ANALYZING THE FRAUDULENT PRODUCT REVIEWS USING MACHINE LEARNING

Project Report

Submitted in partial fulfilment of the requirements for the award of the degree of Master of Science in Information Technology

Project Work

Done By

B V PADMASHRI

(Reg. No. 21PITE15)

Guided by

Dr. J.A. Esther Rani , M.C.A., PGDCS., M.Phil., Ph.D

Department of Computer Science

Lady Doak College

(An Autonomous Institution affiliated to Madurai Kamaraj University)

Re-accredited with Grade 'A+' by NAAC (4th Cycle)

Madurai – 625002

(2022-2023)

DEPARTMENT OF COMPUTER SCIENCE

LADY DOAK COLLEGE

MADURAI

BONAFIDE CERTIFICATE

This is to certify that this is a bonafide record of the project work done by B V Padmashri (21PITE15). This is submitted in partial fulfilment for the award of the degree of Master of Science in Information Technology, Lady Doak College, Madurai. Submitted for the project evaluation and Viva Voce held on **25.04.2023** at Lady Doak College, Madurai.

Internal Examiner

Date:

External Examiner

Date:

Head of the Department

Date:

DECLARATION

I, B V Padmashri(Reg. No. 21PITE15), hereby declare that the project work entitled "**ANALYZING THE FRAUDULENT PRODUCT REVIEWS USING MACHINE LEARNING"** submitted to the Department of Computer Science, Lady Doak College, Madurai for the partial fulfilment of the requirements for the award of the degree of Master of Science in Information Technology is a record of my original work. I further declare that this record or any part of this work has not been submitted elsewhere for any other degree.

B V PADMASHRI

(21PITE15)

ACKNOWLEDGEMENT

At the outset, I would like to express my humble thanks to **the God Almighty,** for the kind grace showered on me to complete the project successfully.

I would also like to express my sincere thanks to our beloved Principal **Dr.Christianna Singh, M.A., M.Phil., P.G.D.C.A., Ph.D.** for the commendable support in the achievement of this project successfully.

I would also like to express my profound thanks to **Dr. N. Jayachandra M.Sc., M.Phil., Ph.D.,** Head & Associate, Department of Computer Science for her scholarly guidance, encouragement and valuable motivation throughout the period of my project.

I would also like to express my profound thanks to **Dr. T.R. Sivapriya M.C.A., M.Phil., Ph.D.,** Co-ordinator for B.Sc. Information Technology & Associate Professor, Department of Computer Science, for her scholarly guidance, encouragement and valuable motivation throughout the period of my project.

I also convey my heartfelt immense gratitude to my project guide **Dr.J.A.Esther Rani M.C.A.,PGDCS.,M.Phil.,Ph.D.,** Associate Professor, Department of Computer Science for her valuable guidance, encouragement and support given throughout the project.

I also convey my heartfelt thanks to project incharge **Mrs. E. Sheeba Sugantharani M.C.A.,M.Phil.,(Ph.D.,),** Associate Professor, Department of Computer Science for her valuable support and encouragement to complete the project.

I convey my heartfelt thanks to all the faculty members, Department of Computer Science for their support and encouragement to complete this project. I also convey my heartfelt thanks to all the non-Teaching staff members for rendering their support throughout this project.

I also thank all my Friends and Parents for their encouragement and full support throughout the project. Last but not least I am proud because I got an opportunity to thank all who made me think beyond the world and completed my project successfully.

B V PADMASHRI(21PITE15)

SYNOPSIS

A major information source now is the e-commerce platform. It considers consumer reviews—comments made by customers about goods and services they have purchased from an online store—when making decisions. Internet stores give customers the option to post evaluations of their purchases after they have made them, allowing potential buyers to read what other customers have to say before deciding whether or not to purchase a product or service from the store. In e-commerce, user reviews can have a significant impact on an organization's revenue.

The main objective of the project is to analyze or to classify the reviews into fake reviews and real reviews. To implement different machine learning algorithms. To enhance the overall performance for all classification algorithms.

CONTENTS

1. INTRODUCTION

1.1 ABSTRACT

Consumer reviews play an important role for consumers' online shopping activities. Customers will pay more attention to the products with positive reviews and avoid the negative ones. Nowadays the usage of Internet and online marketing has become very popular. Millions of products and services are available in online marketing that generate a huge amount of information. The content in the form of reviews, ratings, and comments can be analyzed. Sentiment analysis of reviews extract and aggregate fake and real opinions from product reviews. Product reviews are the data extracted from Amazon reviews. The various data mining techniques are used to detect the fraudulent review based on different features. The fraudulent reviews dataset is taken from kaggle. The main objective is to analyze and classify the product review into fraudulent reviews and real reviews by using the machine learning algorithms such as Xgboost and Logistic regression, and evaluating the performance using metrics such as accuracy, precision, recall and f1 measure.

Keywords:Fraudulent review,Xgboost ,Logistic regression

1.2 ABOUT THE PROJECT

The development of the modern world has accelerated the trend towards online purchases made through e-commerce websites. As the world becomes more digital, the availability of internet connectivity in both urban and rural locations is expanding. The majority of consumers purchase their daily necessities, such as goods or services, from online stores, thus before making a purchase, they read written reviews to learn about other customers' perceptions of the goods or services. A major information source now is the e-commerce platform. It considers consumer reviews—comments made by customers about goods and services they have purchased from an online store—when making decisions. Internet stores give customers the option to post evaluations of their purchases after they have made them, allowing potential buyers to read what other customers have to say before deciding whether or not to purchase a product or service from the store. In e-commerce, user reviews can have a significant impact on an organization's revenue. Before purchasing any product or service, online users rely on reviews. Such information may consist of opinions expressed either positively or negatively by customers who have already utilized the product. As a result, the credibility of online reviews is critical for businesses and can have a direct impact on their reputation and profitability. That is why some companies pay spammers to post fake reviews**.** Customer reviews and ratings can be used to get insightful information using sentiment analysis. The social media revolution has compelled the online community to use online reviews for posting feedback on goods, services, and other issues, as well as helping people to study customer feedback for making purchase decisions and businesses to raise the caliber of produced goods. The spread of fraudulent reviews has turned into a worrying problem since it deceives online shoppers as they make purchases and boosts or degrades the reputation of rival firms. Despite the fact that many actual customers post product reviews to share their opinions and shopping experiences with others, more fake reviews are published on e-commerce websites for monetary gain. Reviews on products and services, such as hotels and restaurants, are now crucial to consumers' online purchase decisions. Before making a buying decision, a lot of people check these reviews. A product may typically receive favorable or bad reviews. Consumers will pay greater attention to and steer clear of products

