Fake Product Review

Monitoring System

Submitted in partial fulfillment of the requirements

of the degree of

BACHELOR OF ENGINEERING

In

COMPUTER ENGINEERING

By

Group No: 16

Roll No.	Name
1804007	Musharraf Alam
1804008	Aryan Asawa
1804016	Yash Balchandani

Guide:

PROF. VAISHALI SURYAWANSHI

(Assistant Professor, Department of Computer Engineering, TSEC)

Computer Engineering Department

Thadomal Shahani Engineering College

University of Mumbai

2021-2022

CERTIFICATE

This is to certify that the project entitled "**Fake Product Review Monitoring System**" is a bonafide work of

Roll No.	Name
1804007	Musharraf Alam
1804008	Aryan Asawa
1804016	Yash Balchandani

Submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of "**BACHELOR OF ENGINEERING**" in "**COMPUTER ENGINEERING**".

Prof. Vashali Suryawanshi

Guide

Dr. Tanuja Sarode

Head of Department

Dr. G.T. Thampi

Principal

Project Report Approval for B.E.

Project report entitled (***Fake Product Review Monitoring System***) by

Roll No.	Name
1804007	Musharraf Alam
1804008	Aryan Asawa
1804016	Yash Balchandani

is approved for the degree of "***BACHELOR OF ENGINEERING"*** **in "*COMPUTER ENGINEERING*".**

Examiners

1. ____________________

2. ____________________

Date: ____________________

Place: ____________________

Declaration

We declare that this written submission represents my ideas in our own words and where others 'ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will because for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1. ______________________________
(Signature)

____
(Name of student and Roll No.)

2. ______________________________
(Signature)

____
(Name of student and Roll No.)

3. ______________________________
(Signature)

____
(Name of student and Roll No.)

Date:

Abstract

People require unfeigned information about the online product. Before they can purchase the product itself for one's personal use, a short evaluation from a consumer who has handled the product first-hand elevates the skepticism and enforce the buyer about what to expect when purchasing the product.

But to boost sales and increase profit such evaluations can be forged and generated using a computer-based system. Natural language processing algorithm can help us to find the difference between the real and fake consumer, to ensure the authenticity of the consumer and to maintain integrity between customer-seller relationship. This project will use datasets of reviews of products and predict that if this is a machine generated review or a review written by a real consumer.

TABLE OF CONTENTS

List of Figures

Chapter 1

Introduction

1.1 Introduction

Over the past few years, E-commerce has been one of the fastest-growing markets. This market reached its peak during the global pandemic of 2020-2021. During this time many business-to-consumer businesses shifted online as maintaining sales targets proved difficult in a traditional method. Before purchasing anything, it is normal human behavior to do research and survey on that product to check its reliability. The most common way to finalize the selection of the product is to check the rating of the product which is done mostly on the scale of 1 to 5 and the customer review from the customer who has used the product first hand and compares the same category of the product with different brands and select the most appropriate product based on an individual need. But this review and customer feedback satisfaction loop can be exploited by any brand and use to manipulate the sales of the product of the target brand by introducing fake reviews into the system from multiple accounts. This can lead to a decrease in the popularity of the brand or an increase in the popularity of one brand product while discrediting the competitor's products. These fake reviews can often be misleading and can directly influence the opinion of tens of thousands of customers who may eventually end up buying something that they did not necessarily desire.

Machine learning concepts like text classification, sentiment-analysis, information retrieval, text summarization, next-word prediction, language translation and detection, etc. can be used to analyze the sentiment of the review text and find out whether the review was genuine or false. Our algorithm uses the data from Amazon and searches for the keyword that have been repeated many times it also checks the sentiment of each review with its rating. And analyze it further to test the integrity of the review posted by a user on the particular product.

1.2 Aim & Objectives

The main aim of this project is to develop a system that can identify a 'fake' and 'genuine' review for any product in the dataset. The project should be able to analyze the ‘review_text’ in the dataset and assign a suitable sentimental score which can be further used to classify a review as positive, negative or neutral. It should be able to rate the sentiment of the product based on its star rating. Also, be able to compare the difference between the two for better identification. The application should also be able to detect spam reviews and sort the reviews which were found helpful by the consumer and which were not among the reviews posted also check the integrity of the project.

1.3 Scope

The Project scope pertains to the work necessary to deliver a working 'Fake product review monitoring system'. The project should be able to work on the larger chunks of the dataset and successfully classify the reviews as genuine and fake reviews. The project will provide businesses and brands to find out the flaws in their review system and help them develop a review collecting system. Also, it will help in reducing the significant amount of misleading information about a product or a brand. Furthermore, it will promote fair competition among the seller which is eventually a win-win condition for both consumer and seller where the former can get a quality product and later can make a quality product and an influential presence in the free-market economy.

Chapter 2

Review of Literature

2.1 Domain Explanation

- **Machine Learning**

Machine Learning (ML) is that field of computer science with the help of which computer systems can provide sense to data in much the same way as human beings do. In simple words, ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method. The key focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention. Machine learning has many applications one of which is to be able to detect sentiments from the texts through mining and analysis techniques. Using these techniques, a machine learning model can find the tone of speech of a text. Also, it can be used for contextual mining of the text which makes textually data comparable in a numerical sense.

The main feature for this project is to classify the text data in the review as fake or genuine while also taking in consideration of the data from the rating and helpfulness of the rating to formulate a final opnion on the review.

The data operation that will be perform on the dataset includes

1. **GridSearch:** Exhaustive search over specified parameter values for an estimator.
2. **Cross-Validation:** In the train-test split, only 20% of the data for testing is used. The performance metric on that 20% test data may not be accurate. So, Cross-Validation allows a developer to consume 100% of the data for training and testing both.

The project uses the following algorithms for the comparative study:

1. GaussianNB: GaussianNB implements the Gaussian Naive Bayes algorithm for classification. In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell-shaped curve that is symmetric about the mean of the feature values.

$$P(X_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(\frac{-(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

2. Logistic Regression: Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt), or the log-linear classifier

3. Decision tree: DecisionTreeClassifier is a classifier capable of performing both binary and multi-class classification on a dataset. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

It also makes use of concepts like:

1. *True Positive*: the truth is positive, and the test predicts a positive. e.g., The person is sick, and the test accurately reports this.

2. *False Positive*: the truth is negative, but the test predicts a positive. e.g., The person is not sick, but the test inaccurately reports that they are. It is also called a Type I error in statistics.

- ## Sentimental Analysis

Sentiment Analysis is a Natural Language Processing technique to determine the sentiment or opinion of a given text. A sentiment analysis model can predict whether a given text data is positive, negative, or neutral by extracting meaning from the natural language and assigning it to a numerical score. There are various ways to develop or train a sentiment analysis model, in this article 5 different ways are discussed:

- Custom Trained Supervised Model
- TextBlob
- Word-Dictionary based model
- Bert
- Named Entity based Sentiment Analyzer

Sentiment analysis is used by various organizations to understand the sentiment of their customers, using reviews, social media conversations, and to more fast and accurate business decisions accordingly

This project will use Textblob to process textual data and allows to specify which algorithms to use for the faster classification and identification of the reviews. This project implements aspect-based sentiment analysis in different ways. The most famous is Sentiment Intensity Analyzer (commonly known as SIA) from vaderSentiment.

But it is found that sentiment from Textblob gives better results than SIA here. Thus, the result sentiment from Textblob have been used for this project.

Apart from sentimental analysis the project also utilizes the important concept from Natural Language processing.

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it may possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

$$TF(t) = \frac{No. of\ times\ term\ t\ appears\ in\ a\ document}{Total\ no. of\ terms\ in\ the\ document}$$

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, to weigh down the frequent terms while scaling up the rare ones, by computing the following:

$$IDF(t) = \log_e\left(\frac{Total\ no. of\ documents}{No. of\ documents\ with\ term\ t\ in\ it}\right)$$

For the ‘textual analysis’ of the given dataset, the project utilizes Lemmatization, *Lemmatization* is a process of extracting a root word by considering the vocabulary. For example, “good,” “better,” or “best” is lemmatized into good. The part of speech of a word is determined in lemmatization. It will return the dictionary form of a word, which must be valid while stemming just extracts the root word.

Stemming is not used as it doesn’t return the dictionary-based synonyms of the word.

