FAK-DET: FAKE REVIEW DETECTION

A PROJECT REPORT

Submitted by

BASIL JASHEEM

Reg.No: LMCT18MCA013

to

The APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the Degree

of

Master of Computer Applications

Department of Computer Applications

MOHANDAS COLLEGE OF ENGINEERING AND TECHNOLOGY

ANAD, NEDUMANGAD, THIRUVANTHAPURAM-695544

2021

FAK-DET: FAKE REVIEW DETECTION

A PROJECT REPORT

Submitted by

BASIL JASHEEM

Reg.No: LMCT18MCA013

to

The APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the Degree

of

Master of Computer Applications

Department of Computer Applications

MOHANDAS COLLEGE OF ENGINEERING AND TECHNOLOGY

ANAD, NEDUMANGAD, THIRUVANTHAPURAM-695544

2021

DECLARATION

I undersigned hereby declare that the project report “FAK-DET: FAKE REVIEW DETECTION”, submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Prof. **Jayanthi T**. This submission represents my ideas in my own words and where ideas or words of others have been included; I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Nedumangad

Date:

BASIL JASHEEM

DEPARTMENT OF COMPUTER APPLICATIONS

MOHANDAS COLLEGE OF ENGINEERING & TECHNOLOGY

Anad, Nedumangad, Thiruvananthapuram- 695544

CERTIFICATE

This is to certify that the report entitled **FAK-DET**: FAKE REVIEW DETECTION submitted by **BASIL JASHEEM** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Master of Computer Applications is a bonafide record of the project work carried out by her under my/our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor(s)

Project Coordinator

Head of the Department

External Examiner

CONTENT

ACKNOWLEDGMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

At the outset, I thank God Almighty for standing by me throughout the project and making it possible for me to complete the project within the stipulated time.

I wish to record my deep sense of gratitude to our **Director, Dr. Ashalatha Thampuran**, and **Principal**, **Dr. S. Sheela**, for their extensive support and guidance throughout the course of my project.

I owe my deep gratitude to my project guide **Prof. Jayanthi T** and **HOD Prof. Sreeja K** who took keen interest on my project work and guided me all along till the completion of our project work by providing all the necessary information for developing a good system.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of MCA which helped me in successfully completing the project work. Also, I would like to extend my sincere esteems to all staff in laboratory for their timely support.

Finally, I wish to express my sincere gratitude to all our friends, who directly or indirectly contributed in this venture.

BASIL JASHEEM

ABSTRACT

FAK-DET: FAKE REVIEW DETECTION

E-commerce proved its importance based on the fact where time is the essence. People are relying on e-commerce more than before. With e-commerce comes a huge amount of user feedback based on the products they buy. As the internet has become cheaper and easy to get, more people are getting connected through different social media and platform where they are expressing product-related feedbacks. With the rise of e-commerce, people are relying more on product reviews to get a clear view and user experience. But there is no convincing way to authenticate the reviews posted on products on e-commerce websites. To generate more revenue and fulfill some immoral benefits, some sellers are making investments and hiring people to post fake reviews. These fake reviews are generated to convince people to buy the product. The idea proposed here is to create application capable of identifying these reviews by employing two learning instead of one (active and supervised) along with PCA (Principal component analysis) to create the best possible dataset and will go through four Classification Algorithms each giving results over 80% of accuracy score. The application can be employed to verify live or real time product reviews from sites like Amazon, Google etc.

1. INTRODUCTION

1.1 OVERVIEW OF PROJECT

The usage of internet is increasing day by day as the world is becoming more digital and therefore the internet is easily accessible in both rural and urban areas. This has also brought commercial affairs to the web where not only the consumer but also the business is getting benefitted. People can easily post product reviews, views, experiences in blogs, discussion forums and on social platforms. These are addressed as user generated contents. As people have the liberty to write whatever they want, there is no monitoring available. Sharing personal opinion, experience with a product is known as reviews. These reviews can attract and influence people to buy a product because people are getting a real-life experience on the product from someone else. These reviews have become a part and parcel to the buyers while buying a new product or an existing one. As these reviews make an influence on the buyers' side, some people provide fake reviews to increase the sale of the products found on e-commerce websites. These people are mainly known as opinion spammers and their activities are known as opinion spamming. The number of fake reviews is increasing day by day. Some of the sellers are taking the chance to grow the business quickly by paying opinion spammers to write fake reviews. In fact, there are many websites that are paying to write fake reviews on different platforms. Therefore, detecting these fake reviews has become a serious issue to maintain the trust factor between the buyer and the customer. In this project "FAK-DET: FAKE REVIEW DETECTION" proposes the idea of a socio-commercial platform were the existence of fake or intentional promotional reviews can be compromised.

2. SYSTEM ANALYSIS

2.1 PROBLEM DEFINITION

Existence of fake reviews makes it difficult to analyze the quality customer opinion of the product in a virtual market. A review that works similar to the concept of "word of mouth" from a fellow consumer is been forged, the issue affects both main entities in a transaction as in the consumer and the retailer since the fake review works in both ways. These can be promotion stunts of the retailer or a demotion tactic by a competitive retailer. This will give the consumer the misconceptions about the product which will very well later lead to decline in costumer satisfaction which is considered one of the main pillars of ecommerce.

2.2 PROBLEM SOLUTION

The idea proposed for overcoming the above problem is to employ an entity to analyze and scrutinize the provided reviews and their attributes to evaluate the reviews as fake or not, it is to recognize the patterns and inconsistencies occur in reviews that points out the review to be on reliable.

2.3 EXISTING SYSTEM

The current methods employed for tackling this issue is manual assessing by a certain individual or a group, where all the disadvantages of human limitations are present and will be clearly visible, though there are automated methods available for this purpose they all employ a bulk of pseudo reviews as a dataset for training the model.

DRAWBACKS OF EXIXTING SYSTEM

- Inaccuracies
- Modification
- Inefficiency

- Time and effort

2.4 PROPOSED SYSTEM

There are many researchers who have come up with different spam detection techniques. But the major issue here is to find out an enriched real-life labeled dataset. Consequently, the existing solutions depend on pseudo fake reviews. Even some researchers are using a different psychological approach to detect a pattern of fake reviews and create a dataset, but the ideal solution will be to have a Enriched and custom labeled dataset: The dataset that has been prepared is labeled manually. From a chunk of 800,000 reviews collected from the dataset of the authors R. He, J. McAuley. of “Ups and downs: Modeling the visual evolution of fashion trends with one class collaborative filtering” & J. McAuley, C. Targett, J. Shi, A. van den Hengel. “Imagebased recommendations on styles and substitutes”, and sorted out based on review writers, writing pattern, keyword-based search such as “Honest review”, “Fake review” and pseudo fake reviews written and combined them together to create an enriched dataset that adds more versatility and efficacy in real-life fake review detection. And complimenting the feature selection methods like vectorizers and n-grams values to labels dataset with Four Clustering Algorithms Namely Logical Regression, Decision Tree, Random Forest, and Gradient Boosting.

ADVANTAGES OF PROPOSED SYSTEM

- User Friendly:
- Speed and Accuracy:
- Efficiency and flexibility:
- Automation:
- Availability:

2.5 IDENTIFICATION OF NEED

- To minimize the time.
- Greater efficiency
- Betters service to the consumers

- Better information retrieval
- More reliable system.

2.6 FEASIBILITY STUDY

Feasibility study is a test of system proposed regarding its workability, impact on the organization, ability to meet the needs and effective use of resources. Thus, when a new project is proposed, it normally goes through a feasibility study before it is approved for development.

A feasibility study is made to see if the project on completion will serve the purpose of the organization of the amount of work, effort and the time that is spend on it. Feasibility study lets the developer foresee the future of the project and its usefulness.

ECONOMICAL FEASIBILITY

The employing of the proposed system will not cost the organization any more than what they spent for an e-commerce platform, there are no much usage of any additional usage of database, and can be performed by the means of an existing server, the expenditure focus should be to collect and update the dataset contents according to the change of time, employing the proposed system complements the credibility of the organization.

OPERATIONAL FEASIBILITY

The main problem faced during development of a new system is getting acceptance from the user. People are inherently resistant to changes and computers have been known to facilitate change. It is mainly related to human organizational and political aspects.

The proposed system in the project is completely automated as in the user will be handling their same old familiar operations, the process of the proposed module to verify the user reviews along with their operation without hindering the user’s process or performance.

TECHNICAL FEASIBILITY

Technical feasibility centers on the existing computer system (hardware, software, etc.) and to what extend it can support the proposed addition. This involves financial considerations to accommodate technical enhancements. If the budget is a serious constraint, then the project is judged not feasible.

The proposed system is considered a light weight module as it can be embedded along to the existing application without much changes to the existing computer system. Especially due to the absence of any additional software or hard ware the end user will not have any addition expenditure to use the proposed system.

2.7. SYSTEM SPECIFICATION

2.7.1 HARDWARE REQUIREMENTS

Hardware is a set of physical components, which performs the functions of applying appropriate, predefined instructions. In other words, one can say that electronic and mechanical parts of computers constitute hardware. This web application is designed on a powerful programming language JAVA. The backend is MYSQL, which is used to maintain the database. It can run on almost all the microcomputers.

Processor	: processor of 2.0 GHz or more
RAM	: 2 GB of RAM or above
Cache Memory	: 1 MB or above
Hard Disk	: 10.2 GB or above

2.7.2 SOFTWARE REQUIREMENTS

A major element in building systems in compatible application. The system analyst has to determine what software package is best for the system. It begins with requirements Analysis, followed by a request for proposal and vendor evaluation.

OPERATING SYSTEM	: Windows 7 or above
LANGUAGE (Back-End)	: Python
DATABASE (ORM)	: SQLAlchemy
SERVER	: CMD (localhost:5000- FLASK)-Python 3.7.3 or above
IDE(or TextEditor)	: Sublime_Text3
FRAMEWORK	: FLASK
FRONT-END	: Bootstrap (HTML, CSS, JavaScript)

2.7.3 SOFTWARE DEVELOPMENT

PYTHON:

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Features in Python:

There are many features in Python, some of which are discussed below –

i. **Easy to code**

ii. **Free and Open Source**

iii. **Object-Oriented Language**

iv. **GUI Programming Support**

v. **Extensible feature**

vi. **Python is Portable language**

vii. **Python is Integrated language**

viii. **Interpreted Language**

ix. **Large Standard Library**

x. **Dynamically Typed Language**

Libraries Employed:

PANDAS

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals

SKLEARN

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

NUMPY

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.[5] The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

PIL

Python Imaging Library (PIL) is a free and open-source additional library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats. It is available for Windows, Mac OS X and Linux.

SECRETS

The secrets module is used for generating random numbers for managing important data such as passwords, account authentication, security tokens, and related secrets that are cryptographically strong. This module is responsible for providing access to the most secure source of randomness.

WTFORMS

WTForms is a flexible forms validation and rendering library for Python web development. It can work with whatever web framework and template engine you choose. It supports data validation, CSRF protection, internationalization (I18N), and more. There are various community libraries that provide closer integration with popular frameworks.

FLASK-SQLALCHEMY

Flask-SQLAlchemy is an extension for Flask that adds support for SQLAlchemy to your application. It aims to simplify using SQLAlchemy with Flask by providing useful defaults and extra helpers that make it easier to accomplish common tasks.