with positive evaluations. This influence will either result in increased business or significant financial losses. As more people make purchases online, fake consumer review detection has gained a lot of attention in recent years. Internet usage and internet marketing are increasingly common these days. Online marketing offers millions of goods and services, which creates a tonne of information. So, it can be challenging to discover the best services or goods that meet the need. Consumers directly base their decisions on reviews or opinions given by others based on their experiences. Anyone may write anything in this cutthroat society, which increases the prevalence of fake reviews.There is a need for a solution to identify these fraudulent reviews because this procedure provides incorrect information to prospective customers who want to purchase similar goods. Sentiment Analysis (SA) has emerged as one of the most intriguing topics in text analysis, owing to the potential commercial benefits. Sentiment analysis is textual contextual mining that identifies and extracts subjective information from source material, assisting businesses in understanding the social sentiment of their brand, product, or service while monitoring online conversations. However, most social media analysis is limited to basic sentiment analysis and count-based metrics.. One of the major challenges for SA is determining how to extract emotions from opinion reviews, as well as how to detect fake positive and fake negative reviews. Furthermore, user opinion reviews can be classified as positive or negative, which can be used by a consumer to choose a product. In this project ,to evaluate review text and determine if consumer reviews are favorable or bad, machine learning's branch of natural language processing is applied. The usage of Machine Learning Algorithms will help to vectorize the data and visualize the data. Then propose a method based on supervised learning for the identification of false reviews in e-commerce sites. The study uses machine learning classifiers to distinguish between fraudulent and real reviews. Experimental findings are assessed using several assessment criteria , and the performance of the suggested system is contrasted with earlier efforts. The main objective of the project is, to analyze or to classify the reviews into fake reviews and real reviews.To implement the different machine learning algorithm.To enhance the overall performance for all classification algorithms.

2.BACKGROUND

2.1 LITERATURE REVIEW

An Empirical Study on Detecting Fake Reviews Using Machine Learning Techniques ,2017

The study of this paper states that the reputation of the e-commerce site plays a significant role in allowing various parties to achieve mutual benefits by establishing relationships. The reputation systems are intended to assist consumers in deciding whether or not to negotiate with a specific party. Many factors have a negative impact on customers' and vendors' perceptions of the reputation system. For example, users may create phantom feedback from fake reviews to support their reputation due to a lack of honesty or effort in providing feedback reviews. Furthermore, user opinions can be classified as positive or negative, which can be used by a consumer to choose a product.Moreover, phony good and fake negative reviews were both caught using detection processes, as evidenced by the findings.This study used both the without-stopwords and with-stopwords approaches to apply the sentiment classification algorithms. They discovered that the stopwords strategy is more effective in identifying fake reviews as well as text categorization. Finally, they would like to expand on this work in future work by using additional datasets, such as the Amazon dataset, the eBay dataset, or a separate dataset of a movie review, and by using various feature selection methodologies.

Application:

- They applied both with and without stop words approach and identification of better results with supervised learning algorithms

Drawback:

- One of the Disadvantages are usage of less data of movie reviews in research

Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques, 2017

The study of this paper states that previous publications had one of the major challenges for Sentimental analysis in determining how to extract emotions from opinions, as well as how to detect fake positive and fake negative reviews from opinion reviews. Furthermore, user opinion reviews can be classified as positive or negative, which can be used by a consumer to choose a product.they proposed several methods to analyze a dataset of movie reviews , also presented sentiment classification algorithms to apply a supervised learning of the movie reviews located in two different datasets v2.0 movie dataset and v1.0 movie dataset.The first dataset is known as the movie reviews dataset V2.0, and it contains 2000 movie reviews, 1000 of which are positive and 1000 of which are negative. The second dataset is known as the movie reviews dataset V1.0, and it contains 1400 movie reviews, 700 of which are positive and 700 of which are negative. The experimental approaches studied the accuracy of all sentiment classification algorithms, and how to determine which algorithm is more accurate. They discovered that the SVM algorithm is the most accurate at correctly classifying the reviews. Also, the best strategy employed in this study affects the detecting procedures for fake positive reviews and fake negative reviews . Future research will expand on this study by using more datasets, such as the Amazon or eBay datasets, and different feature selection techniques. Additionally, usage of tools like Python, R, or R studio to apply sentiment classification algorithms to identify fraudulent reviews.

Application:

- ➔ They are successful in achieving the drawback of previous publication
- ➔ Usage of two different datasets for research

Drawback:

- ➔ The disadvantages of this paper are even though usage of two different datasets, the data are less.
- ➔ Accuracy is low compared to previous publication

Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques , July 2018

The study of this paper states that usage of many strategies to examine a dataset of movie reviews that were proposed in this study. Also, they provided sentiment classification methods for supervised learning of movie reviews from three independent datasets.Five main performance evaluation measures have been introduced for Classification algorithms. These include Fake Positive Reviews predictive value, Fake Negative Reviews predictive value, Real Positive Reviews predictive value, Real Negative Reviews predictive value, and evaluating metrics. Also, they were able to use detection techniques to find fake negative and good reviews. The first dataset is known as the movie reviews dataset V1.0, and it contains 1400 movie reviews, 700 of which are positive and 700 of which are negative. The second dataset is known as the movie reviews dataset V2.0, and it contains a total of 2000 movie reviews, 1000 positive and 1000 negative.The third dataset is known as the movie reviews dataset V3.0, and it contains 10662 movie reviews, 5331 of which are positive and 5331 of which are negative. They discovered that the SVM method is the most accurate for correctly classifying the reviews in movie reviews datasets, i.e., V1.0, V2.0, and V3.0, using the accuracy analysis . The best strategy employed in this study will also influence the detecting procedures for fake negative and positive evaluations. Futureresearch will expand on this study by using more datasets, such as the Amazon or eBay datasets, and different feature selection techniques. Additionally, using tools like Python or R studio, they might apply sentiment classification algorithms with stopwords removal and stemming techniques to identify false reviews.

Application:

- ➔ Usage of more data compared to previous publications
- ➔ Usage of three different dataset for research purposes

Drawback:

- ➔ Accuracy has become low compared to both previous publications on this study

Fake Reviews Identification Based on DeepComputational Linguistic Features, Jan 2020

The study of this paper states that the e-commerce platform has evolved into an important information resource. It considers consumer feedback about products and services purchased from the online website, which are referred to as reviews. Consumers can write on online websites. After-purchase product or service reviews, so that when new customers decide to purchase products or services from an online website, they read the recommendations or reviews written by people who have used the product or service. Those reviews, on the other hand, could be trusted (real) or spam (fake). E-commerce website fraudsters who intentionally mislead potential customers and defame businesses can write fake reviews. The primary goal of this paper is to analyze, identify, and detect fake reviews of electronic products in datasets relating to various cities in the United States. They investigated several feature extraction techniques in this paper, including LIWC, sentiment analysis, POS, and subjectivity. They extracted a set of features from the review text using these methods, including authenticity, analytic thinking, polarity, objective, subjective, and counts of adjectives, verbs, nouns, and adverbs. Use IG (Information Gain) to select discriminative and highest features for feature selection.Three different supervised machine-learning techniques are used to classify the reviews as fake or trusted: Decision tree, Random forest, and Adaptive boosting. They discovered that the adaptive boosting algorithm outperforms all other algorithms considered for study.In their research, they addressed the issue of fake reviews in the customer electronics domain using deep linguistics features related to centric reviews. Future research will attempt to take into account review and reviewer-centric features in order to detect fake reviews in online e-commerce websites.

