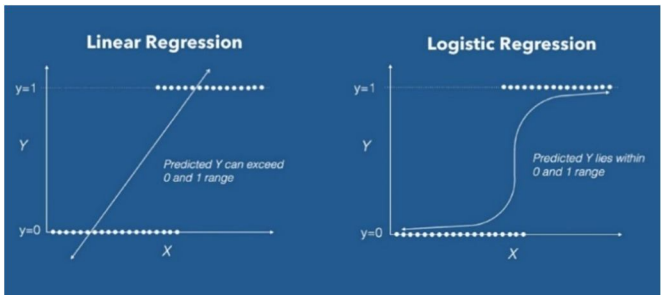
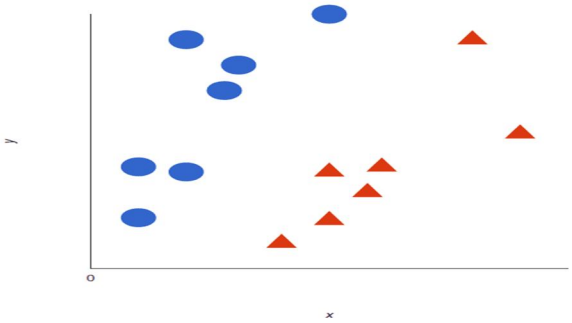


Fig: Linear Regression VS Logistic Regression Graph

3) *Support Vector Machine:* A support vector machine is the type of a supervised machine learning model; it is similar to linear regression but more advanced. SVM algorithm uses to train and classify text within our sentiment polarity model, taking it a step beyond X/Y prediction. For a visual explanation, we will use two tags that are red and blue with two data features that are X and Y. We will train our classifier to give output an X/Y coordinate as either blue or red.



The SVM then assigns a hyper plane that best separates the tags. In 2D, this is simply a line. Anything on one side of the line is blue and anything on the other side is red. In sentiment analysis this would be positive and negative.

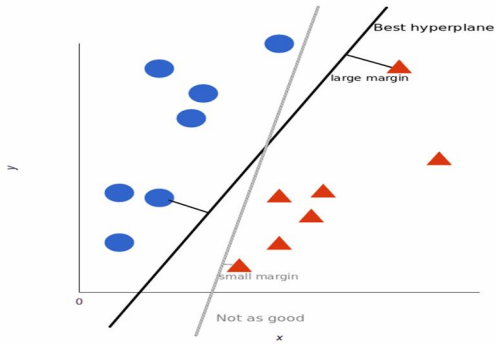


Fig: - SVM

4) *Decision Tree Classifiers:* Decision Tree algorithm belongs to the Supervised Learning technique. In this tree structure is used for classification. The algorithm is used for Classification as well as Regression. But in general, used for classification. In the Tree like structure thus formed, in which, the top most node is called the Root node, internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the target value. In decision nodes, decisions are focuses attention on the given dataset. Decision Tree is easy to understand because, usually mimic human thinking ability while making a decision.

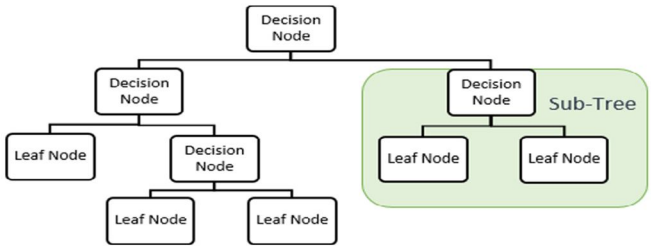
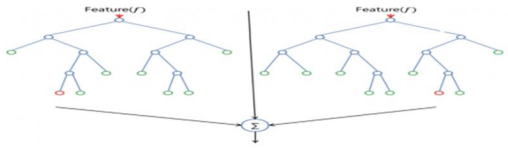


Fig. Decision Tree Classifiers

5) *Random Forest:* Random Forest algorithm comes under Supervised Learning technique, used for Classification and regression problem. It is based on ensemble Learning method, which means combining multiple classifiers to solve a large and complex problem and get the results with much higher accuracy. As the name of the algorithm suggests a random number of Decision Trees are created which when taken in a group resembles as a forest. An average of the decision tree algorithm performed on a given subsets of the dataset is taken, which is further used for prediction of the target values, taking the average of number of such decision trees gives more accuracy in prediction. The forest formed by this is an ensemble of Decision Tree.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$



IV. EXPERIMENTATION

The method followed here to provide the desired results is that, the machine is given the first chance to provide what it understands from the provided data, where the machine uses the libraries in python and finalizes these five columns shown in the figure below. And removes the other columns from the data as it is either useless or unrequired for the calculations to be performed.
After obtaining the desired dataset to work on, then we analyze that the data at the different stages i.e., 25%, 50%, 75%, etc. and also look for its max and min limits in the overall data, using python libraries.