2.2 Existing Solution

Different processes were proposed in past to understand counterfeit audits dependent on styles of information like marked records (as an example, directed learning), unlabeled data (as an instance, unaided learning), and in element named records (for instance, semi regulated discovering) that is portrayed underneath. A. Directed Learning approaches make use of administered studying calculation for counterfeit audit discovery.[5]

Opinion Mining Using Ontological Spam Detection:

For Spam Detection Duhan & Mittal proposed a paper "Opinion Mining Using Ontological Spam Detection" that will help us to find out fake reviews by using Naïve Bayes as an algorithm. To find out fake reviews on the website this "Fake Product review Monitoring System" system is introduced. [2] This system will find out fake reviews made by the customers and it will block the users. To find out the review is fake or genuine, the models will use some classifications such as:

- ***Tracking the IP:*** address of the user to detect if the reviews are from a Spammer. If multiple reviews are from the same IP address, then the Reviews are considered Spam.[1]
- ***Using Account Used***: to check whether the reviews are done using the same account.[1]
- ***Brand only Review detection:*** i.e., whether the reviews are on only Brand, not the product. It's not helpful to consider only the Brand value to judge a product.[3]
- ***Using Negative Dictionary:*** i.e., the negative words are identified in the review. If there are more than five Negative Words then the review is Spam. [4]

2.3 H/W & S/W requirement

The software will run of on any system that fulfills the minimum hardware requirements such as:

- Hardware Requirements: -
    1) Core i5/i7 processor
    2) At least 8GB RAM
    3) At least 60 GB of Usable Hard Disk Space

Apart from the hardware additional software installation is essential so that the client can run the software. Also, python dependencies for Machine Learning and Sentimental Analysis will be required and the packages must be bundled in a (virtual environment) venv which is essential for ML applications this will make error handling easier and avoid any conflicts amidst other applications sharing the resource.

- Software Requirements: -
    1) Python 3.x
    2) Google Colab or Jupyter Notebook
    3) Operating System
    4) Anaconda
    5) Pandas
    6) Natural Language Toolkit
    7) Matplotlib
    8) TextBlob

Chapter 3
Analysis

3.1 Functional Requirement

Functional requirements are the functions or features that must be included in any system to satisfy the business needs and be acceptable to users. Since, the project utilizes Machine Learning Model is used to classify the category to which the review belongs and highlight the importance of which feature holds the utmost importance in classification. Based on this, the functional requirements that the current system will require are as follows: -

- The system should be able to simplify the early raw input data so that it could be preprocessed and mined further.
- The model should be able to experiment, automatic scale and apply normalization techniques to the dataset by default, for assisting the algorithms sensitive to feature.
- The system will use multiple algorithms (3 algorithms in this case study) to find out the best possible for the classification of the reviews.
- The system should be able to highlight the significance of the feature used in classification.
- Since this project is based on the problem of binary classification (helpful or not helpful), roc_auc_score is used to evaluate the models.
- The training speed and prediction speed for a sample size of 40,000 which will be the ideal requirement for passing the test.
- The system should be able to rate the sentiment based on the rating of the product and be able to compare it to the review sentiment of the 'text review' of the product.

Once classification is done it should be able to evaluate the final sentiment and compare it to the written sentiment and differentiate the reviews as fake or genuine.

3.2 Non-Functional Requirement

Non-Functional Requirements are a description of features, characteristics and attributes of the system as well as any constraint that may limit the boundaries of the proposed system. The Non-Functional requirements are essentially based on the performance, information, economy, control and security efficiency and services. Based on this, the non-functional requirements are as follows: -

- **Performance Requirements**: -As for this project version performance will be evaluated on detecting if the system is crashed, hanged or an operating system error has occurred. Also detecting the performance of the system in terms of efficiency of the integration of different components

- **Safety Requirements**: - For the safety requirements operation of weekly backups for the database will take place. Thus, ensuring the safety and integrity of the data.

- **Security and Privacy Requirements**: -There is no specific security placed for now since it will be used directly by an expert at hand requirements.

- **Software Quality Attributes**: - The software is reliable and works under the minimal hardware requirement with the optimal performance efficiency of 90%.

- **Reliability** – The solution provided reliability to the oncologist conducting the test & the product will run all the features mentioned in this document perfectly. It has been tested and debugged completely. All exceptions are well handled.

3.3 Proposed System

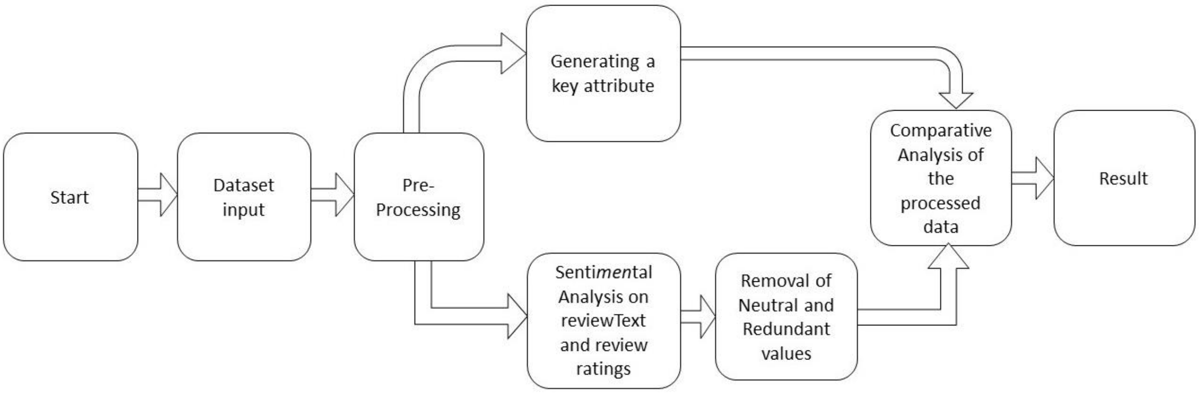


Fig 1. Block diagram of the Proposed System

Introductions:

"Fake review monitoring system" will help understand the true market sentiment on a particular product that belongs to any particular brand for a particular category. it will also eliminate the stress and resources required to identify the impractical reviews and help brands identify their consumer's genuine problems related to the product which will in return help in the improvement of the brand.

Resources Required*:*

The brief details about resources are mentioned in *chapter 2, section 2.3 H/W and S/W requirement*. Since this is more like an identification system for the organization a separate storage unit is unnecessary addition. A separate system will run the application and identify the review authenticity by connecting itself to the data storage unit that stores every data about the project including the reviews.

Benefit to the organization:

1. Qualitative data processing for improvements in the product.
2. Better understanding about consumer needs.
3. Removal of futile data thus saving space for storage and processing.
4. Improvement in the brand reputations and credibility for delivery up to the expected demands.

Chapter 4

Design

4.1 Design Consideration

The project design consists of two parts implementations**:**

Part 1: Prediction of Helpfulness from given data.

The discrepancy to look after here is: Amazon Reviews of poor quality at the top of forums despite of the helpfulness rating system of Amazon. This problem mainly arises due to the new reviews are directly placed on top of the forum which would give the community a chance to rate them.

The solution to this problem is to create a model using machine learning techniques that would pre-rate the helpfulness of a given customer review before they are posted on the top of the forum. This pre-rating acts as a filter for poor quality reviews to be not shown on top of forums.

The model will be trained on Amazon Reviews for Electronic Category to predict if a given review is helpful or not helpful.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	AO94DHGC771SJ	0528881469	amazdnu	[0, 0]	we got this gps for my husband who is an otr o...	5	Gotta have GPS!	1370131200	06 2, 2013
1	AMO214LNFCEI4	0528881469	Amazon Customer	[12, 15]	im a professional otr truck driver and i bough...	1	Very Disappointed	1290643200	11 25, 2010
2	A3N7T0DY83Y4IG	0528881469	C. A. Freeman	[43, 45]	well what can i say ive had this unit in my tr...	3	1st impression	1283990400	09 9, 2010
3	A1H8PY3QHMQQA0	0528881469	Dave M. Shaw "mack dave"	[9, 10]	not going to write a long review even thought ...	2	Great grafics, POOR GPS	1290556800	11 24, 2010
4	A24EV6RXELQZ63	0528881469	Wayne Smith	[0, 0]	ive had mine for a year and here what we got i...	1	Major issues, only excuses for support	1317254400	09 29, 2011

Fig 2. Snippet of the dataset

Based on the above dataset it can be concluded that:

Input Features: reviewText, overall

Output labels: helpfulness

Reason for selecting input features: When a review of a product is posted; along with textual feedback(reviewText) of the review, a rating in stars(overall) is also given for numerical analysis of customer satisfaction.

A brief explanation about the helpful column: helpful column given above is a list containing two values 'no of helpful ratings' and the 'total no of ratings' separated by a comma.
Int this project the helpful column is divided into two parts i.e.

helpful_numerator => contains no. of helpful rating.

helpful_denominator => contains total no. of ratings.

and then the helpful column is dropped.

Problem Statement 2: Classification of genuine and fake/sarcastic reviews.