Application:

- ➔ Usage of various feature selection techniques

Drawback:

- ➔ No usage of large scale labeled dataset for training the classifier

Data Analytics for the Identification of Fake Reviews Using Supervised Learning, 2022

The study of this paper states that fake reviews, also known as deceptive opinions, are used to mislead people and have recently gained prominence. This is due to the rapid increase in online marketing transactions such as selling and purchasing. Before making a purchase decision, new customers typically read the posted reviews or comments on the website. However, the current challenge is how new people can tell the difference between genuine and fake reviews, which later deceive customers, cause losses, and tarnish companies' reputations.Fake review detection is a subfield of natural language processing. Its goal is to analyze, detect, and classify product reviews on online e-commerce domains as fake or true. Due to the significant impact on customers and e-commerce businesses, many researchers have conducted studies on fake/spam review identification. Their paper aims to create an intelligent system that can detect fake reviews on ecommerce platforms by analyzing the review text's n-grams and the sentiment scores provided by the reviewer. This study's proposed methodology used a standard fake hotel review dataset for experimenting and data preprocessing methods, as well as a term frequency-Inverse document frequency (TF-IDF) approach for extracting features and representing them. For detection and classification, n-grams from review texts were fed into the built models to determine whether they were false or true. Based on testing accuracy and the F1-score, the classification results of these experiments revealed that naive Bayes (NB), support vector machine (SVM), adaptive boosting (AB), and random forest (RF) were the best. In their study, the random Forest outperforms the other algorithms.The drawback was that they found that experiments' dataset was restricted to the hotel domain. For future research, a large-scale dataset should be combined with several textual and behavioral features for detecting fake reviews on various e-commerce domains.

Application:

- ➔ Usage of various feature selection techniques

Drawback:

- ➔ Less features were used to train the proposed models

Fake Reviews Detection using Supervised Machine Learning ,2021

The study of this paper states that with the ongoing evolution of E-commerce systems, online reviews are increasingly regarded as a critical factor in establishing and maintaining a positive reputation. Identifying fake reviews is thus an active and ongoing research topic. Detecting fake reviews is dependent not only on the key features of the reviews but also on the reviewers' behaviors.The paper proposes a machine learning approach for detecting fake reviews. In addition to the review features extraction process, the paper employs several feature engineering techniques to extract various reviewer behaviors. The paper compares the results of several experiments performed on a real Yelp dataset of restaurant reviews with and without features extracted from user behaviors.This dataset contains 5853 reviews of 201 Chicago hotels written by 38, 063 reviewers. The reviews are divided into four categories: real (4, 709 reviews) and fake (1, 144 reviews). Yelp categorizes reviews as genuine or fake.Each review instance in the dataset includes the review date, review ID, reviewer ID, product ID, review label, and star rating.Researchers compared the performance of several classifiers in both cases: KNN, Naive Bayes (NB), SVM, Logistic Regression, and Random Forest. During the evaluations, different n-gram language models, specifically bi-gram and tri-gram, are also considered.The results show that the KNN(with K=7) classifier outperforms the other classifiers in the detection of fake reviews. Future work may consider including other behavioral features, such as features that depend on how frequently reviewers do the reviews, how long it takes reviewers to complete reviews, and how frequently they submit positive or negative reviews.

Application:

➔ Usage of different n-gram language models

Drawback:

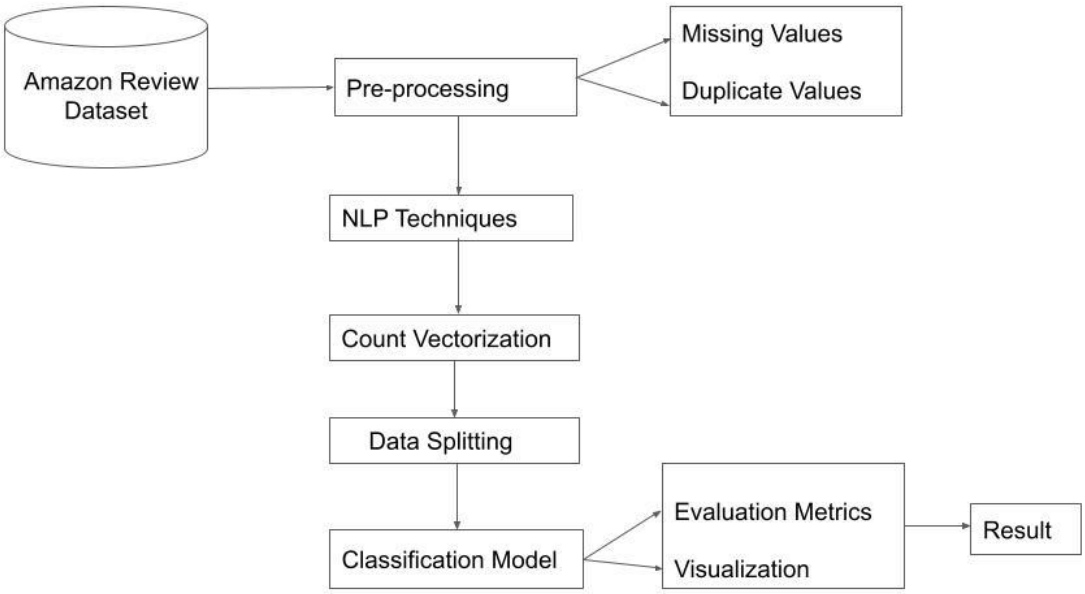
➔ In the current work, not all reviewers' behavioral characteristics have been considered

2.2 PROBLEM STATEMENT

Based on literature survey's future enhancements and drawback ,

- This project aims to work on amazon product review dataset which is large –scale dataset
- Using different machine learning algorithms

2.3 WORKFLOW DIAGRAM



3.SYSTEM ENVIRONMENT

3.1 HARDWARE CONFIGURATION

System : Pentium IV 2.4 GHz

Hard Disk : 200 GB

Mouse : Logitech.

Keyboard : 110 keys enhanced

RAM : 4 GB

3.2 SOFTWARE CONFIGURATION

O/S : Windows 10

Language : Python Programming

Front End : Google Colab

4.SYSTEM SPECIFICATION

4.1 SOFTWARE SPECIFICATION

Python is one of the few programming languages that can be both simple and powerful. Users can be pleasantly surprised to discover how simple it is to focus on the solution to a problem rather than the syntax and structure of the programming language. Python's official introduction is that it is a simple yet powerful programming language. It has high-level data structures that are efficient and a straightforward but effective approach to object-oriented programming. Python's elegant syntax, dynamic typing, and interpreted nature make it an ideal language for scripting and rapid application development across many platforms.