FLASK-LOGIN

Flask-Login provides user session management for Flask. It handles the common tasks of logging in, logging out, and remembering your users’ sessions over extended periods of time.

HTML:

Hypertext Markup Language (HTML) is the standard mark-up language for creating web pages and web applications. With Cascading Style Sheets (CSS) and JavaScript, it forms a triad of cornerstone technologies for the World Wide Web. Developed by **Tim Berners-Lee**.

Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

CSS:

Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript.
CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. This separation can improve content accessibility, provide more flexibility and control in the specification of presentation characteristics, enable multiple web pages to share formatting by specifying the relevant CSS in a separate .css file, and reduce complexity and repetition in the structural content.

JAVASCRIPT:

JavaScript, implemented as an integrated component of the web browser, allowing the development of enhanced user interface and dynamic website. JavaScript is a dialect of the ECMA Script was influenced by many languages and was designed to look like Java, but to be easier for non-programmers to work with. JavaScript, despite the name, is essentially unrelated to the Java programming language even though the two do have superficial similarities. Both languages use syntaxes influenced by that of C syntax, and JavaScript copies many Java names and naming conventions.

BOOTSTRAP:

Bootstrap is a web framework that focuses on simplifying the development of informative web pages (as opposed to web apps). The primary purpose of adding it to a web project is to apply Bootstrap's choices of color, size, font and layout to that project. As such, the primary factor is whether the developers in charge find those choices to their liking. Once added to a project, Bootstrap provides basic style definitions for all HTML elements. The end result is a uniform appearance for prose, tables and form elements across web browsers. In addition, developers can take advantage of CSS classes defined in Bootstrap to further customize the appearance of their contents.

FLASK:

Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries.[2] It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

SQLALCHEMY:

SQLAlchemy is an open-source - SQL - Toolkit and ORM - Framework for the programming language Python with the aim of Object-relational impedance mismatch in the way of Java's Hibernate to circumnavigate.

SQLAlchemy offers a number of design patterns for efficient persistence of data in a relational database. The motivation behind SQLAlchemy is based on the fact that SQL databases are less similar to object collections, the more extensive the data stock and the more power is required, while object collections behave less like relations and tuples, the more is abstracted between data representation and mini-world.

3. METHODOLOGY: SCRUM PLAN, ROLES, REVIEWS

3.1 AGILE METHODOLOGY

Agile software development refers to software development methodologies centered round the idea of iterative development, where requirements and solutions evolve through collaboration between self-organizing cross-functional teams. The ultimate value in Agile development is that it enables teams to deliver value faster, with greater quality and predictability, and greater aptitude to respond to change. Scrum and Kanban are two of the most widely used Agile methodologies.

3.2 SCRUM PLAN (Sprint planning)

Sprint planning is an event in scrum that kicks off the sprint. The purpose of sprint planning is to define what can be delivered in the sprint and how that work will be achieved. Sprint planning is done in collaboration with the whole scrum team. In scrum, the sprint is a set period of time where all the work is done. However, before you can leap into action you have to set up the sprint. You need to decide on how long the time box is going to be, the sprint goal, and where you're going to start. The sprint planning session kicks off the sprint by setting the agenda and focus.

3.3 ROLES

Scrum defines three roles: Scrum Master, Product Owner, and Development Team. Together all three roles make up a Scrum Team.

The **Product Owner** defines the what--as in what the product will look like and what features it should contain. The Product Owner is expected to incorporate stakeholder feedback to create the highest value product increments each and every sprint.

The **Development Team** decides how to accomplish the work set forth by the Product Owner. Development Teams are structured and empowered to organize and manage their own work.

The **Scrum Master** helps the Scrum Team perform at their highest level. They also protect the team from both internal and external distractions. Scrum Masters hold the Scrum Team accountable to their working agreements, Scrum values, and to the Scrum framework itself**.**

3.4 REVIEWS

The **Sprint Review** meeting is the before-last Event of the Sprint and should take place every Sprint. Only the Retrospective follows (right after) the Sprint Review, and after the Retrospective, a new Sprint begins.

A Sprint Review is held at the end of the Sprint to inspect the Increment and adapt the Product Backlog if needed. During the Sprint Review, the Scrum Team and stakeholders collaborate about what was done in the Sprint. Based on that and any changes to the Product Backlog during the Sprint, attendees collaborate on the next things that could be done to optimize value. This is an informal meeting, not a status meeting, and the presentation of the Increment is intended to elicit feedback and foster collaboration.

3.5 PRODUCT BACKLOG

Product Backlog is an ordered list of everything that is known to be needed in the product. It is the single source of requirements for any changes to be made to the product. The Product Owner is responsible for the Product Backlog, including its content, availability, and ordering.

FAKE-DET BACKLOG

Date of Submission : 2021, June First Week　　Priority Scale : MoSCoW Method

SL	User Stories	Scrum Master cmt	Additional Info	Priority
1	**Problem identification** • Deeply study the problem in different platforms	Refer different sources to gather ideas for the proposed system		Must
2	**Data Collection**	Gather data from multiple choices	Bulk DataSets of the sizesof 3-4 Gigabytes	Must
3	**Filtering a Train_Dataset**	A training Dataset below 10000 reviews to be retrived from the Bulk Dataset		Must
4	**Convert the Dataset** • CSV conversion	Convert the Dataset in to the suitable Format as in txt->json or json->csv		Must
5	**Install the Software Requirements** • IDE • LANGUAGE • DATABASE • SERVER		Setting up a suitable IDE,Language etc	Must
6	**Download Python libraries** • Pandas • Sklearn • Flask • Numpy • Secret • Socket..	Libraries to be downloaded pre to codi ng to ensure smooth development	Library version combatabiliy is to be noted	Must
7	**Vectorizarion** • Count Vectorizer		CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words.	Should

8	**Performing PCA** Principal component analysis	To reduce Dimensionality		Must
9	**Classification and Training** • Logistic Regression • Decision Tree Classification • Gradient Boosting Classification • Random Forest Classification	4 Different Classifiers to Increase Efficiency		Must
10	**Manual Entry Model Module**		Model Testing With Manual Entry	Must
11	**Design Forms** • REGISTRATION • LOGIN • REVIEW • PRODUCT		Registration Form, Login Form, Admin Form, Review Form, Change Form, Forget Form, Product Form.	Must
12	**Database Design**		Admin table, Product table, User tableReview table.	Must
13	**Design Routes**		Branching network that connects all pages/interface of the project	Must
14	**Design Interface pages**		Interface built using Bootstrap	Must
15	**Embed Review Analyzing module**	Review is classified to fake or not by an auto- mated module in the system on real time		Must
16	**Testing** • Testing without Web Frame Work • Testing with Web Frame Work	Perform all the final test with and without inputs		Must
17	**Documentation**		The documentation either explains how the software operates or how to use it, and may mean different things to people in different roles.	Should

3.6 SPRINT BACKLOG

Sprint Backlog is the set of Product Backlog items selected for the Sprint, plus a plan for delivering the product Increment and realizing the Sprint Goal. The Sprint Backlog is a forecast by the Development Team about what functionality will be in the next Increment and the work needed to deliver that functionality into a "Done" Increment.

SPRINT BACKLOG				
SL. NO.	**TASK/ SPRINT**	**STATUS**		
		NOT STARTED	***IN PROGRESS***	***COMPLETED***
1	**Problem identification**			✓
2	**Collect Bulk Review Dataset**			✓
3	**Filtering a Train_Dataset** **&** **Convert the Dataset**			✓
4	**Install the Software Requirements** **&** **Download Python libraries**			✓
5	**Vectorizarion, Performing PCA** **&** **Classification and Training**			✓
6	**Manual Entry Model Module**			✓
7	**Design Forms**			✓
8	**Database Design**			✓
9	**Design Routes**			✓
10	**Design Interface pages**			✓
11	**Embed Review Analyzing module**			✓

12	**Testing**			✓
13	**Documentation**			✓

3.7 SCRUM BOARD

A Scrum Board (also called Scrum Task Board) is a tool that helps Teams make Sprint Backlog items visible. The board can take many physical (i.e. whiteboard and stickers) and virtual forms (i.e. software tools) but it performs the same function regardless of how it looks. A Scrum Board is the focal point of any agile project and serves as a good place at which to hold the stand-up meeting. The board is updated and referred by the Team and shows all items during the Daily Scrum keeps the team focused on the tasks that remain and their priorities.

SCRUM BOARD					
SL.NO	**TASK/SPRINT**	**START**	**FINISH**	**DURATION**	**STATUS**
1	SPRINT -1 Problem identification	10-03-2021	17-03-2021	8 days	Done
2	SPRINT-2 Collect Bulk Review Dataset	18-03-2021	05-04-2021	20 days	Done
3	SPRINT-3 Filtering a Train_Dataset & Convert the Dataset	06-04-2021	18-04-2021	13 days	Done
4	SPRINT-4 Install the Software Requirements & Download Python libraries	19-04-2021	20-04-2021	2 day	Done
5	SPRINT-5 Vectorizarion, Performing PCA & Classification and Training	21-04-2021	25-04-2021	5 days	Done

6	SPRINT-6 Manual Entry Model Module	26-04-2016	29-04-2021	3 days	Done
7	SPRINT-7 Design Forms	30-04-2021	02-05-2021	3 days	Done
8	SPRINT-8 Database Design	03-05-2021	05-05-2021	2 days	Done
9	SPRINT-9 Design Routes	06-05-2021	13-05-2021	8 days	Done
10	SPRINT-10 Design Interface pages	14-05-2021	19-05-2021	6 days	Done
11	SPRINT-11 Embed Review Analyzing module	20-05-2021	25-05-2021	6 days	Done
12	SPRINT-12 Testing	26-05-2021	31-05-2021	6 days	Done
13	SPRINT-13 Documentation	01-05-2021	07-06-2021	7 days	Done

3.8 CODE SAMPLE

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report

import re
import string

df_manual_testing = pd.read_csv("review.csv")

df_manual_testing.head(10)

df_manual_testing.shape

df_manual_testing.columns

df = df_manual_testing.drop (["_id/$oid", "reviewerID", "asin", "reviewerName",
        "helpful/0", "overall", "summary", "unixReviewTime", "reviewTime",
        "category"], axis = 1)

df.head(5)

df.reset_index(inplace = True)

df.drop(["index"], axis = 1, inplace = True)

df.columns

def wordopt(text):

  text = text.lower()

  text = re.sub('\[.*?\]', '', text)

  text = re.sub("\\W"," ",text)

  text = re.sub('https?://\S+|www\.\S+', '', text)

  text = re.sub('<.*?>+', '', text)

  text = re.sub('[%s]' % re.escape(string.punctuation), '', text)

  text = re.sub('\n', '', text)

  text = re.sub('\w*\d\w*', '', text)

  return text


df["reviewText"] = df["reviewText"].apply(wordopt)

x = df["reviewText"]

y = df["class"]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)

from sklearn.decomposition import SparsePCA
pca = SparsePCA()
xv_train = pca.fit_transform(xv_train.toarray())
xv_test = pca.transform(xv_test.toarray())

from sklearn.linear_model import LogisticRegression
LR = LogisticRegression()
LR.fit(xv_train,y_train)
pred_lr=LR.predict(xv_test)
LR.score(xv_test, y_test)

from sklearn.tree import DecisionTreeClassifier
DT = DecisionTreeClassifier()
DT.fit(xv_train, y_train)
pred_dt = DT.predict(xv_test)
DT.score(xv_test, y_test)


from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=0)
```

```
GBC.fit(xv_train, y_train)

pred_gbc = GBC.predict(xv_test)