	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

Fig 1: - Analysis of Raw Data

Then as regular work-flow the data is split into training and testing dataset. Now to start working with the training dataset, Asin's columns is considered the most important column of the entire dataset, being unique for every record. It helps to identify how many distinct products we have and what is their buying rate and respective ratings.

```
array(['B01AHB9CN2', 'B00VINDBJK', 'B005PB2T0S', 'B002Y27P3M',
       'B01AHB9CYG', 'B01AHB9C1E', 'B01J2G4VBG', 'B00ZV9PXP2',
       'B0083Q04TA', 'B018Y229OU', 'B00REQKWGA', 'B00IOYAM4I',
       'B018T075DC', nan, 'B00DU15MU4', 'B018Y225IA', 'B005PB2T2Q',
       'B018Y23MNM', 'B00OQVZDJM', 'B00IOY8XWQ', 'B00LO29KXQ',
       'B00QJDU3KY', 'B018Y22C2Y', 'B01BFIBRIE', 'B01J4ORNHU',
       'B018SZT3BK', 'B00UH4D8G2', 'B018Y22BI4', 'B00TSUGXKE',
       'B00L9EPT8O,B01E6AO69U', 'B018Y23P7K', 'B00X4WHP5E', 'B00QFQRELG',
       'B00LW9XOJM', 'B00QL1ZN3G', 'B0189XYY0Q', 'B01BH83OOM',
       'B00BFJAHF8', 'B00U3FPN4U', 'B002Y27P6Y', 'B006GWO5NE',
       'B006GWO5WK'], dtype=object)
```

Fig 2: - ASINS id's

From this we conclude that our data set has total 42 unique products and review their ratings in sorted order. Now we apply the SVM model for the project for actual prediction and analysis results.

V. WHY SVM IS BETTER THAN OTHER MODELS?

SVM gives excellent performance with the noise less data and is relatively memory efficient.
SVM provides the clear margin with of separation between classes.
Also, it is more effective in high dimensional spaces. Thus, SVM is chosen as the model for the classification results of this project.

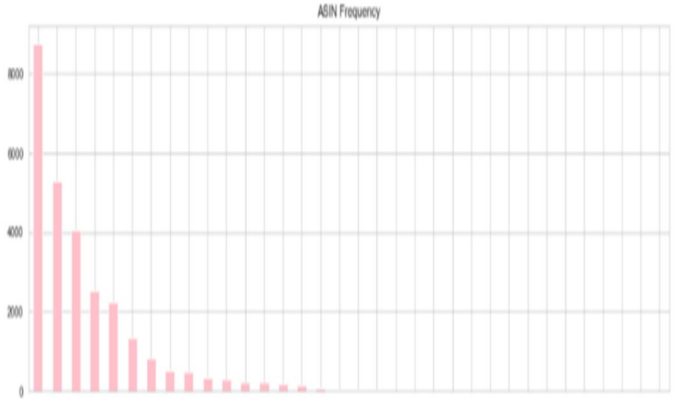


Fig 3: - ASINS Frequency

Depending on the single parameter (ratings) for the accurate results is delicate to work with, thus not only the ratings but also the frequency is taken as another parameter to analyze which products have greater sale and thus can be passed-down to provide most accurate results.

Fig 4: - ASINS Frequency (Log10)

In the 32000-row data we have 42 unique products and all the products buys multiple time and customer provides their rating also multiple time so this graph basically shows highest mean rating for all 42 unique products, where the overall ratings lie between 3-5, supportive enough to accept the data is of high quality.