Since a customer cannot examine the product first hand, they have to depend on reviews of the given product.
This dependency is exploited by some people who do not give genuine review or give a sarcastic review.
Here are some examples:

Fig 3. Review snippet from amazon's website

Having such reviews displayed in the review section negatively affects Amazon's business, as stated above the people are willing to buy consumer goods online without examining the items firsthand and for that they have to access to other people's opinion of the item.
The area concern which is being addressed here is if Amazon Review is fake or not genuine. Sometimes some people give fake reviews on forums. This could negatively affect the sale of the product if such a review came on top.

What will be considered as a fake review are: the review that are given a 5-star rating but in textual feedback of the review the person gives a negative review and since the star rating sentiment conflicts the textual feedback the review it will be considered as fake.

There exists a possibility that such reviews be given by competitors to harm the sale of the product.

Model which uses machine learning techniques to classify given customer's review based on sentiment analysis before they are posted on the top of the forum is a possible solution to the pertaining problem. Through this fake reviews would be identified before being shown on top of forums.

4.2 Design Details

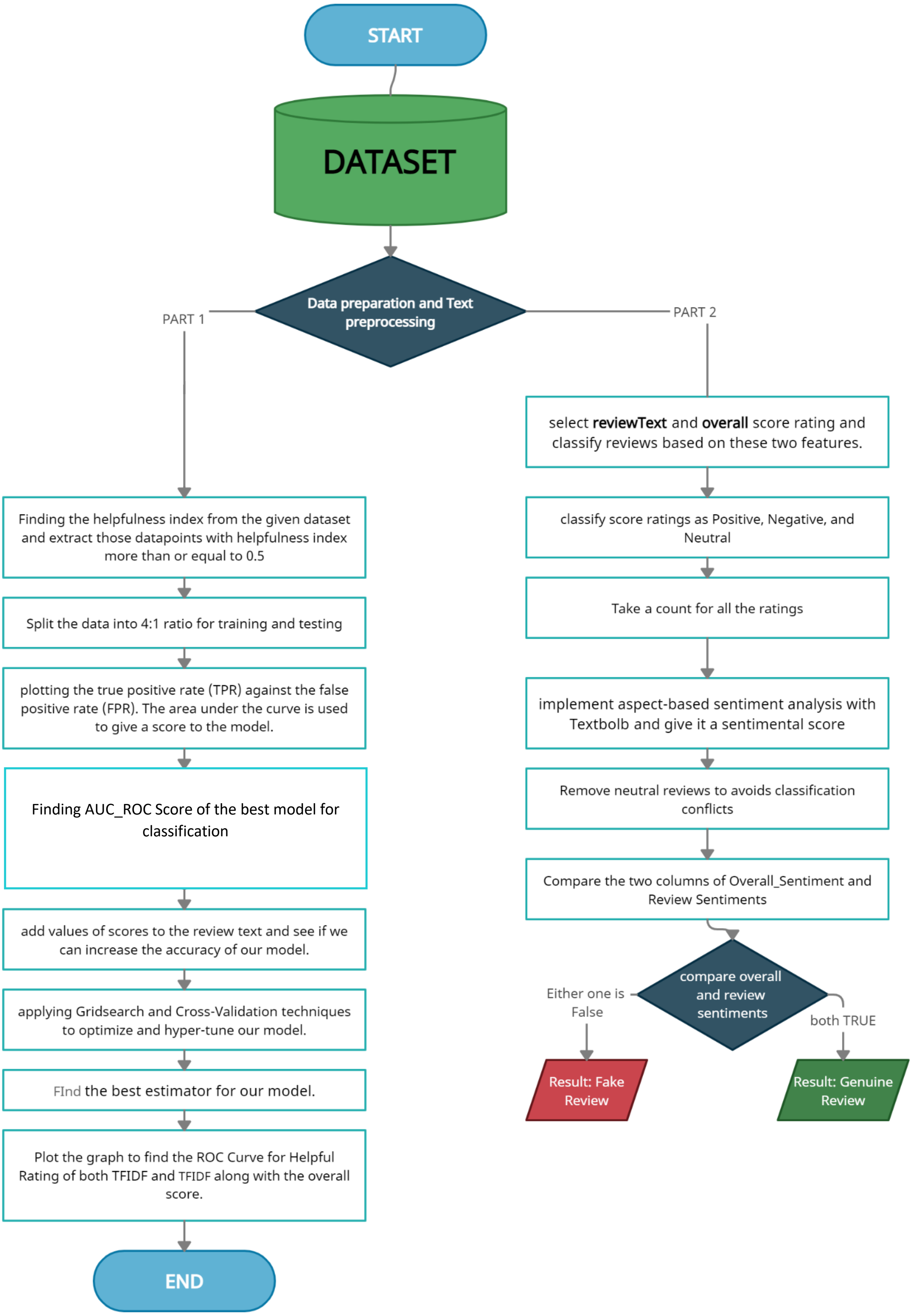


Fig 4. Flowchart for the Project

Preprocessing of the data:

Before starting with the problem statement, data preparation must be completed.

First, let's import all required dependencies that will be used in the project as stated in chapter 2 section 2.3.

Now read the dataset.

The dataset is a JSON file so read_json() function of Pandas is used. Here, lines=True is used to read the file as a JSON object per line, else it will give an error.

This dataset contains - 5-core dataset of product reviews from Amazon Electronics category from May 1996 - July 2014. One set of the total dataset contains total of 1689188 entries. The dataset also contains at the least 5 reviews per Customer and also 5 reviews each product.

Dataset Attributes are:

- asin - ID of the product assigned by Amazon, Eg: B00LGQ6HL8
- helpful - helpfulness rating of the review - example: 2/3.
- overall - rating of the product.
- reviewText - text of the review (heading).
- reviewTime - time of the review (raw).
- reviewerID - ID of the reviewer, Eg: AO94DHGC771SJ
- reviewerName - name of the reviewer.
- summary - summary of the review (description).
- unixReviewTime - unix timestamp.

Text Preprocessing:

Since, this project implements the solution in two sub-problems, both of which require the same kind of preprocessing. The preprocessing will be done before starting with the problem statements.

Preprocessing before going to problem statements will save a lot of time since just to preprocess once it requires about an hour.

In preprocessing, first 'reviewText' attribute is converted into lowercase. Then further preprocessing on the dataset is done. And finally, lemmatization is applied to sort and select sentences which matches positive, neutral and negative experience.

Here, Lemmitization is chose over stemming as lemmatization uses the dictionary for matching

the textual input and root word extraction unlike stemming which just performs string operations on any given textual data.
An example of it would be Lemmatization handles matching “trucks” to “truck” along with matching “truck” to “vehicle” while Stemming handles matching “truck” to “trucks”.

```
Out[40]: 0    we got this gps for my husband who is an otr o...
         1    im a professional otr truck driver and i bough...
         2    well what can i say ive had this unit in my tr...
         3    not going to write a long review even thought ...
         4    ive had mine for a year and here what we got i...
         Name: reviewText, dtype: object
```

Fig 5. Lemmatization on the reviewText Attribute

Part 1 Implementation:

To create a model using machine learning techniques that would pre-rate the helpfulness of a given customer review before they are posted on the top of the forum. This way poorly formulated reviews would not be displayed on top of the review section of product.

Based on the above dataset, the conclusion for input features and output labels will be:

Input Features: reviewText,overall

Output labels: helpfulness

Reason for selecting input features: When a given review is posted it is done so along with textual feedback(reviewText) of the review and rating in stars on the scale of 1 to 5(overall).

A brief explanation about the helpful Attribute: helpful column given above is a list containing two values ‘no of helpful ratings’ and the ‘total no of ratings’ separated by a comma.

the helpful attribute will be divided into two sub-attributes i.e.

1. helpful_numerator => contains no. of helpful rating.
2. helpful_denominator => contains total no. of ratings.

and then the helpful column will be dropped from the processing of the dataset.

Out[43]:

	overall	helpful_numerator	helpful_denominator
count	1.687169e+06	1.687169e+06	1.687169e+06
mean	4.222604e+00	3.129790e+00	3.747861e+00
std	1.185743e+00	3.865623e+01	4.035785e+01
min	1.000000e+00	0.000000e+00	0.000000e+00
25%	4.000000e+00	0.000000e+00	0.000000e+00
50%	5.000000e+00	0.000000e+00	0.000000e+00
75%	5.000000e+00	1.000000e+00	2.000000e+00
max	5.000000e+00	3.073500e+04	3.145300e+04

Fig 6. Dataset after splitting helpful column

Here, since the dataset use is huge (about 1687169 records), only those records that have at least 20 ratings in total are taken under for evaluation. (This is under the assumption that people will vote for review they found helpful)

Here to get the output label helpfulness, the ratio of helpful_numerator to helpful_denominator should be calculated. The result is compared with a threshold value (Here it is taken as 50%, but it can be changed as per the requirement).

- If result > threshold ==> helpful = 1
- if result < threshold ==> not helpful = 0

Out[45]:

	overall	reviewText	helpful_numerator	helpful_denominator	Helpful
2	3	well what can i say ive had this unit in my tr...	43	45	1
211	3	i purchased this mount for my inch lcd tv inst...	70	92	1
221	5	do not listen to other reviewer screw that thi...	18	23	1

Fig 7. Count on the total helpful and non-helpful review

Now, let's do the count of data to get an idea about the distribution of helpfulness.

Count:

Helpful	overall	reviewText	helpful_numerator	helpful_denominator
0	4896	4896	4896	4896
1	45289	45289	45289	45289

Fig 8. Count on the total helpful and non-helpful review stats

Since, dataset is already prepared. TF-IDF Vectorizer is applied to generate more features and also to find the statistical measure of how important a word is to document in a collection or corpus. It is generally used in text mining and information retrieval.

Since a separate dataset for testing is not required, the dataset is splitted as 80% for training and 20% for testing.

Since, the problem is of binary classification (helpful or not helpful), roc_auc_score is used to evaluate the models used.

The roc_auc_score calculates the area under the receiver operating characteristic (ROC) curve which is also denoted by AUC or AUROC. The curve information is summarized in one number through the calculated area under the curve.

This curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). The area under the curve is for assigning a score to the model.

If AUC = 0.5 => TPR = FPR, the model is computing the values randomly (not an ideal model).

If AUC= 1.0 => TPR = 100%, and it is an ideal model.

To make a baseline model for this project 3 of the following algorithms will be used:

- **Gaussian Naïve Bayes**
- **Logistic Regression**
- **Decision tree**

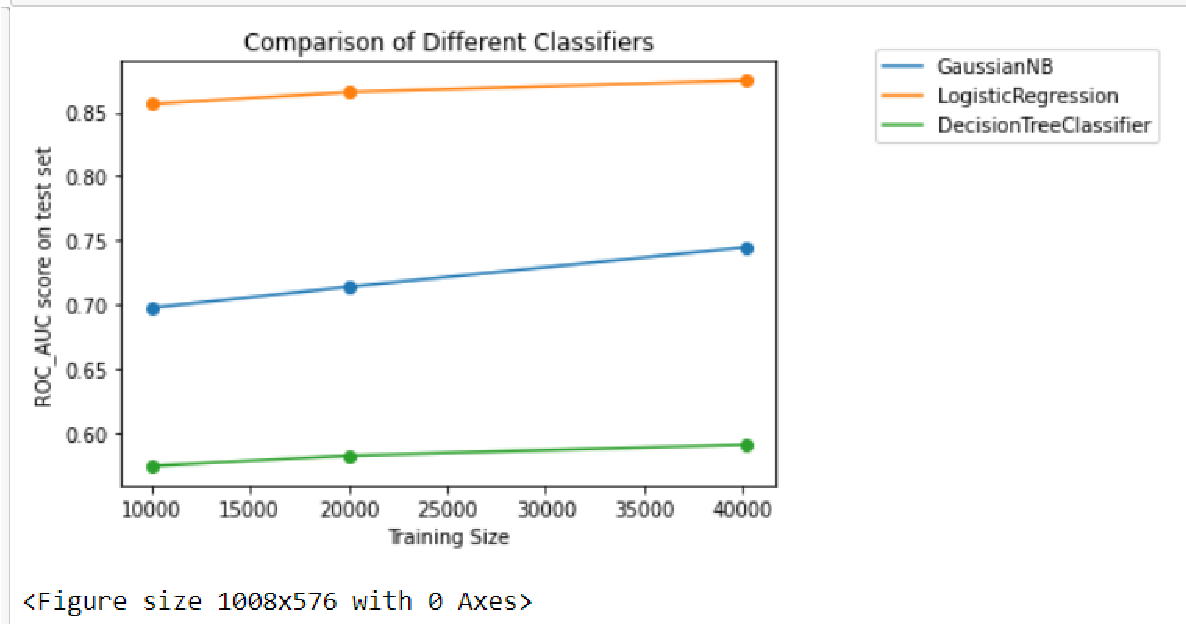


Fig 9. AUC_ROC score for all 3 algorithms

Logistic Regression gives us the best accuracy. Its final score for the area under the ROC curve was 0.8704 and a sample size of ~40,000. Besides, it is the fastest. The training speed and prediction speed were 19.993s and 0.955s for a sample size of 40,000. Since our model has to consider the accuracy and speed, the Logistic Regression algorithm represents the ideal model for us.

Now, let's add values of scores to the review text and see if the accuracy of the model can be increased.

Now the dataset can be split for optimization of the initial model.

```
X_train2, X_test2, y_train, y_test = train_test_split(features, df1['Helpful'], test_size=0.2, random_state=RAN_STATE)
```

Hyper-parameters are parameters that are not directly learned within estimators. In scikit-learn, they are passed as arguments to the constructor of the estimator classes. Hyperparameter tuning helps us in optimizing our model.

We will now be applying Gridsearch and Cross-Validation techniques to optimize and hypertune on the model.

The best estimator for the model are depicted in Fig.

```
{'C': 1, 'class_weight': None}
```

Fig 10. The best estimator.

Here the optimized classifier is a Logistic Regression with a 'C' parameter of 1 and a 'class_weight' = 'None'. This is the same as default, meaning the optimization step did not change the parameters of the model. Let's find the ROC_AUC Score.

```
ROC_AUC Score: 0.9044773279046178
```

90 % ROC_AUC Score. That means the model has been trained well.

Plotting the ROC Curve for Helpful Rating of both TFIDF and TFIDF with the overall score for comparative understanding of the which one should be used in implementation.

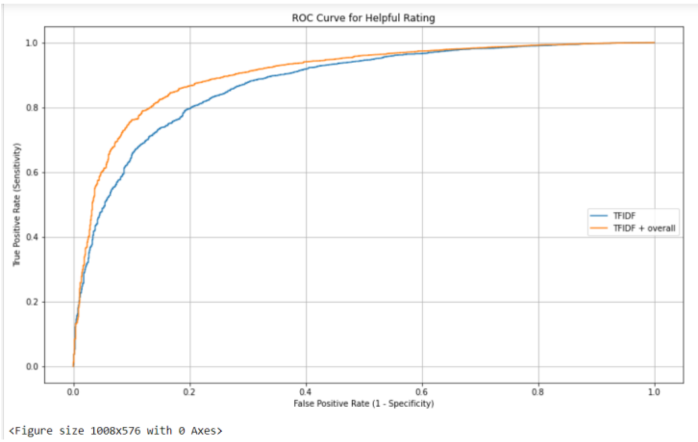


Fig11. Roc score before and after the addition of TFIDF to the overall result

Implementation of Part 2:

To create a model using machine learning techniques that would classify given customer reviews based on sentiment analysis before they are posted on the top of the forum. Thus, fake reviews would be filtered before being displayed in the review section of the Amazon Webpage of the product.

The model is trained on Amazons Electronic Reviews Category to classify if a given review is genuine or not.

Here, **reviewText** and **overall** score rating is selected since to classify the reviews on those two features will be required.

Out[60]:

	overall	reviewText
0	5	we got this gps for my husband who is an otr o...
1	1	im a professional otr truck driver and i bough...
2	3	well what can i say ive had this unit in my tr...
3	2	not going to write a long review even thought ...
4	1	ive had mine for a year and here what we got i...
5	5	i am using this with a nook hd it work a descr...
6	2	the cable is very wobbly and sometimes disconn...
7	5	this adaptor is real easy to setup and use rig...
8	4	this adapter easily connects my nook hd to my ...
9	5	this product really work great but i found the...

Fig 12. Dataset with overall and reviewText column.

Now, the score ratings are classified as Positive, Negative, and Neutral.

- If the **score rating => 4 or 5**, it is taken as **Positive**.
- If the **score rating => 3**, it is taken as **Neutral**.
- And if the **score rating => 1 or 2**, it is taken as **Negative**.

Then the result is saved in a dataset newly defined attribute called '**Overall_Sentiment'**. This would tell us the sentiment of the review based on the star rating.

Out[61]:

	overall	reviewText	Overall_Sentiment
0	5	we got this gps for my husband who is an otr o...	Positive
1	1	im a professional otr truck driver and i bough...	Negative
2	3	well what can i say ive had this unit in my tr...	Neutral
3	2	not going to write a long review even thought ...	Negative
4	1	ive had mine for a year and here what we got i...	Negative
...	...	...	...
1689183	5	burned these in before listening to them for a...	Positive
1689184	5	some people like dj style headphone or earbud ...	Positive
1689185	5	i m a big fan of the brainwavz s actually all ...	Positive
1689186	5	ive used thebrainwavz s in ear headphone and t...	Positive
1689187	5	normally when i receive a review sample i can ...	Positive

1687169 rows × 3 columns

Now, let's find out how the number of Positive, Negative, and Neutral reviews.

Fig 13. Dataset after sentimental analysis on overall.

```
Positive     1354351
Negative      190693
Neutral       142125
Name: Overall_Sentiment, dtype: int64
```

Fig 14. The total categorical count.

Here, 1354351 reviews are Positive,190693 are Negative while 142125 are Neutral.

But it may also be possible that some data may be missing or null.

This 'null values' will be dropped here if there are any present in current dataset.

A rechecking is done to reassure the number of reviews.

The above step leads us to:

Positive 1354351

Negative 190693

Neutral 142125

Name: Overall_Sentiment, dtype: int64

There are no missing or null values.

Thus, let move to the next step. For sentiment analysis of review text.

Sentiment analysis is a process which we computationally analyzes and identifies opinions and judgments from a piece of text, in this project the piece of text would be the reviewText. One can understand if a piece of text is positive, negative, or neutral, based on their sentiment analysis.

There are various types of sentiment analysis, but here aspect-based sentiment analysis is used. Aspect-based sentiment analysis is mostly for analyzing one or more aspects of a service or product i.e., a typical objective review.

For example, if a company that sells cars uses this type of sentiment analysis, it could be for one aspect of cars – like Engine, transmission, interior, etc.

So, they can understand how customers feel about specific attributes of the product.

	overall	reviewText	Overall_Sentiment	Review_Sentiment	sentiment score
0	5	we got this gps for my husband who is an otr o...	Positive	Positive	0.250000
1	1	im a professional otr truck driver and i bough...	Negative	Positive	0.062441
2	3	well what can i say ive had this unit in my tr...	Neutral	Positive	0.086977
3	2	not going to write a long review even thought ...	Negative	Positive	0.047284
4	1	ive had mine for a year and here what we got i...	Negative	Positive	0.002778
5	5	i am using this with a nook hd it work a descr...	Positive	Positive	1.000000
6	2	the cable is very wobbly and sometimes disconn...	Negative	Negative	-0.100000
7	5	this adaptor is real easy to setup and use rig...	Positive	Positive	0.274439
8	4	this adapter easily connects my nook hd to my ...	Positive	Positive	0.297718
9	5	this product really work great but i found the...	Positive	Positive	0.212487

Fig 15. Dataset after sentimental scoring on the reviewText

In the dataset some reviews are of Neutral sentiment based on star rating while others are Positive, based on the sentiment of review text.

Such reviews may be classified incorrectly by the model.

So, all Neutral reviews are removed.

The number of neutral labels has been removed: 176941

Significant amount of the reviews has been removed thus making processing on the remaining

Now, let's classify reviews as true or false.

- ***True review***: review in which the sentiment of star rating matches the sentiment of review text.
- ***False review***: review in which the sentiment of star rating does not matches the sentiment of review text. Comparison done below.

	overall	reviewText	Overall_Sentiment	Review_Sentiment	sentiment score	result
0	5	we got this gps for my husband who is an otr o...	Positive	Positive	0.250000	True
1	1	im a professional otr truck driver and i bough...	Negative	Positive	0.062441	False
3	2	not going to write a long review even thought ...	Negative	Positive	0.047284	False
4	1	ive had mine for a year and here what we got i...	Negative	Positive	0.002778	False
5	5	i am using this with a nook hd it work a descr...	Positive	Positive	1.000000	True

Fig 16. Dataset after labelling of reviews as false and True

Finding the total number of all the fake reviews.

And running the data check on the fake review data-set that is being generated.

	overall	reviewText	Overall_Sentiment	Review_Sentiment	sentiment score	result
1	1	im a professional otr truck driver and i bough...	Negative	Positive	0.062441	False
3	2	not going to write a long review even thought ...	Negative	Positive	0.047284	False
4	1	ive had mine for a year and here what we got i...	Negative	Positive	0.002778	False
22	4	this wall mount doe everything it supposed to ...	Positive	Negative	-0.092143	False
36	5	didnt think it would work a well a it hasbecau...	Positive	Negative	-0.063889	False

Fig 17. Dataset after extraction of false/fake reviews

Here, there are all the reviews that are not genuine and in figures it amounts to 202988 reviews.

The Following reviews will be removed by the project as it has been classified as fake.

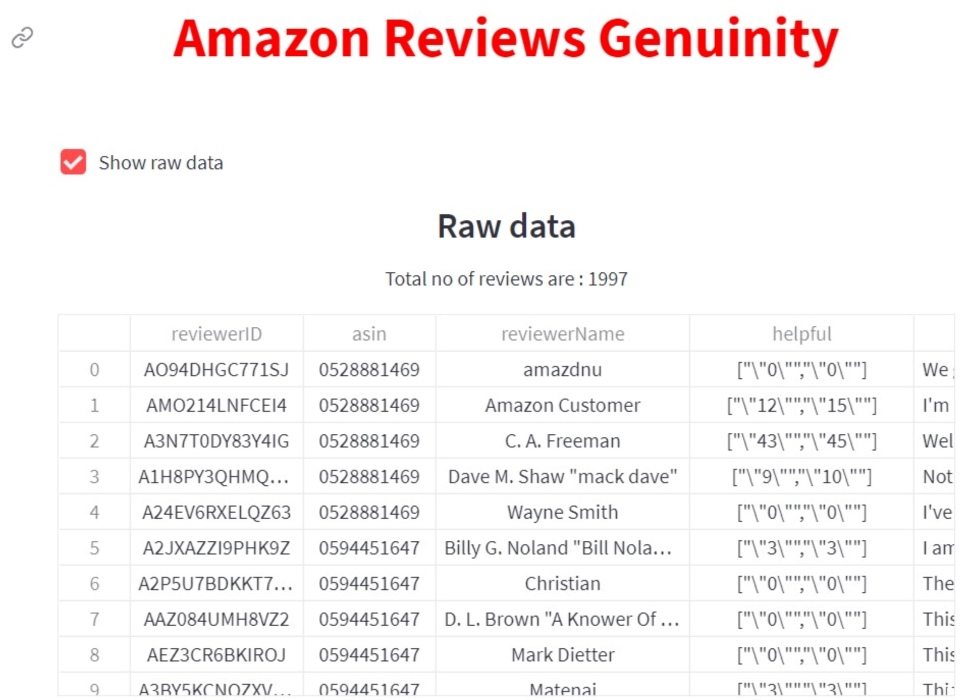
4.3 Front-End Design

The backend of the Model has been running well and is generating the expected result for the user. But a normal user (with zero backend experience) can find it difficult to navigate through all the lines of code. To solve this issue, Steamlit is used to develop the Front-End for this project.

Streamlit architecture allows developer to write apps the same way as plain Python scripts. This can be unlocked by Streamlit apps through a unique data flow: if any updates must be updated on the screen, Streamlit will rerun the entire Python script from top to bottom.

This can happen in two situations:

- Whenever a developer modifies their app's source code.
- Whenever a user interacts with widgets in the app. For example, when dragging a slider, textual input box, or a button for clicking.



	reviewerID	asin	reviewerName	helpful	
0	AO94DHGC771SJ	0528881469	amazdnu	["\"0\"","\"0\""]	We
1	AMO214LNFCEI4	0528881469	Amazon Customer	["\"12\"","\"15\""]	I'm
2	A3N7T0DY83Y4IG	0528881469	C. A. Freeman	["\"43\"","\"45\""]	Wel
3	A1H8PY3QHMQ...	0528881469	Dave M. Shaw "mack dave"	["\"9\"","\"10\""]	Not
4	A24EV6RXELQZ63	0528881469	Wayne Smith	["\"0\"","\"0\""]	I've
5	A2JXAZZI9PHK9Z	0594451647	Billy G. Noland "Bill Nola...	["\"3\"","\"3\""]	I am
6	A2P5U7BDKKT7...	0594451647	Christian	["\"0\"","\"0\""]	The
7	AAZ084UMH8VZ2	0594451647	D. L. Brown "A Knower Of ...	["\"0\"","\"0\""]	This
8	AEZ3CR6BKIROJ	0594451647	Mark Dietter	["\"0\"","\"0\""]	This
9	A3BY5KCNQZXV...	0594451647	Matenai	["\"3\"","\"3\""]	Thi

Fig 18. Raw Data Set Column

The fig.18 panel is a part of the frontend that shows the raw input data used for Amazon and displays the respective columns present in the dataset.