GBC.score(xv_test, y_test)


from sklearn.ensemble import RandomForestClassifier

RFC = RandomForestClassifier(random_state=0)

RFC.fit(xv_train, y_train)

pred_rfc = RFC.predict(xv_test)

RFC.score(xv_test, y_test)


def output_lable(n):

    if n == 0:

        return 0

    elif n == 1:

        return 1



def manual_testing(news):

    testing_news = {"text":[news]}

    new_def_test = pd.DataFrame(testing_news)

    new_def_test["text"] = new_def_test["text"].apply(wordopt)

    new_x_test = new_def_test["text"]

    new_xv_test = vectorization.transform(new_x_test)

    pred_LR = LR.predict(new_xv_test)

    pred_DT = DT.predict(new_xv_test)

    pred_GBC = GBC.predict(new_xv_test)
```

```
pred_RFC = RFC.predict(new_xv_test)

l = pred_LR[0] + pred_DT[0] + pred_GBC[0] + pred_RFC[0]

analysis = [output_lable(pred_LR[0]), output_lable(pred_DT[0]),

            output_lable(pred_GBC[0]), output_lable(pred_RFC[0])]

return analysis
```

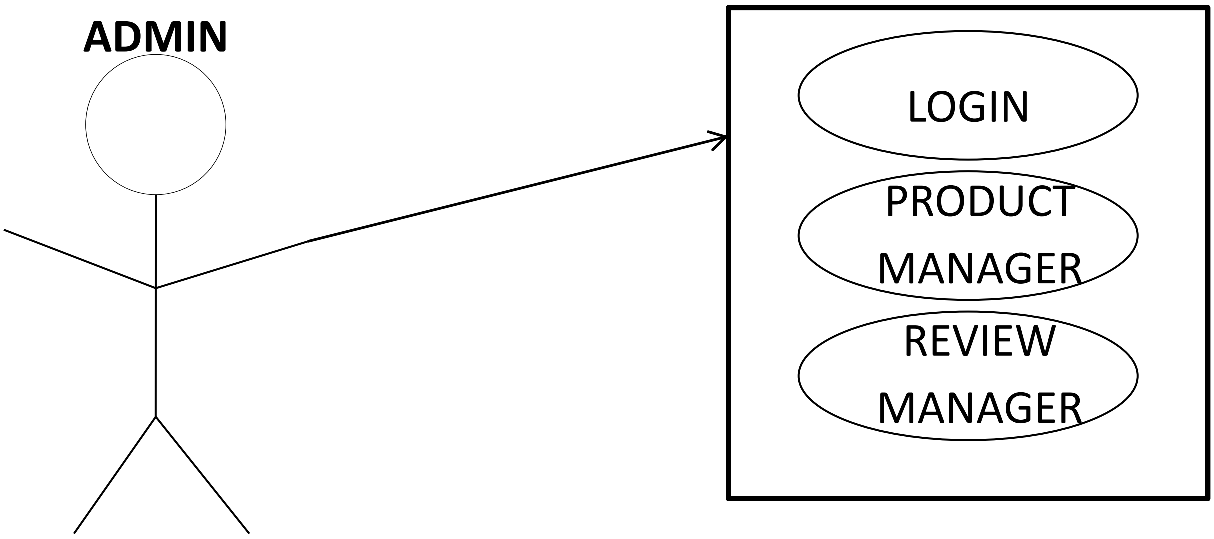
4.SYSTEM DESIGN

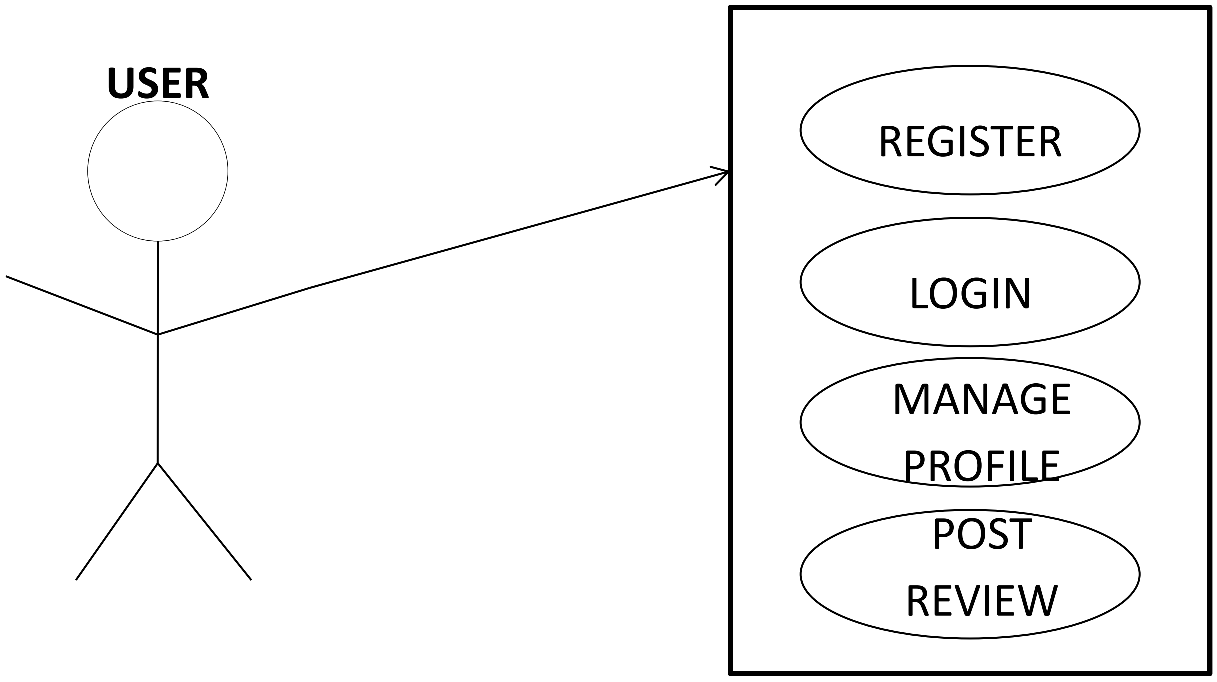
System design is the process of defining the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data goes through that system. It implies a systematic approach to the design of a system. It may take a bottom-up or top-down approach.

System design covers the following:

- Reviews the current physical system.
- Prepares output specifications.
- Prepares input specifications.
- Prepares a logical design walk through of the information flow, output, input, control, and implementation plan.

4.1 USECASE DIAGRAM





4.2 INPUT DESIGN

The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in proving input.
- Methods for preparing input validation and steps to follow when error occur.

4.3 OUTPUT DESIGN

Computer output is the most important and direct source of information to the user. Efficient output design should improve the systems relationship with the user and helps in decision-making. Designing computer output should proceed in an organized, well through manner: the right output be developed while ensuring that each output element is designed so that people will find the system easy to use effectively.

The output from an information system should accomplish one or more of the following objectives:

- View Detections (to user)
- View Detections (to admin in a detailed manner)
- Confirm an action.

4.4 TABLES

Table Name: Admin

Description: This table is used specifically for admin login, where only admin's details are saved and con only be accessed pages with admin authorization.

Primary Key: username

Column Name	Datatype	Desc	Comment
username	VARCHAR	Primary Key	
email	VARCHAR	unique	
password	VARCHAR		

Table Name: User

Description: Used to store user profile information as in email id, password. The password and photo stored in the table is in encrypted format, hence higher level of security.

Primary Key: username

Column Name	Datatype	Desc	Comment
username	VARCHAR	Primary Key	

email	VARCHAR	unique	
password	VARCHAR		

Table Name: Product

Description: This table contains the details of products placed in an e-commerce website where the proposed system is employed, it gives a product description.

Primary Key: item_id

Column Name	**Datatype**	**Desc**	**Com**
item_id	INT	Primary Key	
item_name	VARCHAR	Unique	
details	TEXT		
photo	LONG_BLOB		
Platform	VARCHAR		
category	VARCHAR		

Table Name: Review

Description: The table consist of reviews to the products in the web application, as well as their ratings and detection results from the automated module.

Primary Key: rev_id

COLUMN NAME	DATATYPE	DESC	COM
rev_id	INT	Primary Key	
rev_name	VARCHAR		
date	DATETIME		
item_id	INT	Foreign Key	
rev_text	TEXT		
status	INT	Fake or Not	
rate	FLOAT		
lr	INT	Logical regression	
dt	INT	Decision tree	
rfc	INT	Random forest	
gbc	INT	Gradient Boosting	
mark	INT	label	
ip	VARCHAR	Jp address	
i_mark	INT	Mark IP	
i_ignore	INT	IP status	

4.7 DATASET

A **data set** (or **dataset**) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. Data sets can also consist of a collection of documents or files.

Datasets used in these projects are of formats:

JSON (JavaScript Object Notation) is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays (or other serializable values). It is a very common data format, with a diverse range of applications, one example being web applications that communicate with a server.

A **CSV** (Comma-Separated Values) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

Main signals used for labeling:

- **Lexical features** such as word n-grams, part-of-speech n-grams, and other lexical attributes.

- **Content and style similarity** of reviews from different reviewers.

- **Semantic inconsistency** (we have never used this kind of features). For example, a reviewer wrote "My wife and I bought this car ..." in one review and then in another review he/she wrote "My husband really love ..." .

- **Public data available from Web sites**, e.g., reviewer id, time of posting, frequency of posting, first reviewers of products, and many more. For example, do you see anything wrong with the reviews from this user, Big John? What about after you see the reviews of these two users, Cletus and Jake? In fact, if you browse the reviews of their reviewed products, you will find another suspicious user/reviewer. This is just one example of atypical behaviors that our algorithms are able to discover.

- **Web site private/internal data** (we have not used such data, but they are extremely useful), e.g., IP and MAC addresses, time taking to post a review, physical location of the reviewer, etc (a lot of them).

The data set employed for the proposed project is Amazon product reviews from **Prof. Bing Liu** [1] & **Assis Prof. Naveed Hussain** [2]

Format:

```
"root":{
12 items
"_id":{
1 item
"$oid":
string"5a1321d5741a2384e802c552"
}
"reviewerID":
string"A3HVRXV0LVJN7"
"asin":
string"0110400550"
"reviewerName":
string"BiancaNicole"
"helpful":[
2 items
0
:
int4
1
:
int4
]
"reviewText":
string"Best phone case ever . Everywhere I go I get a ton of compliments on it. It was in perfect condition as well."
"overall":
int5
"summary":
string"A++++"
"unixReviewTime":
int1358035200
"reviewTime":
string"01 13, 2013"
"category":
string"Cell_Phones_and_Accessories"
"class":
int1
}
```

The dataset consists of about 40000 reviews from different segments of e commerce in the Amazon e-commerce website, As in **Electronics, Clothing, Sports, Cell phone and Accessories, Toys, Baby produc**ts etc. Though there are multiple attributes of the dataset is given, at the end these attributes are added up for classifying the rev text to two binary results i.e., 1 as in Not Fake/Spam and 2 as in Fake/Spams. Hence the “**ReviewText**” is taken as the independent Variable and the “**class**” is taken as the dependent variable for vectorization.

4.8 PRINCIPAL COMPONENT ANALYSIS & CLASSIFIERS

Principal component analysis (**PCA**) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

Advantages of PCA on the Project Dataset:

- Removes Correlated Features:
- Improves Algorithm Performance:
- Reduces Overfitting:

Classification is the problem of identifying to which of a set of categories (sub-populations) an observation, (or observations) belongs to. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.).

Why Multiple Classifiers:

A dataset of multiple segments of high Dimensionality is been used by the system that gives of about the result of 80% accuracy score. Which is considered good but more can be done on to increase the precision. As accessing more dataset will get accounted to feasibility of the product. Multiple Classifiers are performed and retrieved an average to the results which can be considered more precise.

Classifiers employed in the project:

LOGICAL REGRESSION:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

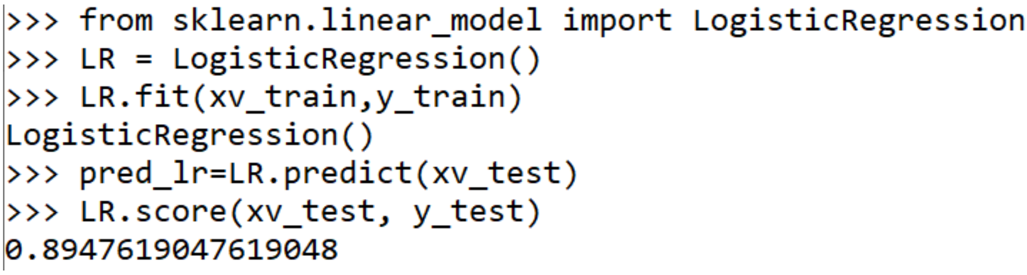
Sigmoid Function:

$$p = 1/1 + e^{-y}$$

Apply Sigmoid function on linear regression:

$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots\ldots \beta nXn)}$$

Model Development and Prediction



```
>>> from sklearn.linear_model import LogisticRegression
>>> LR = LogisticRegression()
>>> LR.fit(xv_train,y_train)
LogisticRegression()
>>> pred_lr=LR.predict(xv_test)
>>> LR.score(xv_test, y_test)
0.8947619047619048
```

DECISION TREE:

Decision tree learning or induction of decision trees is one of the predictive modeling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

The basic idea behind any decision tree algorithm is as follows:

- Select the best attribute using Attribute Selection Measures(ASM) to split the records.

- Make that attribute a decision node and breaks the dataset into smaller subsets.
- Starts tree building by repeating this process recursively for each child until one of the condition will match

Attribute Selection Measures(ASM)

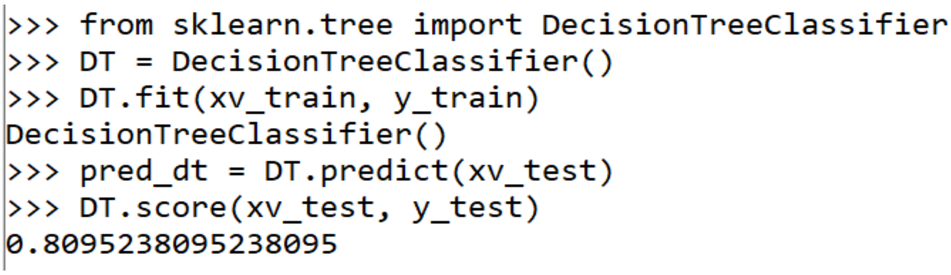
- Information Gain

$$\text{Info(D)}= - \sum_{i=1}^{m} pi \log_2 pi$$

- Gini Index

$$\text{Gini(D)}=1 - \sum_{i=1}^{m} \text{Pi}^2$$

Model Development and Prediction



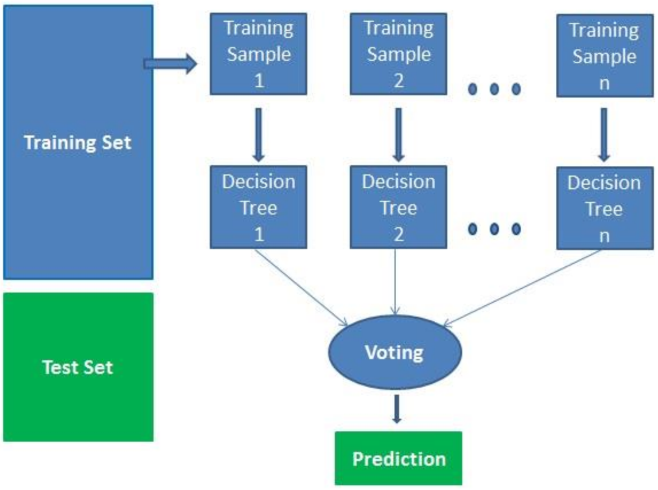
```
>>> from sklearn.tree import DecisionTreeClassifier
>>> DT = DecisionTreeClassifier()
>>> DT.fit(xv_train, y_train)
DecisionTreeClassifier()
>>> pred_dt = DT.predict(xv_test)
>>> DT.score(xv_test, y_test)
0.8095238095238095
```

RANDOM FOREST:

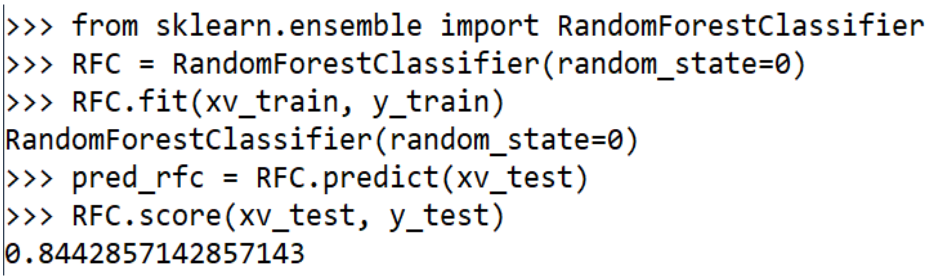
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

It works in four steps:

- Select random samples from a given dataset.
- Construct a decision tree for each sample and get a prediction result from each decision tree.
- Perform a vote for each predicted result.
- Select the prediction result with the most votes as the final prediction.



Model Development and Prediction

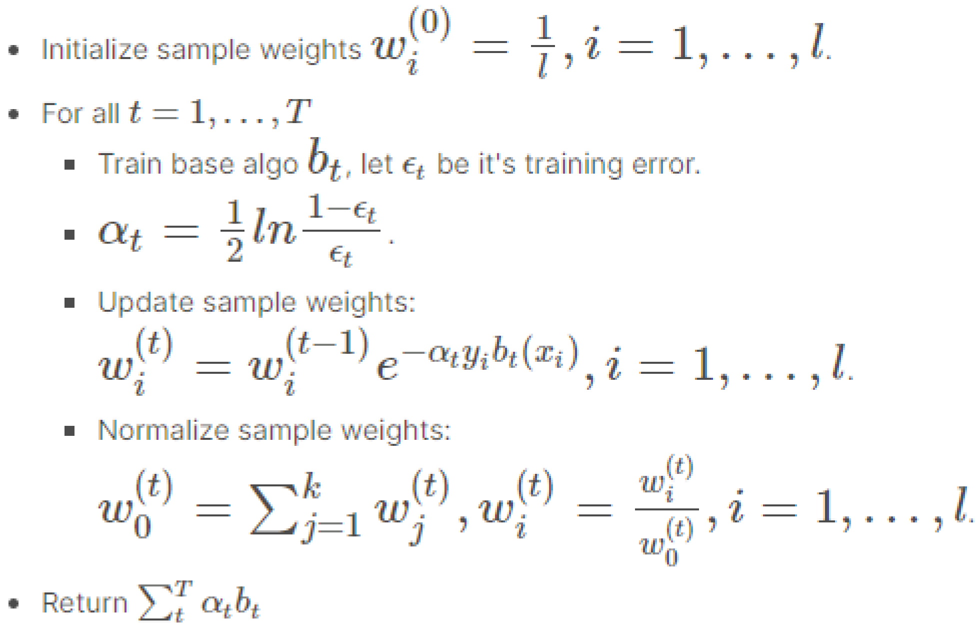


```
>>> from sklearn.ensemble import RandomForestClassifier
>>> RFC = RandomForestClassifier(random_state=0)
>>> RFC.fit(xv_train, y_train)
RandomForestClassifier(random_state=0)
>>> pred_rfc = RFC.predict(xv_test)
>>> RFC.score(xv_test, y_test)
0.8442857142857143
```

GRADIENT BOOSTING:

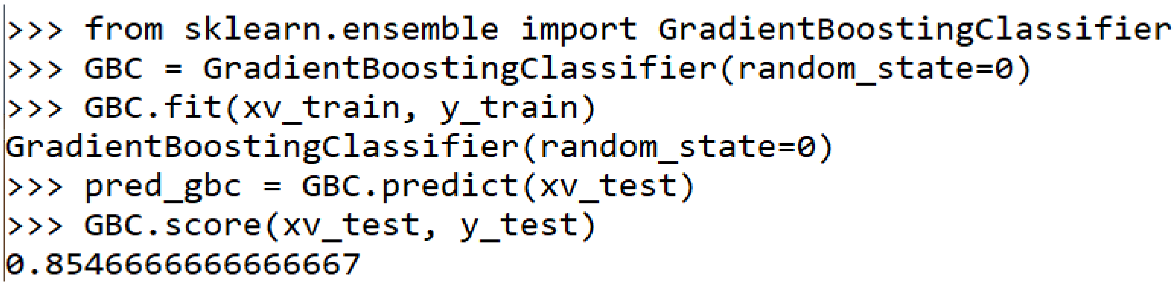
Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Pseudo code:



```
- Initialize sample weights $w_i^{(0)} = \frac{1}{l}, i = 1, \dots, l$.
- For all $t = 1, \dots, T$
    - Train base algo $b_t$, let $\epsilon_t$ be it's training error.
    - $\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$.
    - Update sample weights:
      $w_i^{(t)} = w_i^{(t-1)} e^{-\alpha_t y_i b_t(x_i)}, i = 1, \dots, l$.
    - Normalize sample weights:
      $w_0^{(t)} = \sum_{j=1}^{k} w_j^{(t)}, w_i^{(t)} = \frac{w_i^{(t)}}{w_0^{(t)}}, i = 1, \dots, l$.
- Return $\sum_t^T \alpha_t b_t$
```

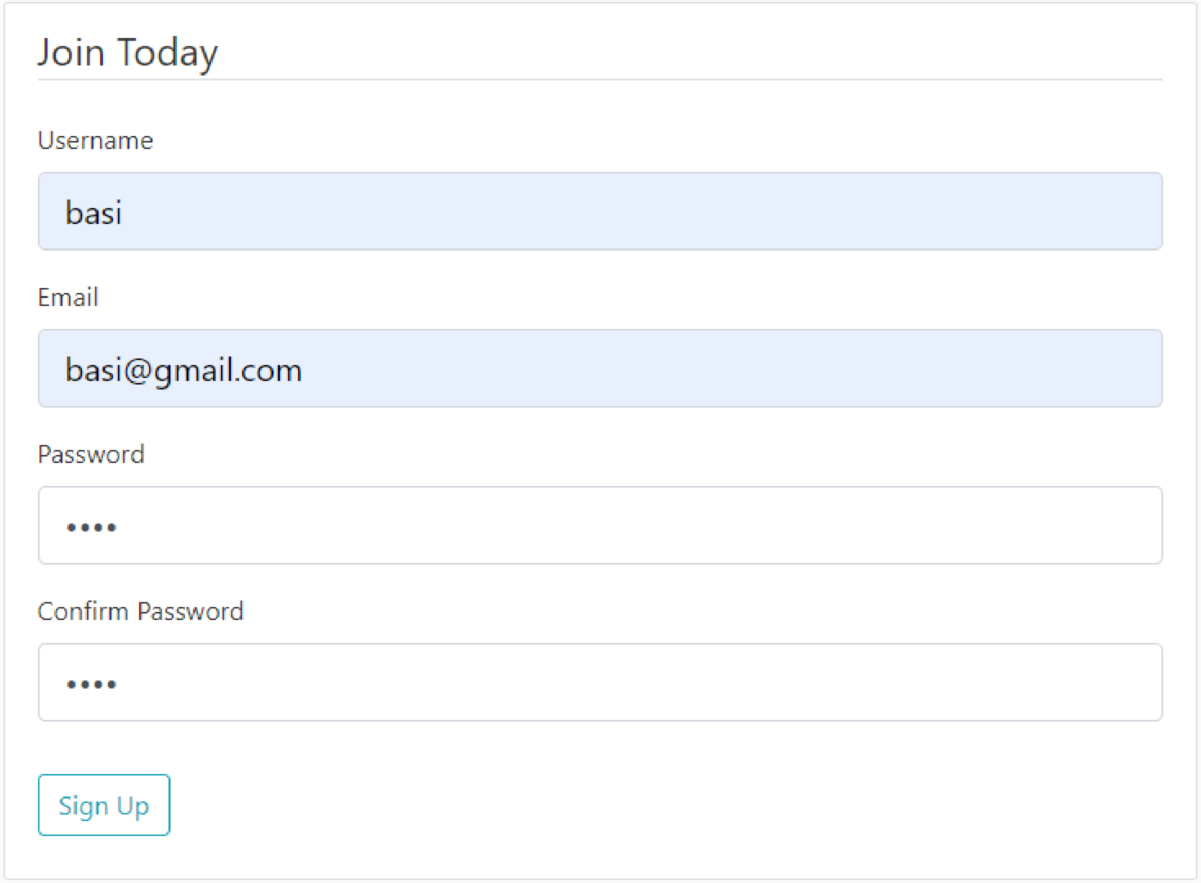
Model Development and Prediction



```
>>> from sklearn.ensemble import GradientBoostingClassifier
>>> GBC = GradientBoostingClassifier(random_state=0)
>>> GBC.fit(xv_train, y_train)
GradientBoostingClassifier(random_state=0)
>>> pred_gbc = GBC.predict(xv_test)
>>> GBC.score(xv_test, y_test)
0.8546666666666667
```

4.9FORMS

REGISTRATION FORM



Join Today

Username

basi

Email

basi@gmail.com

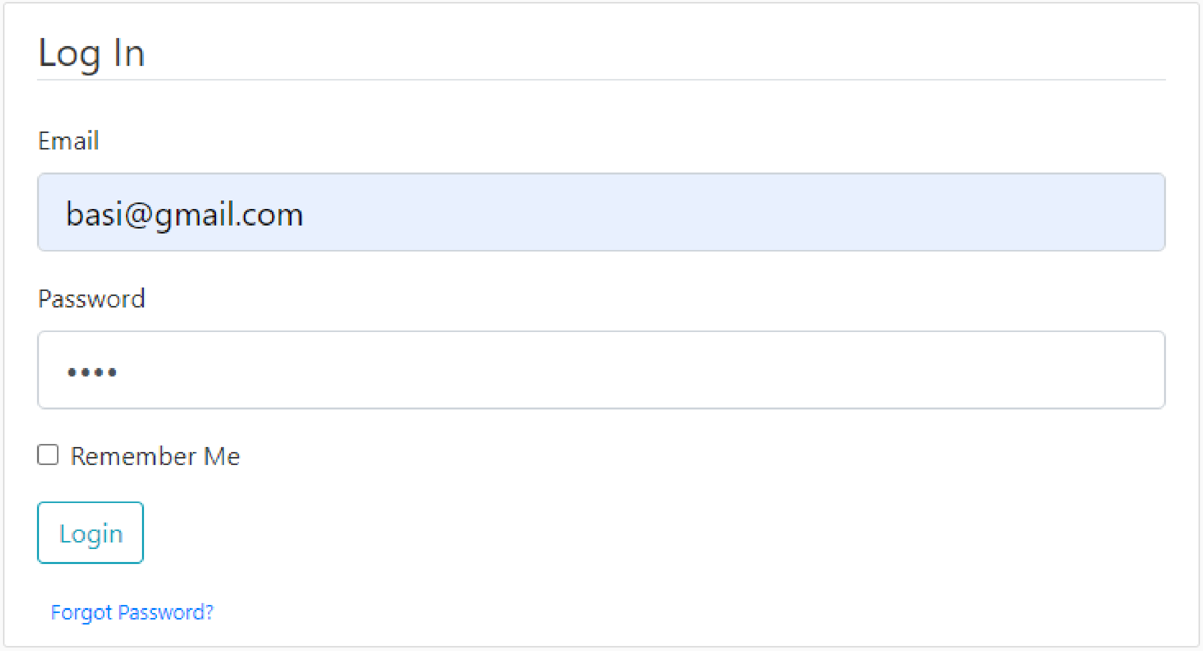
Password

••••

Confirm Password

••••

LOGIN FORM



Log In

Email

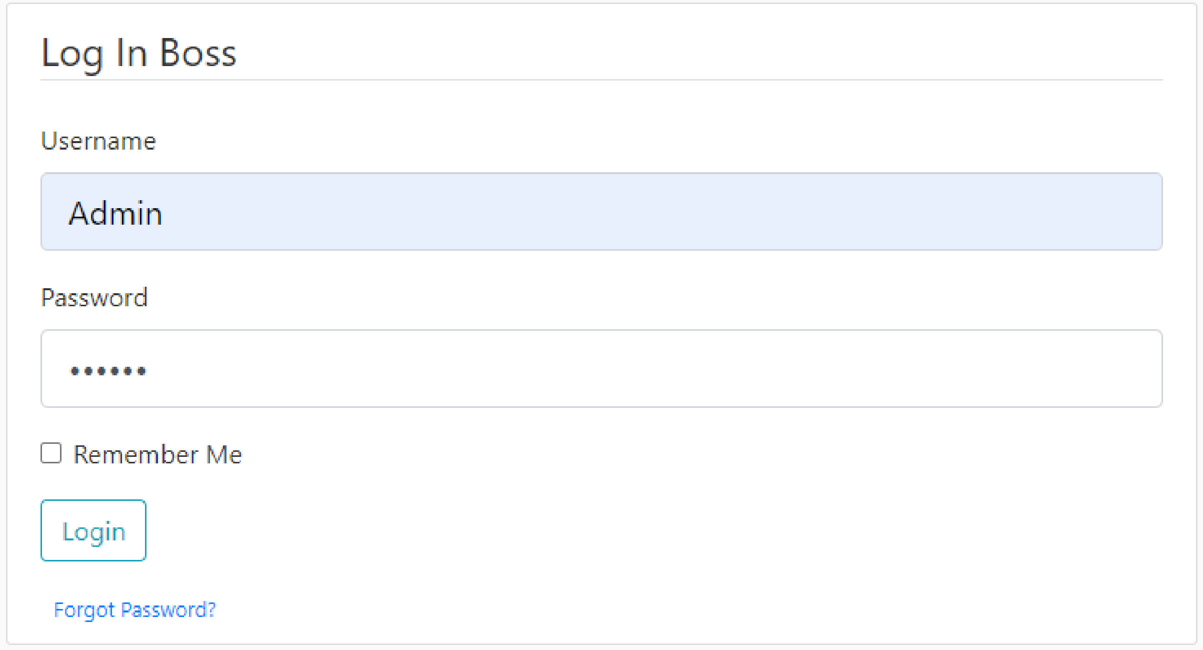
basi@gmail.com

Password

••••

□ Remember Me

ADMIN LOGIN FORM



Log In Boss

Username

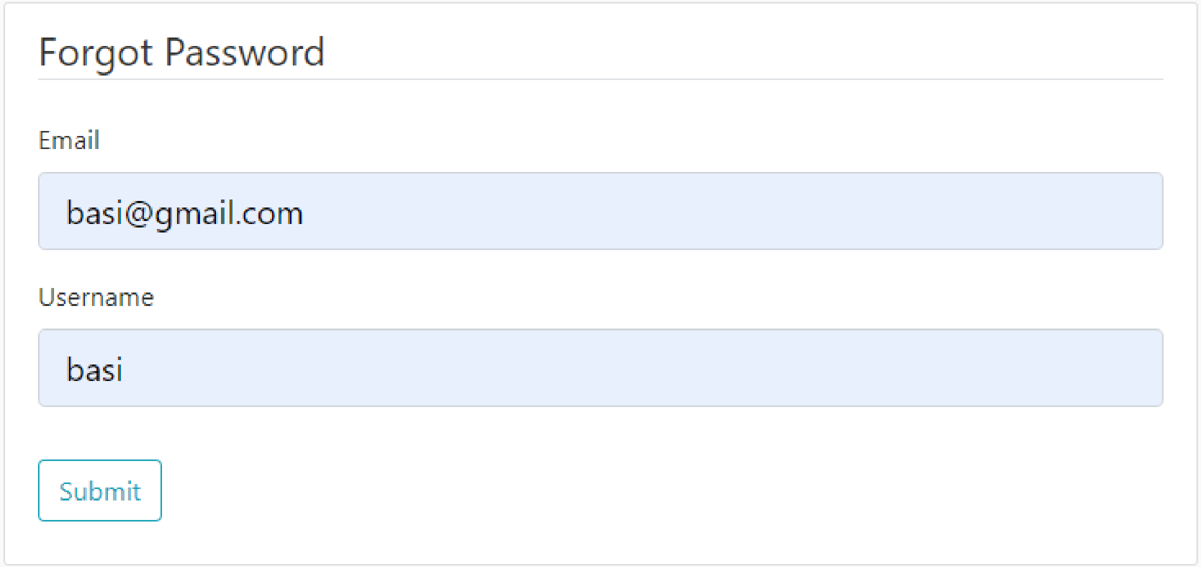
Admin

Password

••••••

□ Remember Me

FORGOT LOGIN



Forgot Password

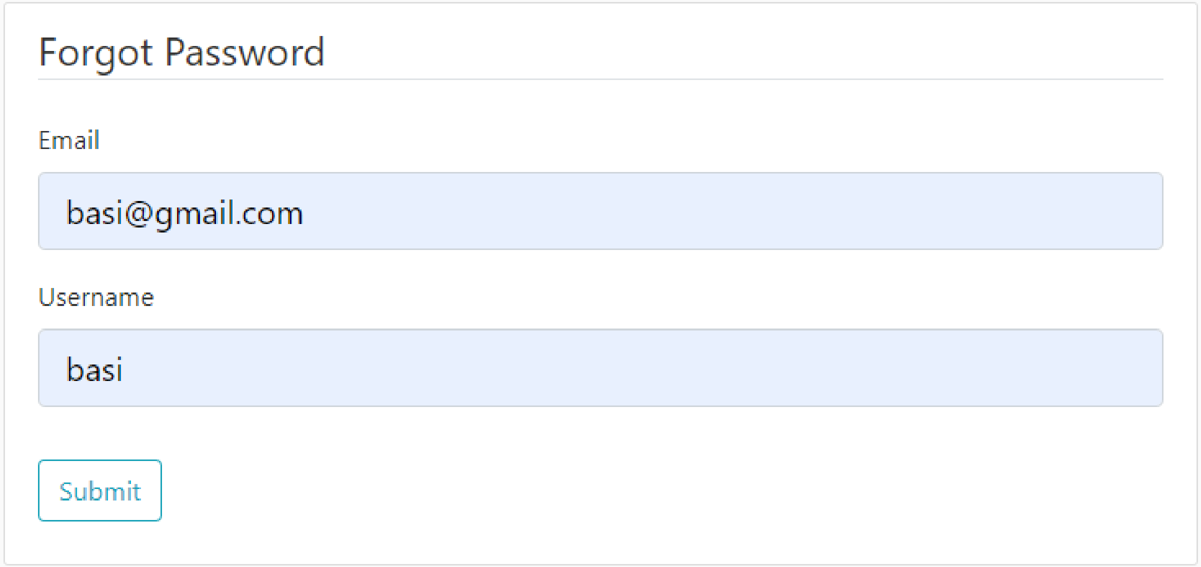
Email

basi@gmail.com

Username

basi

CHANGE FORM



Forgot Password

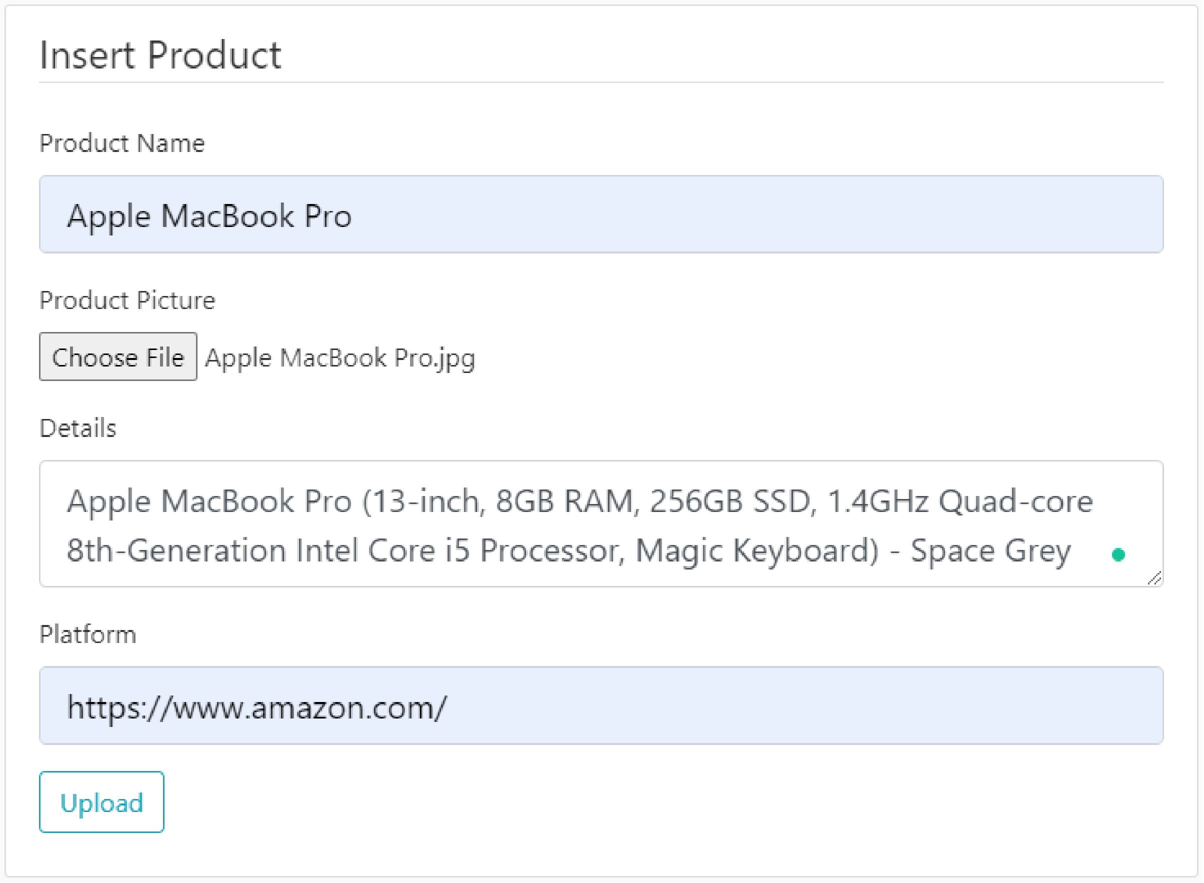
Email

basi@gmail.com

Username

basi

PRODUCT FORM



Insert Product

Product Name

Apple MacBook Pro

Product Picture

Choose File Apple MacBook Pro.jpg

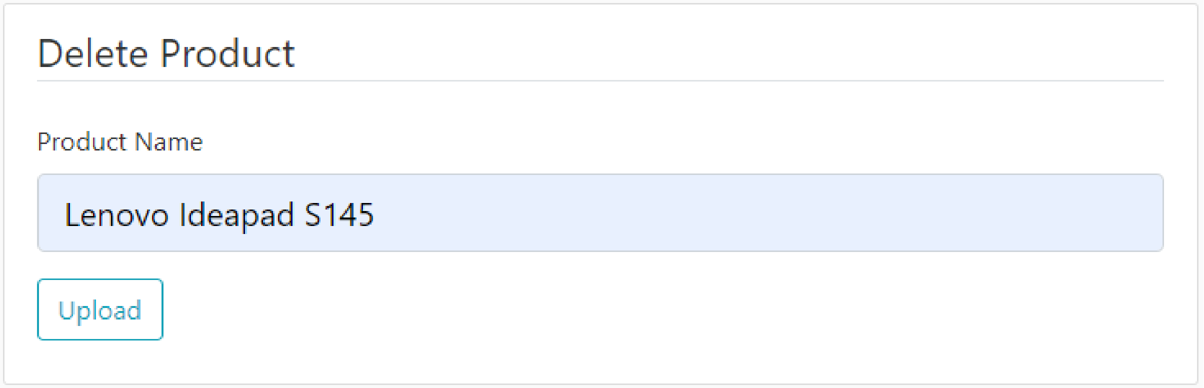
Details

Apple MacBook Pro (13-inch, 8GB RAM, 256GB SSD, 1.4GHz Quad-core 8th-Generation Intel Core i5 Processor, Magic Keyboard) - Space Grey

Platform

https://www.amazon.com/

DELETE FORM

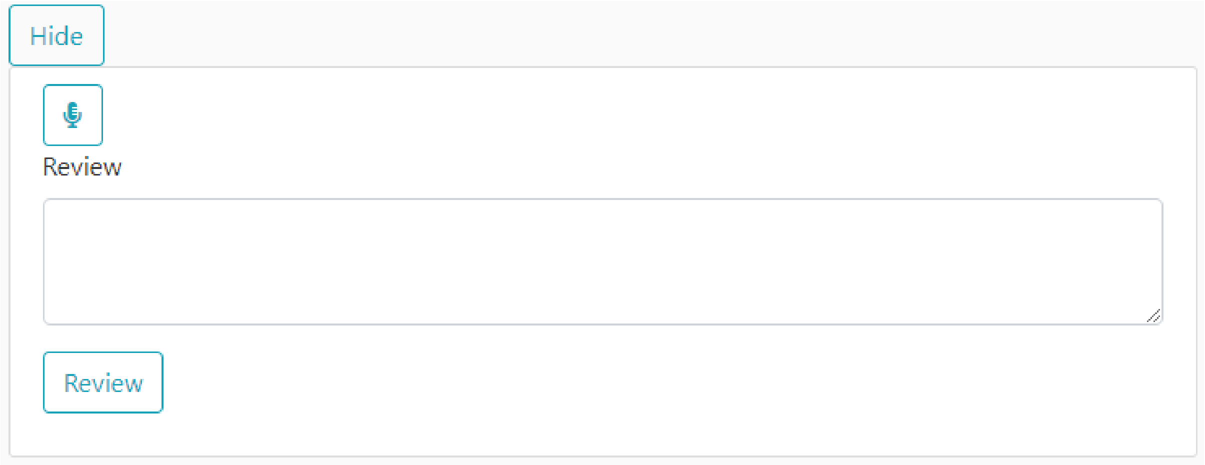


Delete Product

Product Name

Lenovo Ideapad S145

REVIEW FORM



5. TESTING

SL. No	PROCESS	EXPECTED OUTCOME	ACTUAL OUTCOME	REMARKS
1	Data Collection	To collect user review for a training dataset	Successfully Collected	Successfully collected user’s review.
2	Train Dataset	Receive a score of greater than 80%	Successfully received a score of above 80%	Successfully trained the Dataset
3	Creating GUI	Creating a user interface GUI	GUI created as expected	Successfully created GUI
4	Registration	A successful Registration page	Successfully Created a Registration Page	A successful Registration page
5	Login	A successful Login page	Successfully Created a Login Page	A successful Login page
6	Product Manger	Creating a unit to Manipulate Products	Successfully Developed a Unit to Manipulate Product	Successfully Created the Product Manager

7	Review Detection module	Detect reviews by the Trained Unit	Successfully detected reviews	Successfully Created Review Detection module
8	Ip Verification of Reviewer	Verify the reviewer Ip	Successfully Verified the reviewer Ip	Successfully developed Ip Verification of Reviewer

6. CONCLUSION

In this work, the proposed Fake Review Detection module to reduce the forging of false review in E-commerce Websites. The experiments have demonstrated the efficiency and effectiveness of the proposed automated method. The proposed module can validate Reviews to be fake or not in real-time and provides an attachment indicating its proficiency to the normal users on the basis of Four Clustering Algorithms which is also supplemented by a user IP address verifying unit. The proposed module can be easily integrated into any of the best e-commerce websites. And results indicated that the module achieved its goal to a affirmative phase. The project was successfully completed within the time span allotted. Every effort has been made to present the system in more useful manner.

7. FUTURE ENHANCEMENT

Though the aimed goal has been achieved, the result can still be considered inconsistent due to the presence of multiple variables that escalate fake review, the proposed idea for the future is to enable the module to capable of handling all the filter features of recognizing a review as fake or not such as

- Lexical features
- Content and style similarity
- Semantic inconsistency
- Public data available from Web sites
- Web site private/internal data

In an automated mechanism similar to an AI and a Dataset with a higher accuracy score.

8. REFERENCES

[1] Prof. Bing Liu Department of Computer Science University of Illinois at Chicago (UIC), “Detect Fake Reviews and Fake Reviewers - Detect Opinion Spam”.

[2] Assistant Professor. Naveed Hussain The University of Lahore, Lahore Pakistan Lahore, Punjab, Pakistan, “Amazon Product Review (Spam and Non Spam)”.

[3] R. He, J. McAuley. “Ups and downs: Modeling the visual evolution of fashion trends with one class collaborative filtering”. WWW, 2016.

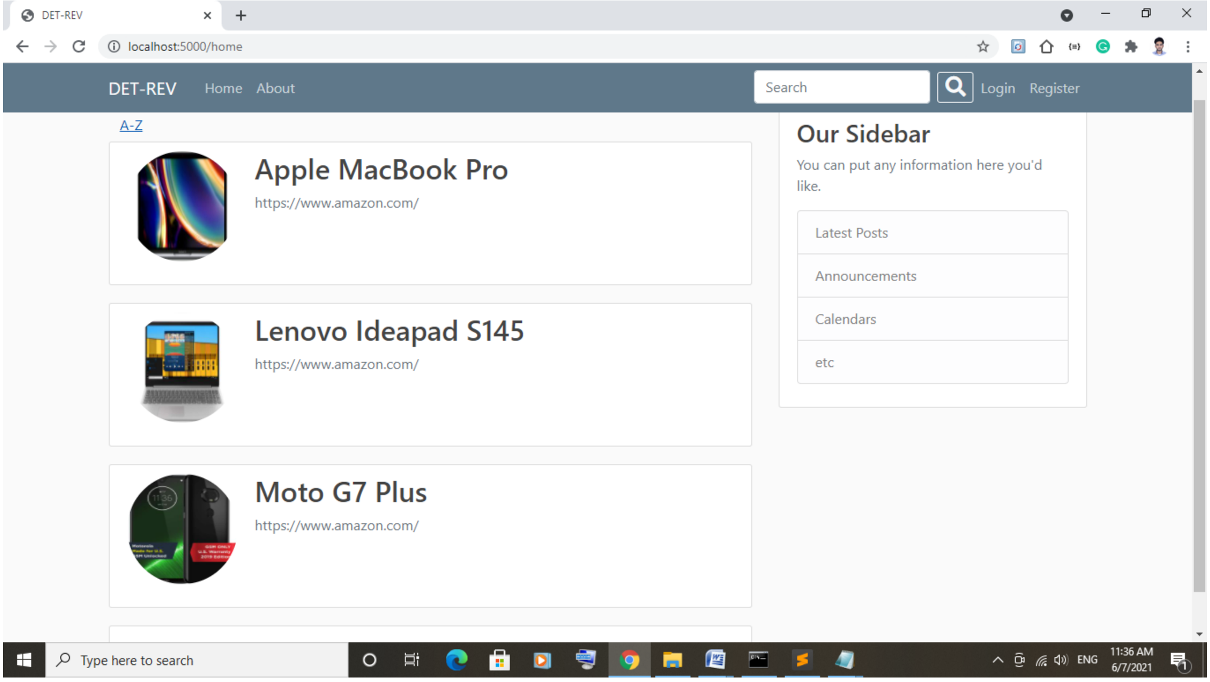
[4] J. McAuley, C. Targett, J. Shi, A. van den Hengel. “Imagebased recommendations on styles and substitutes”. SIGIR, 2015.

[5] Corey Schafer, “Flask Tutorials” https://youtube.com/playlist?list=PL-osiE80TeTs4UjLw5MM6OjgkjFeUxCYH/.

9.APPENDIX

9.1 SCREENSHOTS

HOME PAGE



Apple MacBook Pro

https://www.amazon.com/

Lenovo Ideapad S145

https://www.amazon.com/

Moto G7 Plus

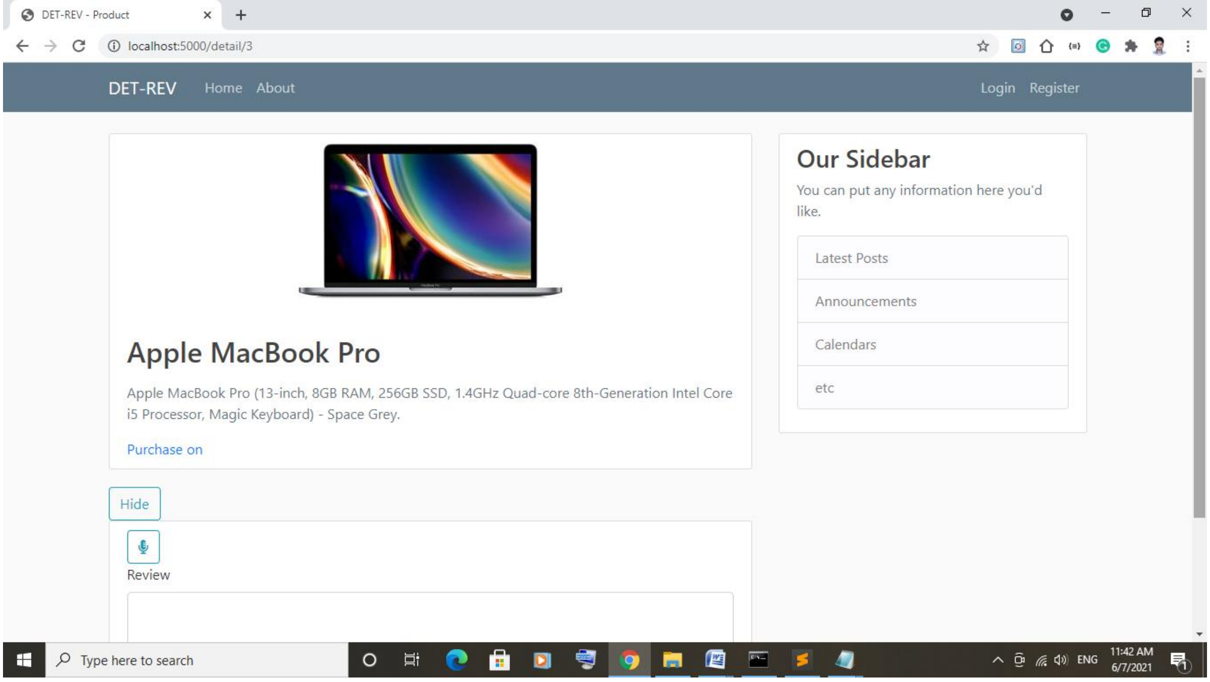
https://www.amazon.com/

Our Sidebar

You can put any information here you'd like.

- Latest Posts
- Announcements
- Calendars
- etc

DETAILS PAGE



Apple MacBook Pro

Apple MacBook Pro (13-inch, 8GB RAM, 256GB SSD, 1.4GHz Quad-core 8th-Generation Intel Core i5 Processor, Magic Keyboard) - Space Grey.

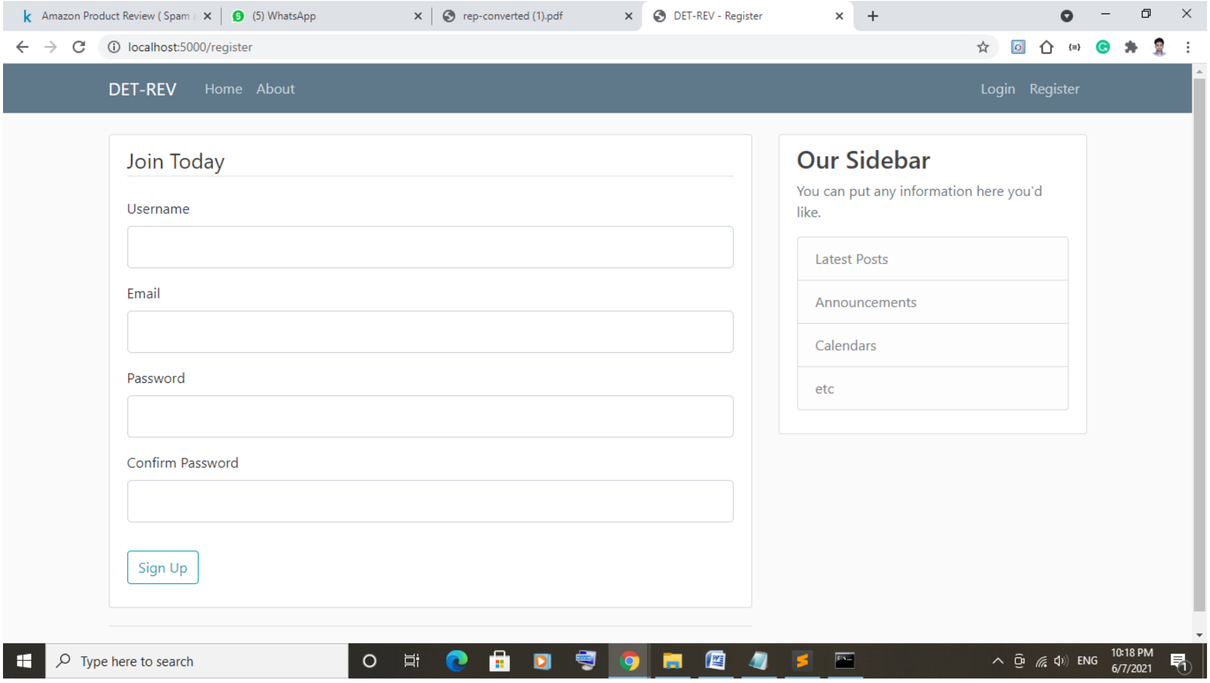
Purchase on

Our Sidebar

You can put any information here you'd like.

- Latest Posts
- Announcements
- Calendars
- etc

REGISTRATION PAGE



localhost:5000/register

DET-REV Home About Login Register

Join Today

Username

Email

Password

Confirm Password

Sign Up

Our Sidebar

You can put any information here you'd like.

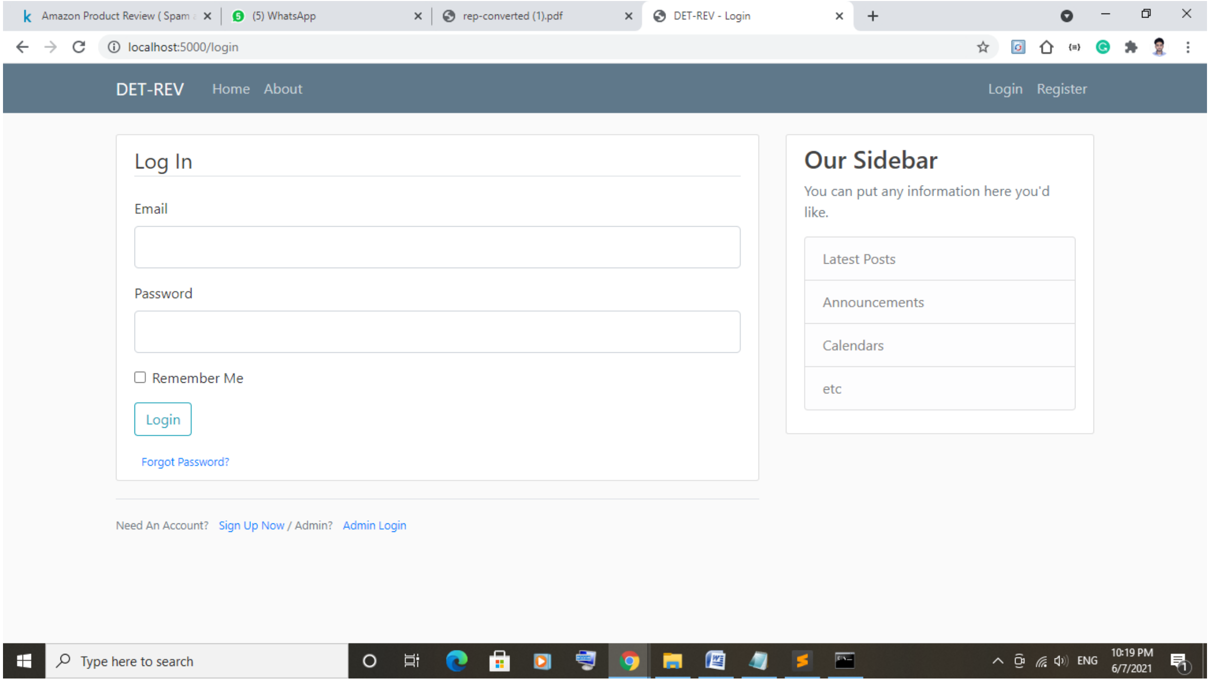
Latest Posts

Announcements

Calendars

etc

LOGIN PAGE



localhost:5000/login

DET-REV Home About Login Register

Log In

Email

Password

□ Remember Me

Login

Forgot Password?

Need An Account? Sign Up Now / Admin? Admin Login

Our Sidebar

You can put any information here you'd like.

Latest Posts

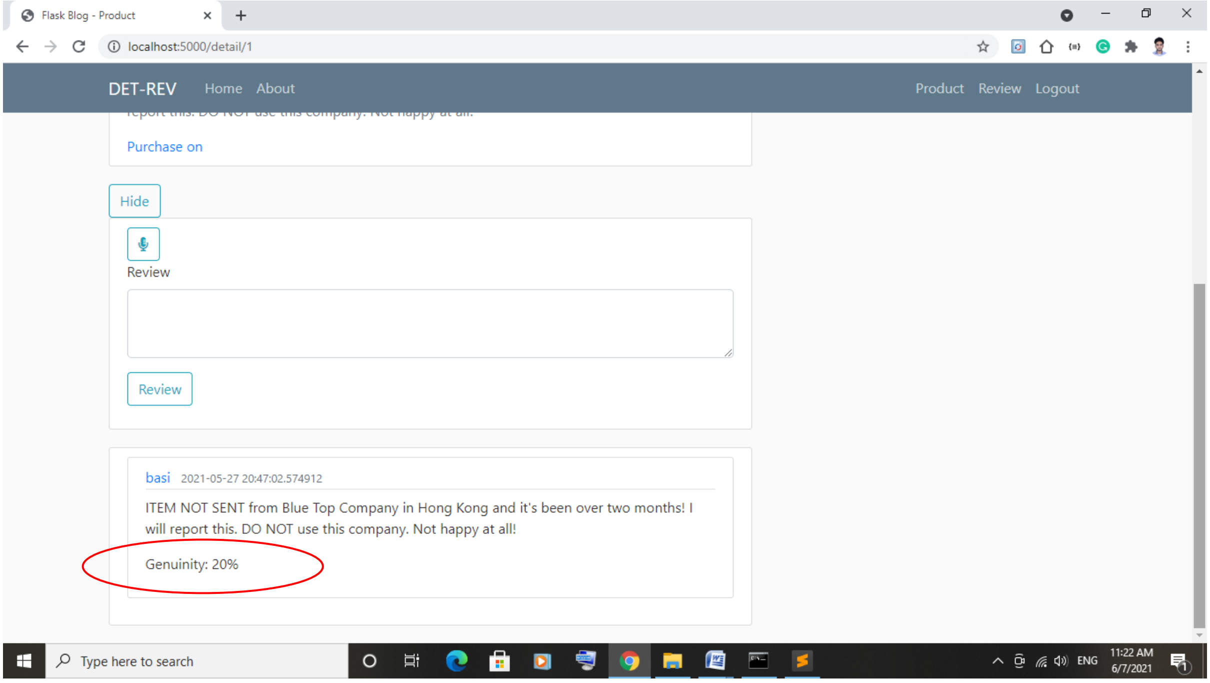
Announcements

Calendars

etc

DETAILS PAGE :

Review form and Scoring- A Genuity score is given to the review in percentage just below every review for user to see.



report this. DO NOT use this company. Not happy at all.

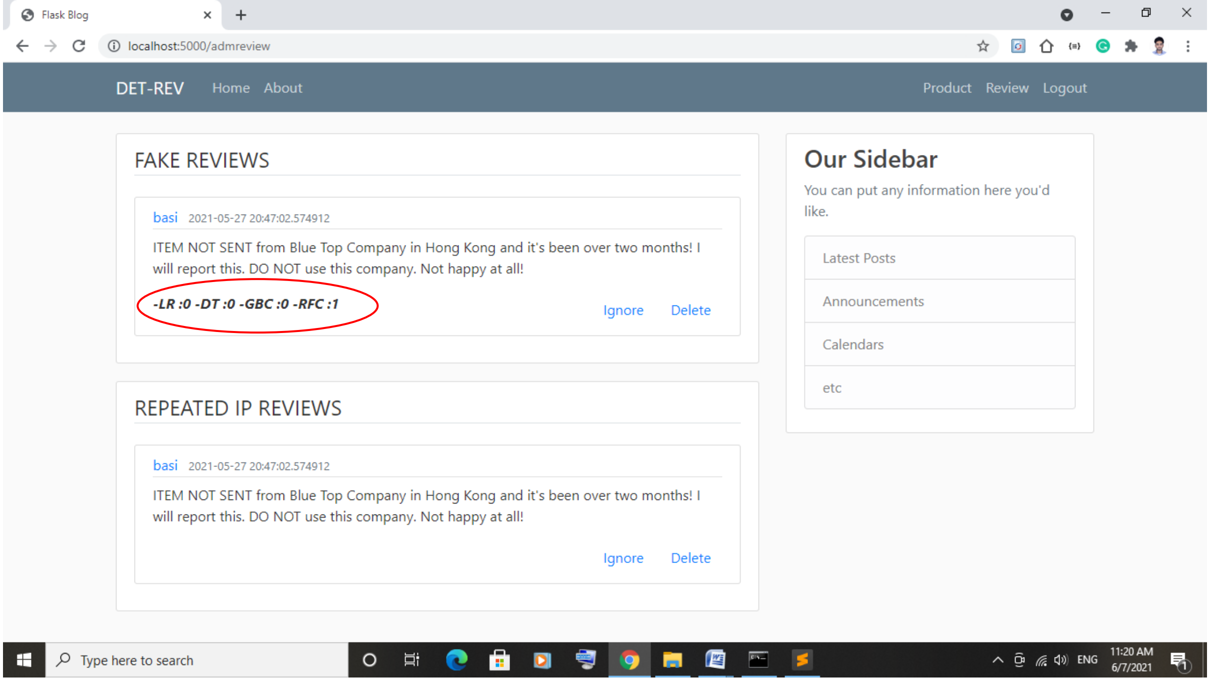
basi 2021-05-27 20:47:02.574912

ITEM NOT SENT from Blue Top Company in Hong Kong and it's been over two months! I will report this. DO NOT use this company. Not happy at all!

Genuinity: 20%

REVIEW MANGER PAGE

This page is only accessible to the admin and shows the result of each classifier given to a review when it is considered fake.



FAKE REVIEWS

basi 2021-05-27 20:47:02.574912

ITEM NOT SENT from Blue Top Company in Hong Kong and it's been over two months! I will report this. DO NOT use this company. Not happy at all!

-LR :0 -DT :0 -GBC :0 -RFC :1

REPEATED IP REVIEWS

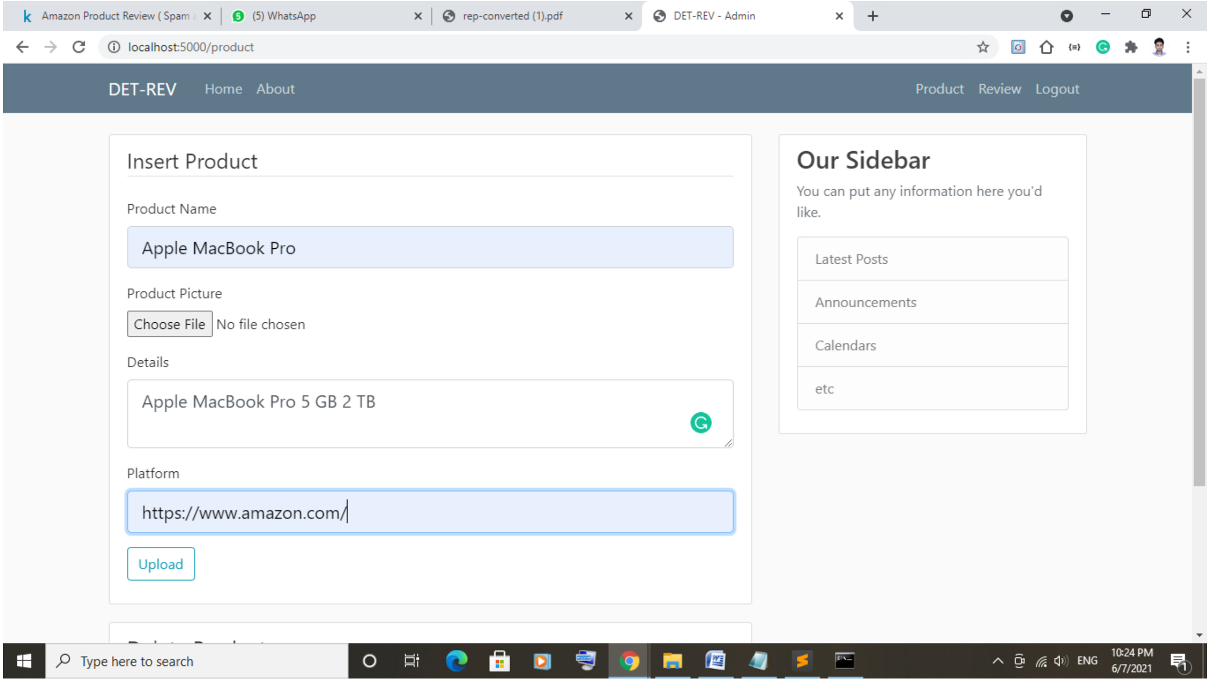
basi 2021-05-27 20:47:02.574912

ITEM NOT SENT from Blue Top Company in Hong Kong and it's been over two months! I will report this. DO NOT use this company. Not happy at all!

Our Sidebar

You can put any information here you'd like.

PRODUCT MANAGER



Insert Product

Product Name

Apple MacBook Pro

Product Picture

Choose File No file chosen

Details

Apple MacBook Pro 5 GB 2 TB

Platform

https://www.amazon.com/

Upload

Our Sidebar

You can put any information here you'd like.