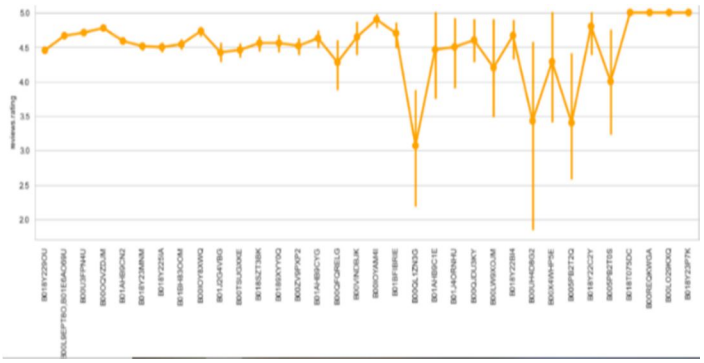


Fig 5: - Rating Analysis

From this analysis, we can see that the 19 ASINs show that consumers recommend the product, which is consistent with the "reviews.rating / ASINs" analysis above, where the 19 ASINs have good ratings between 4.0 to 5.0.

The remaining ASINs have fluctuated results due to lower sample size, which should not be considered.

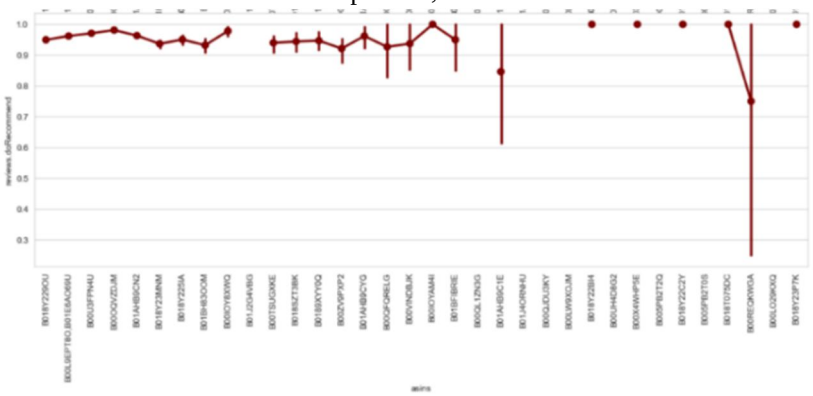


Fig 6: - DoRecommend

All above algorithms were tested on Intel® Core™ i5- CPU @ 1.60 Hz 2.11 GHz /i5+ processor, 64-bit Operating System.