Features of python:

- Easy to Learn-Python code resembles everyday English terms. The code block is defined by the indentations rather than by semicolons or brackets
- Free and Open Source-Python is created under an open source license that has been accepted by OSI. Because of this, using it is totally free—even for profit. Python can be downloaded and used in applications without costing anything
- Object Oriented-When design is centered on data and objects rather than on functions and logic, a programming language is said to be object-oriented. Contrarily, a programming language is procedure-oriented if it emphasizes functions more (code that can be reused). Support for both object-oriented and procedure-oriented programming is a key Python feature
- Extensible-If a programming language can be expanded to other languages, it is said to be extensible. Python is a very extensible language because code may also be written in other languages, such as C++

4.2 PROJECT SPECIFICATION

The Python Standard Library contains all of Python's syntax, semantics, and tokens. It includes built-in modules that provide the user with access to basic functions such as I/O and a few other essential modules. For the most part, the Python libraries have been written in C. The Python standard library contains over 200 core modules. Python is an extremely powerful programming language as a result of all of these factors. The Python Standard Library is essential. Python includes a number of libraries that help programmers.

The following python Library are used in this project ,

Pandas-Pandas is a free library that is distributed under the Berkeley Software Distribution license (BSD). This well-known library is frequently used in the field of data science. Among other things, they're primarily utilized for data analysis, modification, and cleansing.It merge and connect data sets, data sets can be easily reshaped and pivoted and allows for time-series functionality .The ability to split, apply, and combine operations on data sets with the group by functionality.

Matplotlib-This library is responsible for the charting of numerical data. It is employed in data analysis for this reason. It is an open-source library that, among other things, plots high-definition graphs, scatterplots, box plots, and pie charts.

NLTK-NLTK is a standard python library that provides a set of diverse algorithms for Natural language processing .The Natural Language Toolkit (NLTK) is a Python platform for developing programs that work with human language data and can be used in statistical natural language processing (NLP). It is one of the most used libraries for NLP and Computational Linguistics.

For text tokenization, stemming, and stop word removal the NLP is used.

CountVectorizer- A group of text documents are converted into a vector of term/token counts using the CountVectorizer programme. Its capabilities turn it into a very versatile feature representation module for text.

5.RESEARCH METHODOLOGY

5.1 DATA COLLECTION

Machine learning begins with data. However, in order for this data to be useful, several processes must be carried out. One of them is data collection. The process of gathering data relevant to the goals and objectives of the machine learning project is known as data collection. which is to eventually obtain a dataset, which is essentially your collection of data, all ready to be trained and fed into an ML model. The outcomes and performance of an ML model are directly impacted by the data collection process. The role of data collection in machine learning leads to setting the project's objective ,problem statement for ML, processing of data Development of models,model execution,tracking a machine learning model's performance. In this project ,The data is collected from kaggle. Fake review dataset is used in this research. Data selection is the process of identifying or categorizing product evaluations into genuine and fraudulent reviews. The collection includes details about the product, such as reviews, product categories, comments, and product identifiers.

The dataset consists of following feature,

Date, URL, Review_Title, Author, Rating, Review_text, Review_helpful, Sentiment, Subjectivity,Neg_Count,Word_Count,Unique_words,Noun_Count,Adj_Count,Verb_Count,Adv_Count,Pro_Count,Pre_Count,Con_Count,Art_Count,Nega_Count, Aux_Count, Authenticity, AT, Rev_Type

5.2 DATA PREPROCESSING

Preparing raw data to be used by a machine learning model is a process known as data pre-processing or data preparation. The first and most important stage in developing a machine learning model is this one. Real-world data frequently includes noise, missing values, and may be in an undesirable format, making it impossible for machine learning models to be utilized directly on it. Cleaning up the data and preparing it for a machine learning model are necessary steps that also improve the model's accuracy and effectiveness.the pre-processing have following steps, getting or collection of the dataset, importing dataset and libraries,identification of missing values and duplicate values,encoding categorical data ,data splitting and feature scaling.To make data pre-processing successful following thing needs to assessed, data quality assessment ,examining data carefully to determine its overall quality, relevance to your project, and consistency. In almost any data set, there are a number of data anomalies and inherent problems to be aware of. For example, Mismatched data types, Because data is collected from a variety of sources, it may arrive in a variety of formats. While the end goal of this entire process is to reformat your data for machines and replace it with similarly formatted data. Second is data cleaning is the process of adding missing data, filling in any gaps, and deleting inaccurate or unnecessary information from a data set. In order to ensure that data is ready for use for downstream needs, dating cleansing is the most crucial preprocessing step. All of the inconsistent data that data quality check revealed will be fixed via data cleaning. There are a variety of cleaners that you may need to use depending on the type of data you're dealing with. Next is the process of transforming the data into the appropriate forms required for analysis and other downstream operations will start with data transformation.This typically occurs in one or more of the following processes: concept hierarchy development, feature selection, discreditation, normalization, and aggregation.The last one is data reduction, which makes analysis more challenging even after cleaning and transformation. Most of everyday human speech is unnecessary or irrelevant to the needs of the researcher, especially when using text analysis. Data reduction not only simplifies and improves the accuracy of the analysis, but also uses less storage space.

Data pre-processing mainly involved in this project is identification of missing data and duplicate value.Missing data are data loss or corruption that may be the root cause of missing values. Many machine learning algorithms do not allow missing values, the management of missing data is crucial during the dataset's preprocessing.There are many ways to handle missing data, first is delete rows with missing values: rows or columns with null values can be deleted to address missing values. Columns can be completely dropped if more than half of their rows are null. Rows with one or more columns with null values can also be removed.Another way is the mean, median, or mode of the column's remaining values can be used to replace missing values in columns in the dataset that have continuous numeric values. Comparatively speaking to the former procedure, this one can avoid data loss. A statistical method for handling the missing numbers involves replacing the two estimates mean, median.Data is frequently derived from multiple sources, and there is a good chance that a given table or database contains entries that do not belong there. Filtering out of date entries may be necessary in some cases. In others, a more complex data filtering is required. Next pre-processing step involved is to drop duplicate values in the dataset ,duplicate entries are problematic for a variety of reasons. During training, an entry that appears more than once receives disproportionate weight. Models that succeed on frequent entries only appear to perform well. Where identical entries are not all in the same set, duplicate entries can sabotage the split between train, validation, and test sets. This can result in biased performance estimates, which can lead to the model underperforming in production.Duplicate entries in databases can be caused by a variety of factors, including processing steps that were rerun anywhere in the data pipeline. While the presence of duplicates has a significant negative impact on the learning process, it is relatively simple to resolve. One option is to make columns unique whenever possible. Another option is to run a script that detects and deletes duplicate entries automatically.

5.3 PROPOSED METHOD

According to the problem statement, the proposed method for this project will be the Amazon fake reviews dataset taken as input from Kaggle. The first step is to implement the data pre-processing step. This step handles missing values to avoid incorrect prediction and drops duplicate values. The next step is to implement the NLP techniques, which involves cleaning up the texts by removing stop words, punctuation, and stem words. Following that, data splitting implementation is completed, with the data being split into train and test. The train data is used to evaluate the model, while the test data is used to predict the model. The next step is to implement a different machine learning algorithm, such as XGBoost or Logistic Regression.Finally, the experimental results show performance metrics such as accuracy, precision, recall, f1-score, and a comparison graph between the algorithms mentioned above. The results then display a graph in the form of a bar and pie chart to classify the product reviews as genuine or fake.

Novelty

- Consideration of large -scale dataset and effectively classifying the product reviews
- Implementation of algorithms ,which is not in existing in base paper considered for this project

The following project modules will briefly explain proposed system of the work,

- Data collection
- Data preprocessing
- NLP techniques
- Vectorization
- Data splitting
- Classification
- Result Generation

DATA COLLECTION

- The data is collected from kaggle, Fake review dataset is used in this research
- The objective of work is the process of identifying or categorizing product evaluations into genuine and fraudulent reviews
- The collection includes details about the product, such as reviews, product categories, comments, and product identifiers.the dataset consists of following features:

Date, URL, Review_Title, Author, Rating, Review_text, Review_helpful, Sentiment, Subjectivity,Neg_Count,Word_Count,Unique_words,Noun_Count,Adj_Count,Verb_Count,Adv_Count,Pro_Count,Pre_Count,Con_Count,Art_Count,Nega_Count, Aux_Count, Authenticity, AT, Rev_Type

DATA PRE-PROCESSING

- The process of removing unwanted data from a dataset is known as data pre-processing
- The dataset is transformed using pre-processing data transformation techniques into a structure appropriate for machine learning
- This step also includes cleaning the dataset by removing irrelevant or corrupted data that may affect the dataset's accuracy, making it more efficient
- Missing data removal: In this process, null values such as missing values and Nan values are replaced with 0. Missing and duplicate values were removed, and the data was cleaned of any irregularities
- That the majority of machine learning algorithms necessitate numerical input and output variables

In this project ,checking of missing values and dropping duplicates values is done

NLP TECHNIQUES

In the dataset the review text involves reviewer's opinion about the product ,which cannot be directly fed to machine learning algorithms to process to classify the fraudulent reviews and real reviews.so the review text is preprocessed and cleaned using NLTK libraries (NLP).

- NLP is a branch of machine learning that focuses on a computer's ability to understand, analyze, manipulate, and potentially generate human language
- Cleaning (or pre-processing) data usually consists of several steps:
- Remove punctuation: Punctuation can add grammatical context to a sentence, which aids our comprehension
- However, vectorizers that count the number of words rather than the context do not add value, so all special characters are removed. For example: How are you? ->How are you
- Text is divided into units, such as sentences or words, using tokenization. It provides previously unstructured text structure. such as Product is good-> "Product," "is," "good"
- Removal of stopwords: These are frequent words that almost always appear in texts. It doesn't have any crucial data-related terminology in it. like is "product is good" becomes "product" "good", the stopword is removed
- Removal of stem words:Stemming is the process of reducing a word to its stem form. It's common sense to treat related words similarly. It removes suffixes such as "ing", "ly", "s", and so on using a simple rule-based approach

COUNT VECTORIZATION

- Text must be tokenized, or parsed to eliminate specific words, in order to use textual data for predictive modeling
- Next, in order to be used as inputs in machine learning algorithms, these words must be encoded as integers or floating-point numbers
- This method is referred to as feature extraction (or vectorization)

- A collection of text documents is transformed into a vector of term/token counts using Scikit-CountVectorizer. It makes it possible to pre-process text data before creating the vector representation
- It is a very flexible feature representation module for text because of this functionality

DATA SPLITTING

- Data is necessary during the machine learning process in order for learning to occur.
- In this process, consider the 70% of the dataset to be the training data and the remaining 30% to be the testing data
- Data splitting is the act of partitioning available data into two portions, typically for cross-validation purposes
- Data is split into training and testing sets, with the training set being used to construct a predictive model and the testing set being used to assess the model's performance.
- The majority of the data is often utilized for training, and a smaller piece is used for testing when the dataset is divided into a training set and a testing set

CLASSIFICATION

- In this process,to implement the machine learning algorithms of supervised learning such as,
- XGBoost algorithm and Logistic regression for classifying the fraudulent and genuine reviews

RESULT GENERATION

- The Final Result will get generated based on the overall classification and prediction
- The performance of this proposed approach is evaluated using some measures like, accuracy ,precision, recall and f1-score

6. TESTING AND IMPLEMENTATION

6.1 EXPERIMENT AND TESTING

After data splitting into training data and test data at 70:30 ratio. The classification algorithm is implemented such as XGBoost and Logistic Regression to detect the fraudulent review and genuine review

XGBoost Algorithm:

A well-known machine learning toolkit called XGBoost (Extreme Gradient Boosting) applies the gradient boosting technique to supervised learning issues. XGBoost is a well-liked option for a variety of machine learning tasks, including classification, regression, and ranking because of its accuracy, speed, and scalability. It operates by adding decision trees to the ensemble iteratively. At each iteration, the method tries to use a new decision tree to fit the residual errors from the previous iteration. The total of all the individual decision trees' forecasts makes up the final forecast.

Logistic Regression:

Logistic regression is a statistical method for binary classification problems in which the response variable has only two possible values: 0 or 1. It is a type of generalized linear model that models the relationship between one or more predictor variables and the response variable. The goal of logistic regression is to find the best-fit parameters that can predict the likelihood of a binary outcome based on the predictor variables' values. The logistic regression output is a probability score ranging from 0 to 1, which can be converted to a binary outcome using a decision threshold.

6.2 EVALUATION AND VALIDATION RESULTS

ACCURACY:

The ability of the classifier is referred to as accuracy. It accurately predicts the class label, and predictor accuracy measures how effectively a specific predictor can predict the value of an attribute for new data.

ACCURACY = (TP+TN)/ (TP+TN+FP+FN)

METRICS:

Data science places a lot of importance on model evaluation. It makes it simple to present the model to others and aids in understanding how well it performs.

There are 3 main metrics for model evaluation in classification:

- Precision
- Recall
- f1-score

Precision:

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Precision=TP/ (TP+FP)

Recall:

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity..

Recall=TP/ (TP+FN)

F1-score:

It is the harmonic mean of precision and recall, and its value ranges from 0 to 1, with higher values indicating better performance.

F1 = 2 * (precision * recall) / (precision + recall)

6.3 RESULTS AND DISCUSSION

In this proposed system, the comparison of two classification algorithms such as Logistic regression and Xgboost to detect the fraudulent reviews and genuine reviews from fake review detection dataset from kaggle,Logistic regression outperforms the Xgboost in this problem with accuracy metric 94.7% whereas the Xgboost's accuracy metric 93% along with performance metrics.

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
LOGISTIC REGRESSION	**94.70 %**	**93.95 %**	**94.28%**	**94.12%**
XGBOOST	**93.19%**	**90.61%**	**94.67 %**	**92.59%**

7.CONCLUSION

Hence, a method based on machine learning is presented for the detection of actual and fraudulent reviews on the fake reviews dataset. The approach used in the research in this paper, based on the Xgboost and logistic regression, was selected primarily for its simplicity and well-established performance capabilities. The experimental results show that the suggested approach performed better than machine learning algorithms in terms of accuracy, precision, recall, and F1-score.

8.FUTURE ENHANCEMENTS

In future work of this project, it is possible to provide extensions or modifications to the proposed optimization and classification algorithms using hybrid models and also to achieve further increased performance. Also to explore the application of more advanced deep learning methods like Long short term memory model and Gated Recurrent Units model which are used for text processing and also pre-trained models can be used for model's best performance and with possible combinations of machine learning.

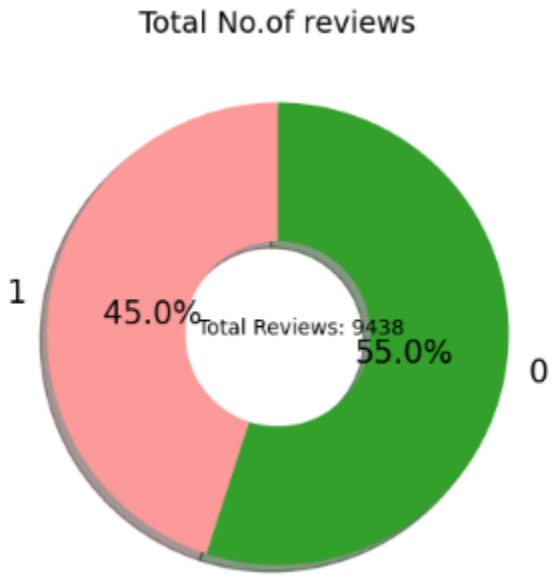
9.BIBLIOGRAPHY

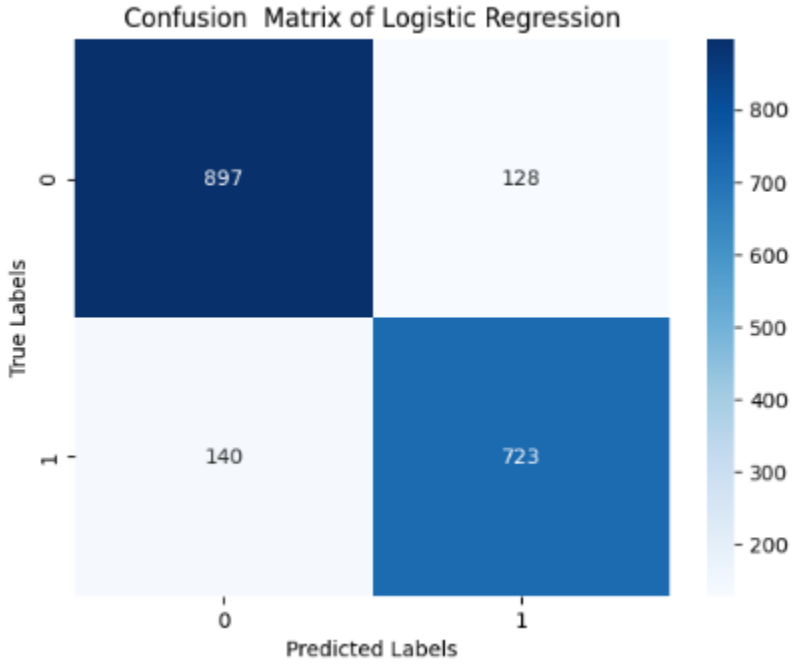
1. Elshrif Elmurngi; Abdelouahed Gherbi,"An empirical study on detecting fake reviews using machine learning techniques",2017 Seventh International Conference on Innovative Computing Technology (INTECH)-16-18 August 2017.
2. Elshrif Elmurngi; Abdelouahed Gherbi,"Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Technique ",DATA ANALYTICS 2017 : The Sixth International Conference on Data Analytics.
3. Elshrif Elmurngi; Abdelouahed Gherbi,"Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques",International Journal on Advances in Systems and Measurements, vol 11 no 1 & 2, year 2018.
4. Saleh Nagi Alsubari , Mahesh B. Shelke , Sachin N. Deshmukh," Fake Reviews Identification Based on DeepComputational Linguistic Features",International Journal of Advanced Science and Technology Vol. 29, No. 8s, (2020), pp. 3846-3856 .
5. Saleh Nagi Alsubari , Sachin N. Deshmukh , Ahmed Abdullah Alqarni , Nizar Alsharif Theyazn H. H. Al Dhyani, Fawaz Waselallah Alsaade and Osamah I. Khalaf," Data Analytics for the Identification of Fake Reviews Using Supervised Learning",Computers, Materials & Continua Tech Science Press DOI:10.32604/cmc.2022.019625.
6. Ahmed M. Elmogy , Usman Tariq , Atef Ibrahim,Ammar Mohammed," Fake Reviews Detection using Supervised Machine Learning",(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 1, 2021.
7. Sun, Huan, Alex Morales, and Xifeng Yan., "Synthetic review spamming and defense", Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 13, 2013.
8. Sharma, Kuldeep, and King-Ip Lin., "Review spam detector with rating consistency check", Proceedings of the 51st ACM Southeast Conference on - ACMSE 13, 2013.
9. Xu, Chang, "Detecting collusive spammers in online review communities", Proceedings of the sixth workshop on Ph D students in information and knowledge management - PIKM 13, 2013.

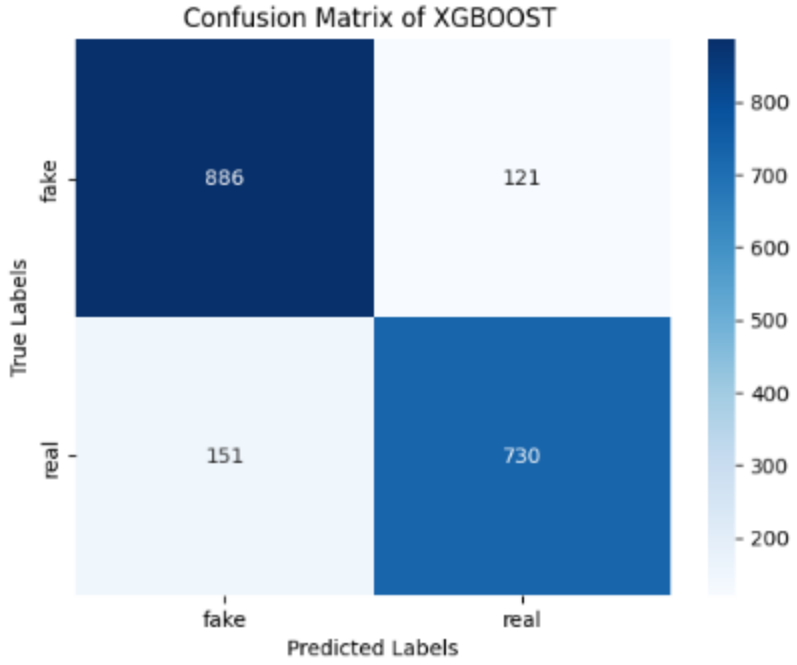
10. Lau, Raymond Y. K., S. Y. Liao, Ron Chi-Wai Kwok,Kaiquan Xu, Yunqing Xia, and Yuefeng Li. "Text mining and probabilistic language modeling for online review spam detection", ACM Transactions on Management Information Systems, 2011.

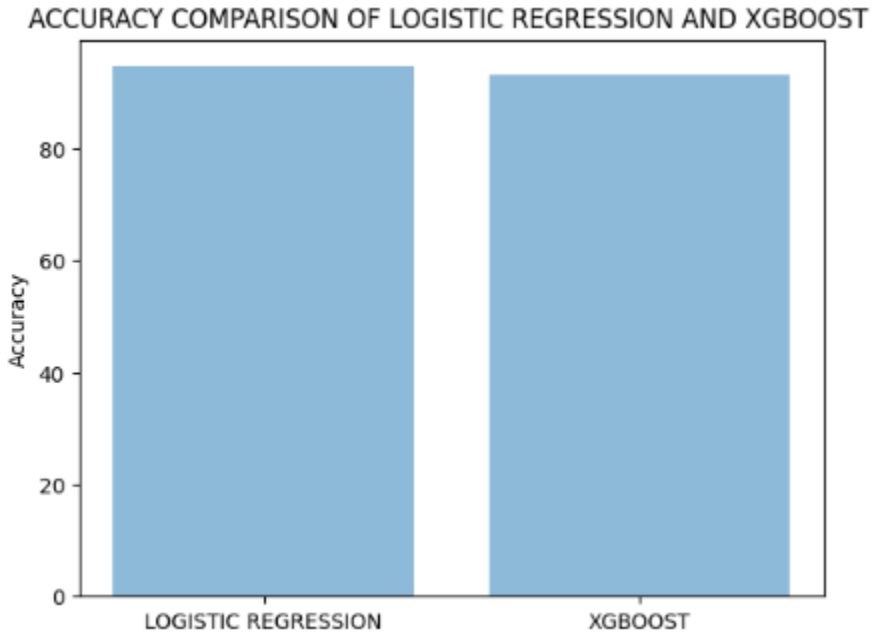
10.APPENDIX

10.1 SAMPLE REPORTS

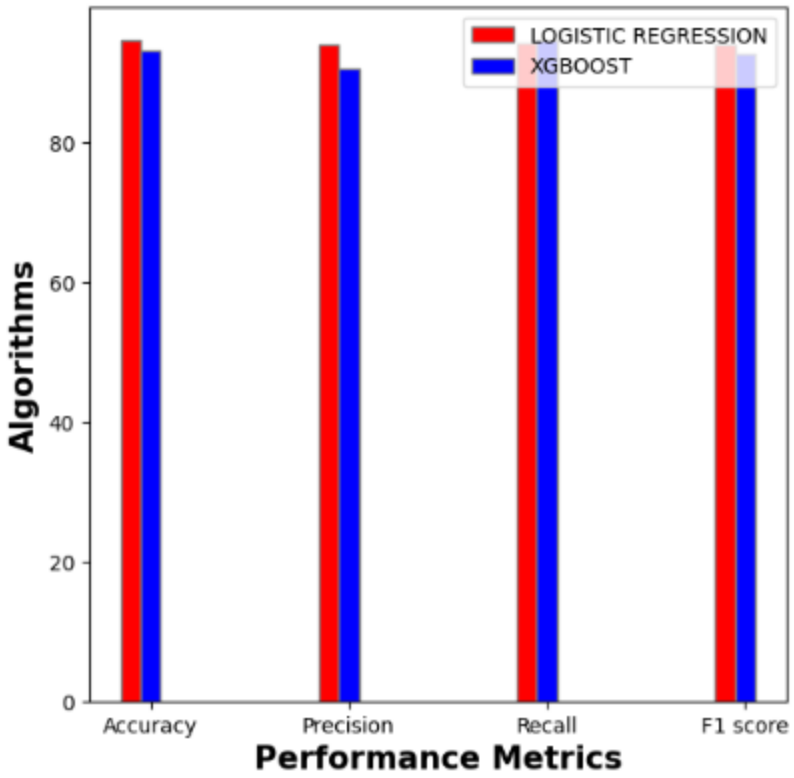





Confusion Matrix of XGBOOST
886
121
151
730
fake
real
True Labels
fake
real
Predicted Labels
800
700
600
500
400
300
200


ACCURACY COMPARISON OF LOGISTIC REGRESSION AND XGBOOST
Accuracy
0
20
40
60
80
LOGISTIC REGRESSION
XGBOOST

```
------------------- Performance Comparison --------------------
```



10.2 SAMPLE CODE

IMPORTING DATASET

```
import pandas as pd
data=pd.read_csv('/content/drive/MyDrive/fake product review.csv')
data
```

PRE-PROCESSING

```
#================== PREPROCESSING ============================

#=== checking missing values ===

print("---------------------------------------")
print("========== Checking missing values  ======")
print("---------------------------------------")
print(data.isnull().sum())
print()

data.drop_duplicates(inplace = True)
```

TEXT CLEANING

```
#=== TEXT CLEANING ====
import re

cleanup_re = re.compile('[^a-z]+')
def cleanup(sentence):
    sentence = str(sentence)
    sentence = sentence.lower()
    return sentence
```

```
print("-------------------------------------------")
print("============ Before Applying NLP  ============")
print("-------------------------------------------")
print()
print(data['Review_text'].head(10))

print("-------------------------------------------")
print("============ After Applying NLP  ============")
print("-------------------------------------------")
print()

data["Summary_Clean"] = data["Review_text"].apply(cleanup)
data["URL"] = data["URL"].apply(cleanup)

print(data["Summary_Clean"].head(10))
```

NLP TECHNIQUES

```
#==== stop words ====

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

import re
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
import string
stop_words = stopwords.words('english')
stemmer    = nltk.SnowballStemmer("english")
stop_words = stopwords.words('english')


def clean_data(text1):
```

```
    text1 = str(text1).lower()
    text1 = re.sub('\[.*?\]', '', text1)
    text1 = re.sub('https?://\S+|www\.\S+', '', text1) # remove urls
    text1 = re.sub('<.*?>+', '', text1)
    text1 = re.sub('[%s]' % re.escape(string.punctuation), '', text1) # remove punctuation
    text1 = re.sub('\n', '', text1)
    text1 = re.sub('\w*\d\w*', '', text1)
    return text1
def preprocess_data(text):
    text = clean_data(text)                                    # Clean puntuation, urls, and so on
      text = ' '.join(word for word in text.split() if word not in stop_words)     # Remove
stopwords
     text = ' '.join(stemmer.stem(word) for word in text.split())                # Stemm all the
words in the sentence
    return text

print("=====================================================")
print("              Before Applying NLP              ")
print("=====================================================")
print()
print(data['Review_text'].head(10))

print("=====================================================")
print("           After Applying NLP                  ")
print("=====================================================")
print()
data["Clean"] = data["Review_text"].apply(preprocess_data)

print(data["Clean"].head(10))
```

COUNT VECTORIZATION

```
#================== VECTORIZATION ================================
from sklearn.feature_extraction.text import CountVectorizer
```

```
X = data["Summary_Clean"]
y = data['Rev_Type']
vector = CountVectorizer(stop_words = 'english', lowercase = True)

#fitting the data
training_data = vector.fit_transform(X)

#transform the test data

print("=============================================================")
print("---------- Vectorization ----------")
print("=============================================================")
print()
print(training_data)
```

DATA SPLITTING

```
#======================== DATA SPLITTING ============================
from sklearn.model_selection import train_test_split
X_train, X_test,y_train, y_test = train_test_split(training_data, y, test_size=0.2, random_state=1)
```

CLASSIFICATION AND RESULT GENERATION

```
from sklearn import linear_model


print("=============================================================")
print("---------- LOGISTIC REGRESSION ----------")
print("=============================================================")
print()

#=== initialize the model ===
lr= linear_model.LogisticRegression()
```

```
#=== fitting the model ===
lr = lr.fit(X_train, y_train)

#=== predict the model ===
y_pred_ada = lr.predict(X_train)
y_pred_te=lr.predict(X_test)


from sklearn import metrics

print()
print("Performances analysis for logistic regression")
acc_lr=metrics.accuracy_score(y_pred_ada,y_train)*100
print(" train Accuracy :",acc_lr,'%')
print()
test_lr=metrics.accuracy_score(y_pred_te,y_test)*100


pre_lr=metrics.precision_score(y_train,y_pred_ada)*100
print(" Precision :",pre_lr,'%')
print()
recall_lr=metrics.recall_score(y_train,y_pred_ada)*100
print("Recall :",recall_lr,'%')
print()

f1_lr=metrics.f1_score(y_train,y_pred_ada)*100
print(" F1 score :",f1_lr,'%')
print()

# Import necessary libraries
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# Create confusion matrix
cm = confusion_matrix(y_pred_te,y_test)

# Define class labels
classes = ['0', '1']

# Create heatmap using Seaborn
sns.heatmap(cm, annot=True,fmt='d',cmap='Blues',xticklabels=classes, yticklabels=classes)

# Add labels and title
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.title('Confusion  Matrix of Logistic Regression')
#=========================PREDICTION=============================
==
for i in range(0,10):
    if y_pred_ada[i]==0:
        print("==========================")
        print()
        print(" The review is fake ")
    else:
        print("==========================")
        print()
        print(" The review is real ")

# Show plot
plt.show()
from xgboost import XGBClassifier
xg= XGBClassifier()
#=== fitting the model ===
xgb = xg.fit(X_train, y_train)
```

```
#=== predict the model ===
y_pred_xgb = xg.predict(X_train)
y_pred_test=xg.predict(X_test)
from sklearn import metrics
print()
print("Performances analysis for XGboost")
test_xgb=metrics.accuracy_score(y_train,y_pred_xgb)*100
print( " train Acuracy :",test_xgb,'%')
print()
tr_xgb=metrics.accuracy_score(y_test,y_pred_test)*100
pre_xgb=metrics.precision_score(y_train,y_pred_xgb)*100
print(" Precision :",pre_xgb,'%')
print()
recall_xgb=metrics.recall_score(y_train,y_pred_xgb)*100
print(" Recall :",recall_xgb,'%')
print()
f1_xgb=metrics.f1_score(y_train,y_pred_xgb)*100
print(" F1 score :",f1_xgb,'%')
print()
# Import necessary libraries
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
# Create confusion matrix
cm = confusion_matrix(y_pred_test,y_test)
# Define class labels
classes = ['fake', 'real']
# Create heatmap using Seaborn
sns.heatmap(cm, annot=True,fmt='d',cmap='Blues',xticklabels=classes, yticklabels=classes)
# Add labels and title
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.title('Confusion Matrix of XGBOOST ')
```

```
# Show plot
plt.show()
#=========================PREDICTION=================================
==

for i in range(0,10):
    if y_pred_xgb[i]==0:
        print("============================")
        print()
        print(" The review is fake ")
    else:
        print("============================")
        print()
        print(" The review is real ")

VISUALIZATIONS
#pie graph
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize = (5,5))
counts =data['Rev_Type'].value_counts()
plt.pie(counts, labels = counts.index, startangle = 90, counterclock = False, wedgeprops =
{'width' : 0.6},autopct='%1.1f%%', pctdistance = 0.55, textprops = {'color': 'black',
'fontsize' : 15}, shadow = True,colors = sns.color_palette("Paired")[3:])
plt.text(x = -0.35, y = 0, s = 'Total Reviews: {}'.format(data.shape[0]))
plt.title('Total No.of reviews', fontsize = 14);
plt.show()
import numpy as np
objects = ('LOGISTIC REGRESSION', 'XGBOOST')
y_pos = np.arange(len(objects))
performance = [acc_lr,test_xgb]
plt.bar(y_pos, performance, align='center', alpha=0.5)
plt.xticks(y_pos, objects)
```

```
plt.ylabel('Accuracy')
plt.title('ACCURACY COMPARISON OF LOGISTIC REGRESSION AND XGBOOST ')
plt.show()
print()
print("------------------- Performance Comparison --------------------")
print()
import matplotlib.pyplot as plt
barWidth =0.1
fig = plt.subplots(figsize =(6, 6))
lr= [acc_lr,pre_lr,recall_lr,f1_lr]
xg= [test_xgb,pre_xgb,recall_xgb,f1_xgb]
lr=[round(num, 1) for num in lr]
br1 = np.arange(len(lr))
br2 = [x + barWidth for x in br1]
plt.bar(br1, lr, color ='r', width = barWidth,
        edgecolor ='grey', label ='LOGISTIC REGRESSION')
plt.bar(br2, xg, color ='b', width = barWidth,
        edgecolor ='grey', label ='XGBOOST')
plt.xlabel('Performance Metrics', fontweight ='bold', fontsize = 15)
plt.ylabel('Algorithms', fontweight ='bold', fontsize = 15)
plt.xticks([r + barWidth for r in range(len(lr))], ['Accuracy', 'Precision','Recall','F1 score'])
plt.legend()
plt.show()
```

10.3 Plagiarism Report

Ouriginal by Turnitin

Document Information

Analyzed document	project doc 21pite15.pdf (D164526013)
Submitted	4/20/2023 12:04:00 PM
Submitted by	J.A. ESTHER RANI
Submitter email	estherrani@ldc.edu.in
Similarity	12%
Analysis address	estherrani.ldc@analysis.urkund.com