See Reviews based on Score Ratings:

5

1 5

	reviewerID	asin	reviewerName	helpful	
0	AO94DHGC771SJ	0528881469	amazdnu	["\"0\"","\"0\""]	We
5	A2JXAZZI9PHK9Z	0594451647	Billy G. Noland "Bill Nola…	["\"3\"","\"3\""]	I am
7	AAZ084UMH8VZ2	0594451647	D. L. Brown "A Knower Of …	["\"0\"","\"0\""]	This
9	A3BY5KCNQZXV…	0594451647	Matenai	["\"3\"","\"3\""]	This
11	A3HICVLF4PFFMN	0594481813	Amazon Customer	["\"0\"","\"0\""]	bou
12	ANSKSPEEAKY7S	0594481813	Gena	["\"1\"","\"1\""]	My
14	ANY6JUFM0GH8U	0594481813	J. Clement	["\"2\"","\"2\""]	I los
18	A20XXTXWF2TCPY	0972683275	<NA>	["\"0\"","\"0\""]	This
19	A2IDCSC6NVONIZ	0972683275	2Cents!	["\"1\"","\"1\""]	This
20	A1EDI0X3GU1SK7	0972683275	AGW	["\"0\"","\"1\""]	Gre

Fig 19. Slider as per ratings in star on a scale of 1 to 5(sorting)

The panel in fig. 19 is a part of the frontend that sorts reviews as per the given star rating by the customer on any particular product.

Division of Reviews based on Score:

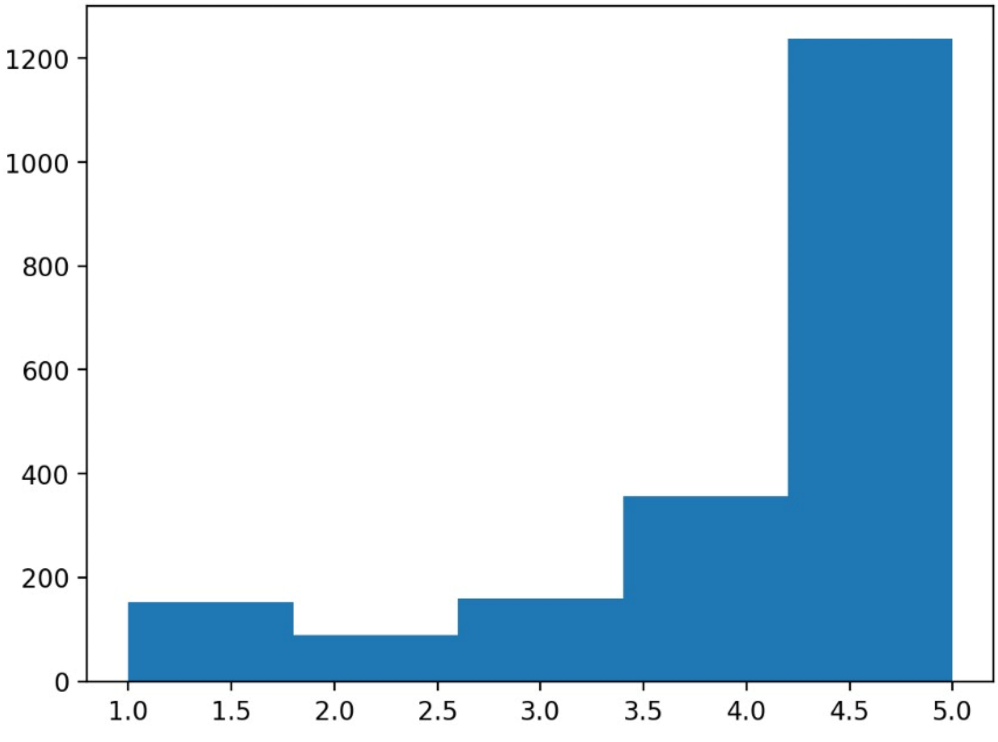


Fig 20. Graphical Division of the review w.r.t ratings

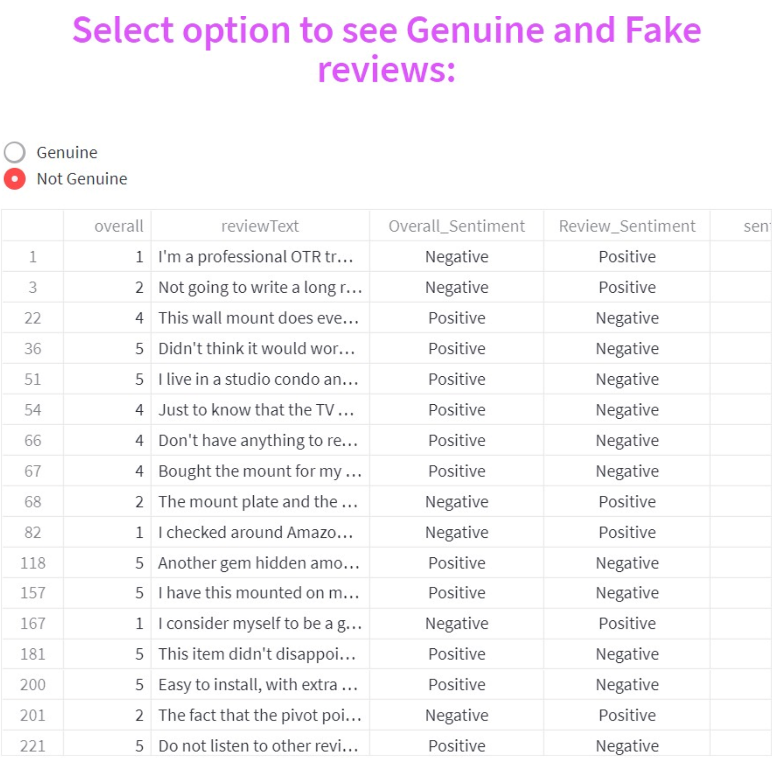
This scale in the graphs Y-axis is in the scale of 10^3. The Graph shows the amount of review that's are of a particular star rating. So, here reviews with 4.5 are 1200, by the scale calculation it would be $1200*10^3$ = 12,00,000 reviews above 4.5.

Select option to see Genuine and Fake reviews:

- (•) Genuine
- ( ) Not Genuine

	overall	reviewText	Overall_Sentiment	Review_Sentiment	sen
0	5	We got this GPS for my hu…	Positive	Positive	
4	1	I've had mine for a year a…	Negative	Negative	
5	5	I am using this with a Noo…	Positive	Positive	
6	2	The cable is very wobbly …	Negative	Negative	
7	5	This adaptor is real easy t…	Positive	Positive	
8	4	This adapter easily conne…	Positive	Positive	
9	5	This product really works …	Positive	Positive	
10	4	This item is just as was de…	Positive	Positive	
11	5	bought for a spare for my …	Positive	Positive	
12	5	My son crewed my HD ch…	Positive	Positive	
14	5	I lost my B&N original cab…	Positive	Positive	
17	4	Works well, a little pricey …	Positive	Positive	
18	5	This is a great buy, compa…	Positive	Positive	
19	5	This mount is just what I …	Positive	Positive	
20	5	Great deal, easy to mount…	Positive	Positive	
21	4	This mount works really …	Positive	Positive	
23	5	for the price you just cant …	Positive	Positive	

Fig 21. Sorting as per Genuineness (Genuine Review)



	overall	reviewText	Overall_Sentiment	Review_Sentiment	sen
1	1	I'm a professional OTR tr…	Negative	Positive	
3	2	Not going to write a long r…	Negative	Positive	
22	4	This wall mount does eve…	Positive	Negative	
36	5	Didn't think it would wor…	Positive	Negative	
51	5	I live in a studio condo an…	Positive	Negative	
54	4	Just to know that the TV …	Positive	Negative	
66	4	Don't have anything to re…	Positive	Negative	
67	4	Bought the mount for my …	Positive	Negative	
68	2	The mount plate and the …	Negative	Positive	
82	1	I checked around Amazo…	Negative	Positive	
118	5	Another gem hidden amo…	Positive	Negative	
157	5	I have this mounted on m…	Positive	Negative	
167	1	I consider myself to be a g…	Negative	Positive	
181	5	This item didn't disappoi…	Positive	Negative	
200	5	Easy to install, with extra …	Positive	Negative	
201	2	The fact that the pivot poi…	Negative	Positive	
221	5	Do not listen to other revi…	Positive	Negative	

Fig 22. Sorting as per Genuineness (Not Genuine review)

The above 2 figures (*21 and 22*) show the sorting of the dataset as per genuine and not genuine review and display each data table distinctively.

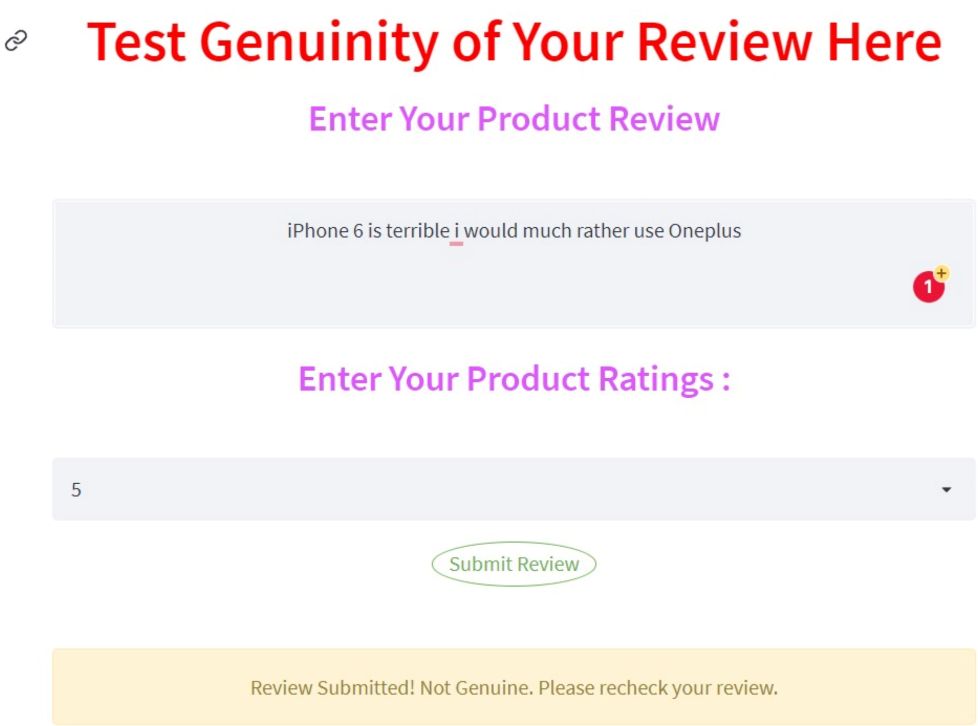


Fig 23. Custom Input of testing Genuinity

The panel in fig. 23 and 24 allows for custom input from user for testing and checking their genuineness of the review.

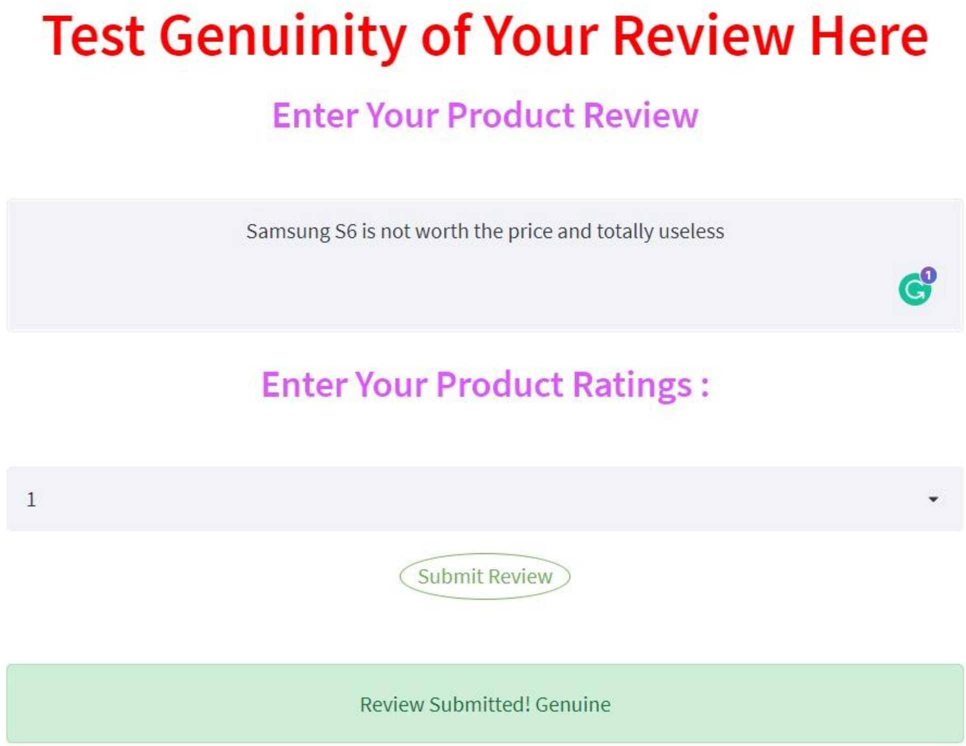


Fig 24. Output of the Custom Input

Chapter 5

Results

The comparative study among the three algorithms shows that logistic regression has the highest roc_auc_score as compared to the other 2 algorithms used. Making Logistic regression an ideal choice for binary classification.

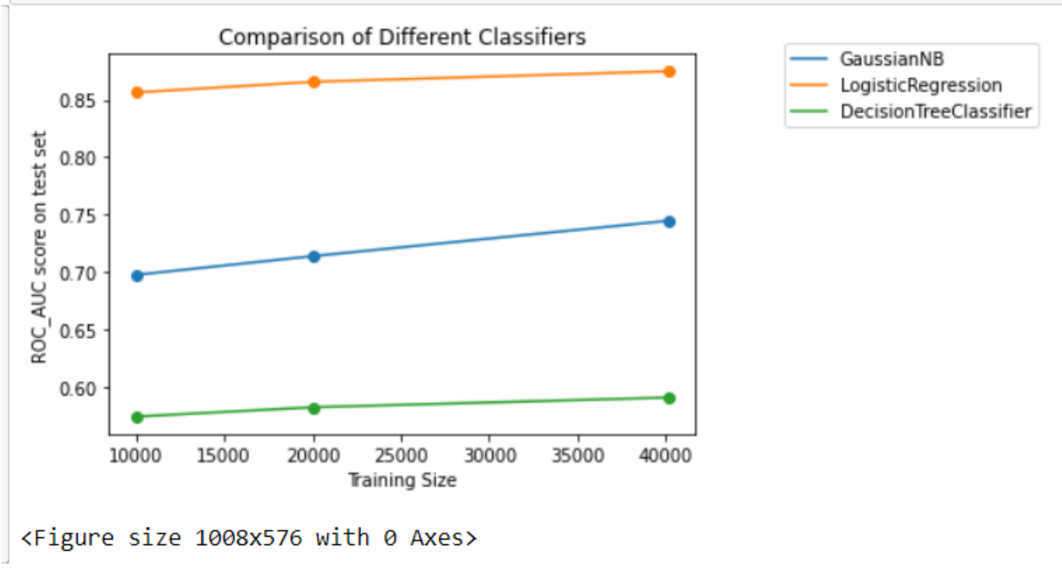


Fig 25. AUC_ROC score for all 3 algorithms

The Logsitic regression model classified only 1 reviews as not fake per 12 fake reviews and 1 review as fake per 11 not fake reviews. As shown in fig.26.

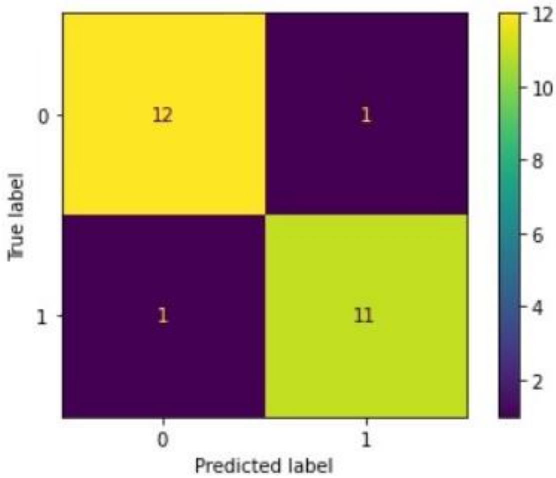


Fig 26. Confusion matrix for logistic regression

The Decision-Tree model classified only 2 reviews as not fake per 12 fake reviews and 1 review as fake per 10 not fake reviews. As shown in fig. 27.

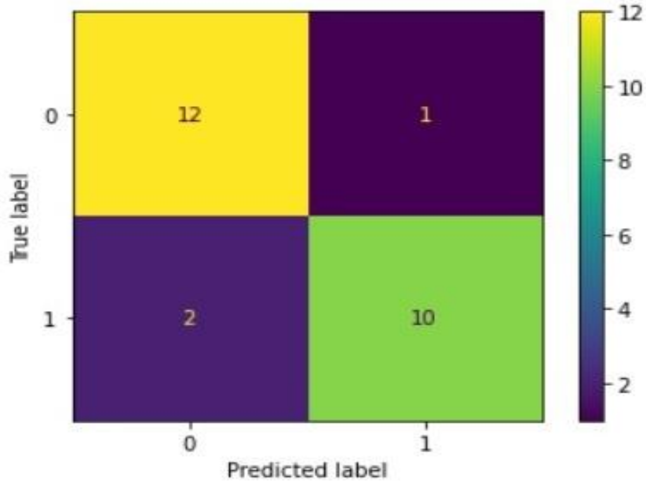


Fig 27. Confusion matrix for decision tree classifier

The Gaussian Naïve Bayes model classified only 3 reviews as not fake per 12 fake reviews and 1 review as fake per 9 not fake reviews. As shown in fig.28.

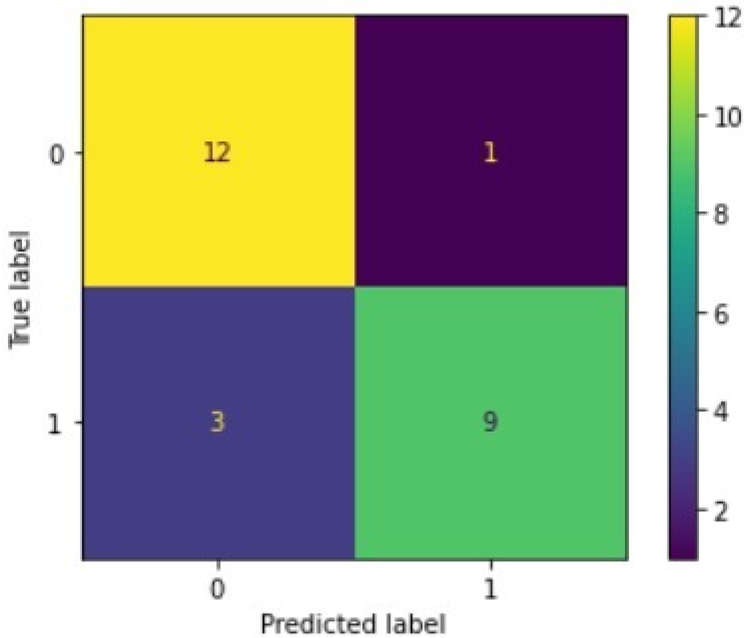


Fig 28. Confusion matrix for Guassian Naïve Bayes

```
Confusion Matrix

[[12  1]
 [ 1 11]]

Accuracy: 0.92

Micro Precision: 0.92
Micro Recall: 0.92
Micro F1-score: 0.92

Macro Precision: 0.92
Macro Recall: 0.92
Macro F1-score: 0.92

Weighted Precision: 0.92
Weighted Recall: 0.92
Weighted F1-score: 0.92

Classification Report

              precision    recall  f1-score   support

     Class 1       0.92      0.92      0.92        13
     Class 2       0.92      0.92      0.92        12

    accuracy                           0.92        25
   macro avg       0.92      0.92      0.92        25
weighted avg       0.92      0.92      0.92        25
```

Fig 29. Logistic regression results

In fig 29. the results about logistic regression are displayed, the accuracy of Logistic Regression is 92% which infers that the model have predicted and classified most of the data in the dataset correctly as 92% is almost closer to 100%. Thus, the success rate of Logistic Regression is the highest so far. The Macro-Precision and Recall for the model turn out to be 92%. This means the model can detect 92 correct samples (fake reviews) in a sample of 100 (reviews).

The weighted average of precision and recall is 0.92 which is 92% highest among all the other models used for this project

```
Confusion Matrix

[[12  1]
 [ 2 10]]

Accuracy: 0.88

Micro Precision: 0.88
Micro Recall: 0.88
Micro F1-score: 0.88

Macro Precision: 0.88
Macro Recall: 0.88
Macro F1-score: 0.88

Weighted Precision: 0.88
Weighted Recall: 0.88
Weighted F1-score: 0.88

Classification Report

              precision    recall  f1-score   support

     Class 1       0.86      0.92      0.89        13
     Class 2       0.91      0.83      0.87        12

    accuracy                           0.88        25
   macro avg       0.88      0.88      0.88        25
weighted avg       0.88      0.88      0.88        25
```

Fig 30. Decision tree classifier results

In fig 30. the results about Decision Tree Classifier are displayed, the accuracy of Decision-Tree Classifier is 88% which infers that the model have predicted and classified most of the data in the dataset correctly as 88% is almost closer to 90% but far from 100. The Macro-Precision and Recall for the model turn out to be 88%. This means the model can detect 88 correct samples (fake reviews) in a sample of 100 (reviews).

The weighted average of precision and recall is 0.88 which is 88%, second highest among all the other models used for this project.

```
Confusion Matrix

[[12  1]
 [ 3  9]]

Accuracy: 0.84

Micro Precision: 0.84
Micro Recall: 0.84
Micro F1-score: 0.84

Macro Precision: 0.85
Macro Recall: 0.84
Macro F1-score: 0.84

Weighted Precision: 0.85
Weighted Recall: 0.84
Weighted F1-score: 0.84

Classification Report

              precision    recall  f1-score   support

     Class 1       0.80      0.92      0.86        13
     Class 2       0.90      0.75      0.82        12

    accuracy                           0.84        25
   macro avg       0.85      0.84      0.84        25
weighted avg       0.85      0.84      0.84        25
```

Fig 31. Gaussian Naïve Bayes classifier results

In fig 31. the results about logistic regression are displayed, the accuracy of Logistic Regression is 92% which infers that the model have predicted and classified most of the data in the dataset correctly as 84% is almost closer to 90% but far from 100. The Macro-Precision and Recall for the model turn out to be 85% and 84%. This means the model can detect 84-85 correct samples (fake reviews) in a sample of 100 (reviews).

The weighted average of precision and recall is 0.85 which is 85%, least among all the other models used for this project.

Further comparative study of the 3 algorithms showed Logistic Regression was a better classifier compare to Guassian Naïve Bayes and Decision Tree Classifier. Since it had better accuracy and yielded better results compared to the other 2 algorithms used.

Models	Micro			Macro			Weighted		
	Precision	Recall	F1-Score	Preci-sion	Recall	F1-Score	Precision	Recall	F1-Score
Logistic Regression	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
Decision Tree	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
Gaussian Naïve Bayes	0.84	0.84	0.84	0.85	0.84	0.84	0.85	0.84	0.84

Fig. 32 Table for Comparison

In fig. 32 All the results and finding are summarized in one table.

The algorithm with highest micro-precision is logistic regression with 0.92 score while decision tree is the second highest and Naïve Bayes the 3^{rd} highest.

Also, Logistic Regression have outperformed the two algorithms with a macro-precision and recall of 0.92.

Thus, making logistic regression an ideal choice for binary classification of dataset.

Chapter 6

Conclusion

The project is successfully able to classify the reviews based on their sentiments and also able to evaluate the results comparatively by analyzing the difference between review rating and review text. The finding from Chapter 5 concludes that for a problem statement with Binary data classification the ideal algorithm for choice should be Logistic Regression, as it outperforms other algorithms that specialize in binary classification as shown in the table (fig. 32).

The drawback of this project is it is not trained to handle slang-based data input which may lead to inaccuracy in predicting the results. Also it only handles dataset files with an extension of .json. Once these drawbacks are resolved the Project will be a perfect system to detect fake reviews. And it will also help multiple MNC's and corporation to clean their database from fake and redundant feedbacks. This in turn will help them with qualitative data analysis for grasping the better understanding of the market and customer satisfaction.

References

[1] “Fake Product Review Monitoring and Removal for Genuine Online Reviews” Journal of Network Communications and Emerging Technologies (JNCET) Volume 8, Issue 4, April (2018).

[2] “Fake Product Review Monitoring Using Opinion Mining” International Journal of Pure and Applied Mathematics (IJPAM) Volume 119 No. 12 2018.

[3] “Fake Product Review Monitoring System” International Journal of Trend in Scientific Research and Development (IJTSRD) Volume: 3 | Issue: 3| Mar-Apr 2019.

[4] “A Framework for Fake Review Detection in Online Consumer Electronics Retailers” Intelligent Systems Group Department of Telematic Engineering Volume 165, 2019.

[5] “Implementation of fake product review monitoring system and real review generation by using data mining mechanism” Journal of Xi'an University of Architecture & Technology Volume XII, Issue II, 2020.

Acknowledgement

We would like to express our gratitude and thanks to Prof. Vaishali Suryawanshi for her valuable guidance and help. We are indebted for her guidance and constant supervision as well as for providing necessary information regarding the project. We would like to express our greatest appreciation to our principal Dr. G.T. Thampi and head of the department Dr. Tanuja Sarode for their encouragement and tremendous support. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of the project.

Musharraf Alam

Asawa Aryan

Yash Balchandani