VI. CONCLUSION & FUTURE SCOPE

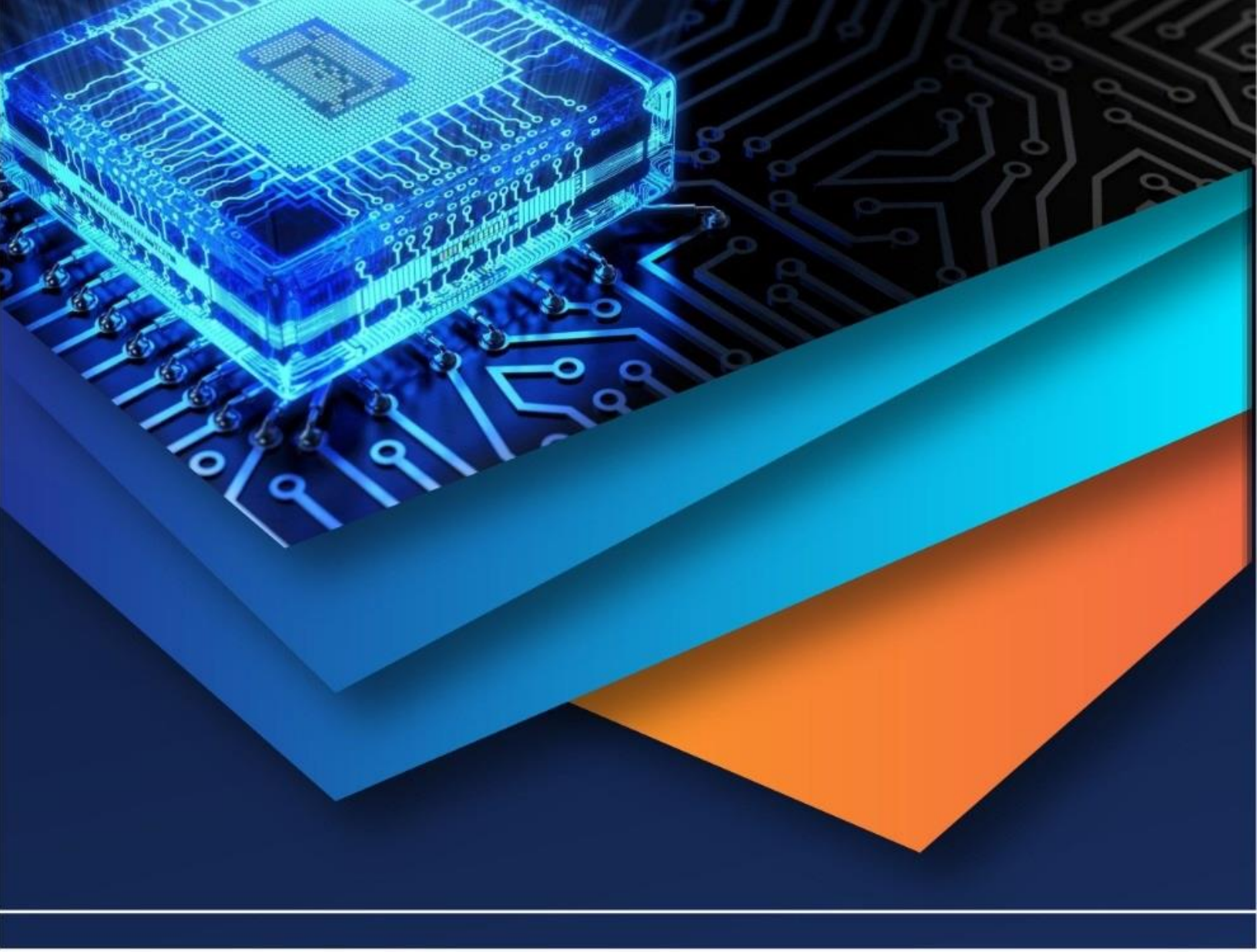
Here in this research work, we provide the prediction for the products, which should be given higher preferences and which should not. Thus, help to improve the business strategy and increase the profits. For this the machine learning approach is used and by making use of five different algorithms we predict the results, which include the Naïve Bayes Classifier (prediction accuracy 93.45%), logistic Regression Classifier (prediction accuracy 93.92%), Support Vector Machine Classifier (prediction accuracy 93.93%), Decision Tree Classifier (prediction accuracy 90.16%), Random Forest Classifier (prediction accuracy 93.50%).
Here by, we conclude that support vector repressor provides the best results and is reliable to provide the output.
Thus, SVM is used to make the further detailing and provide the overall performance analysis which is further increased to 94.08% and finally plot the ROC Curve to show the final results of the overall positive, neutral and negative ratings for the unique products, which can be used for understanding the patterns of the products that should be given higher preferences and which should be eliminated.

Future scope of our research work is as follows:-

A. For Retailers who are new in business
B. Recommended System (for customers)
C. Advertising Brand
D. Competitor Monitoring

REFERENCES

[1] https://becominghuman.ai/sentiment-analysis-of-amazon-product-reviews-93437ad76b59
[2] https://nbviewer.jupyter.org/github/mick-zhang/Amazon-Reviews-using-Sentiment-Analysis/blob/master/Amazon%20Project%20Github.ipynb?flush_cache=true
[3] https://www.kaggle.com/bittlingmayer/amazonreviews
[4] https://towardsdatascience.com/sentiment-analysis-on-amazon-reviews-45cd169447ac
[5] https://towardsdatascience.com/a-complete-sentiment-analysis-algorithm-in-python-with-amazon-product-review-data-step-by-step-2680d2e2c23b
[6] www.rcciit.org.in › projects › csePDF
(Topic: Sentiment Analysis of products-based review using Machine Learning
Author: Dr.Anup Kumar Koyla)
[7] www.cs.rit.edu › pub › ReportPDF
(Topic: A study of Amazon user reviews data using Visualization
Author: Preeti Bamane)
[8] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[9] https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107
[10] https://www.datacamp.com/community/tutorials/decision-tree-classification-python
[11] https://builtin.com/data-science/random-forest-algorithm
[12] https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6
[13] https://www.javatpoint.com/logistic-regression-in-machine-learning
[14] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2
[15] http://cs229.stanford.edu/proj2018/report/122.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY