Fake Product Review monitoring and removal for Genuine Ratings

Submitted By-
Komal Agarwal(2017CS03)
Divyanshi Dixit(2017CS08)
Parul Choudhary(2017CS16)

to the
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY, ALLAHABAD

Acknowledgements

During the construction of my project, We had to take the help and guidelines of some respected persons, who deserve our utmost and sincere gratitude. We would like to show our gratitude to Dr. Mayank pandey sir, MNNIT for giving us a good counsel and guidance for our project throughout its numerous consultations. We also want to thank Mr. Nitin Shukla and Mr. Shabir Ali for their valuable inputs throughout the making of this project.

We would also like to expand our deepest gratitude to all those who have directly or indirectly guided us in making this project a success.

Many people, especially our classmates and seniors themselves, have made many invaluable comment suggestions on this project which gave us the inspiration needed to improve our assignment. We thank all the support staff of our institute and the internet whose help directly and indirectly helped us complete this project.

ABSTRACT

In the era of online marketing e-commerce has become very famous. Users tend to buy more and more things online rather than offline shopping. Online reviews play an important role in the buying decision of the user. Our project is based upon the fact that how could we monitor the difference between fake and genuine product reviews so that user will be able to buy the good quality product. In today's online market, sometimes seller themselves write fake reviews about their products for better sale.

In our project, we have focused upon how to identify these fake reviews and remove them for better user experience.It is a web based application with both client and server support. Clients can register, login and purchase any product whereas at server side we can track our clients by their activity and monitoring the fake and genuine reviews.

***Keywords**:* Reviews, Fake, Sentiment Analysis, Online shopping

Table of Contents

1.Introduction

1.1 About the project

As most of the people require to review about a product before spending their money on the product. So people come across various reviews in the website but these reviews are genuine or fake is not identified by the user.

In some review websites some good reviews are added by the product company people itself in order to make in order to produce false positive product reviews. They give good reviews for many different products manufactured by their own firm. User will not be able to find out whether the review is genuine or fake.

To find out fake review in the website this "Fake Product Review Monitoring and Removal for Genuine Online Product Reviews Using Opinion Mining" system is introduced. This system will find out fake reviews made by posting fake comments about a product by identifying the IP address along with review posting patterns. User will login to the system using his user id and password and will view various products and will give his/her review about the product.

To find out whether the review is fake or genuine, system will trace the IP address of the user if the system observe fake review posted from the same IP Address on same product many a times it will inform the admin to remove that review from the system.This system helps the user to find out correct review of the product.

1.2 Project Features

Our website provides the following features-

1. Admin will able to login in to the website and mark reviews suspected by the system.
2. Admin will also be able to delete the fake reviews identified by our system.
3. A user can register, login,update profile,view his own reviews, buy and review the product.

4. Our system will track the information about IP address of the user's system and the pattern in which reviews are being posted.
5. If our system finds same types of reviews coming from a particular IP address then it will inform the admin and he will remove that fake review.
6. Our system is also capable of identifying suspected reviews by performing sentiment analysis on the review.

1.3 Project Scope

Our project can work with any e-commerce website who sells products and also maintains proper reviews for the products. Now anyone can write any opinion text or review, this can draw the individual's attention, and organizations to give undeserving spam opinions to promote or to discredit some target products. So there is a need to develop a smart system which automatically mine opinions and classify them into spam and non-spam category. Proposed system will automatically classify user opinions into spam or non-spam. This automatic system can be useful to business organization as well as to customers. Business organization can monitor their product sales by analyzing and understanding what the customers are saying about products. Customers can make decision whether he/she should buy or not buy the products. This can helpful to people to purchase valuable product and spend their money on quality products.

1.4 Technologies Used

1.4.1 PHP

PHP stands for Hypertext PreProcessor.It is used in web development.PHP is a server-side scripting language because PHP scripts are executed on server when clients sends a request, and the result are sent back to the browser. PHP supports different databases so compatibility is its strength.PHP web pages are similar to HTML web pages but the main difference is that PHP code is not disclosed to the browser as it is executed on server side only.

1.4.2 HTML/CSS

HTML stands for Hyper text Markup Language. It is the standard and the most basic language for create and structuring web pages. It is the most simple web structuring language. HTML elements are represented by tags,either a combination of starting and ending tags or a self closing tag.With HTML elements, many components like images, interactive forms can be embedded into a web page. CSS stands for Cascading Style Sheet. It describes how the html elements would be rendered on the web page. By using CSS, we can control the layout of multiple web pages at once.

1.4.3 MYSQL

MySQL is the most popular database system that is commonly used with PHP. Data is stored in the form of tables as in any other RDBMS.It is the database system for the websites with huge volumes of data.It is used for web and runs on the server side and can be compiled on the number of platforms.PHP combined with MySQL are cross platform i.e. we can develop in windows and then use in Linux.

1.4.4 XAMPP

XAMPP is and open source and free cross-platform web server solution stack package developed by Apache Friends, consisting of the Apache HTTP Server, MariaDB Database, interpreters for scripts written in PHP or perl. XAMPP stands for Cross-Platform(X),Apache(A), MariaDB(M), PHP(P), Perl(P).

1.4.5 PYTHON

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. It is popular in machine

learning because of many interrelated reasons. Python is simple, elegant, consistent, and math-like. Python code has been described as readable pseudocode.

Python is easy to pick up due to its consistent syntax and the way it mirrors human language and/or their mathematical counterparts. The latter (much due to libraries such as Numpy) is something one will appreciate if he were to implement a machine learning algorithm of which the core is likely just mathematical optimisation.

1.4.6 Sentiment Analysis

SENTIMENT ANALYSIS (sometimes known as OPINION MINING or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).

1.4.7 TEXTBLOB API

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

1.5 Goals

The software which will be developed should have the following capabilities:

1. It should be capable of distinguishing the authorized and unauthorized users.

2. System should be able to trace the IP from which a particular user has posted the review.

3. User gets genuine reviews about the product.

4. User can post their own review about the product.

5. If a particular review is contradictory to the majority of the reviews, then that could be a fake review and can be removed.

6. In this internet dominated world our goal is to stop the misuse of it,like the review is added in the system so that it will help the consumers to decide before buying the product and the can purchase the one they like.But fake review is like cheating to the so our goal is to make ecommerce cheat free sites.

7. Our goal is to make it genuine so that more and more people can take advantage of online shopping knowing that it is genuine,because in our country there are still more than half population who don't buy online.There is some trust issues as well so we want to make it genuine so that people should buy more and more online as it has advantages to both supplier and consumers.

2. Overall Description

2.1 Project Perspective

It is a web based website with both client and server support. Client can register, login and purchase any product whereas at server side we can track our clients by their activity and monitoring the fake and genuine reviews.

System works as follows:-

- User need to register first and then can login to the website.

- User once access the system, user can view product and can post review about the product.
- For admin the website will list all the fake reviews and Admin can delete the review which is fake.

2.2 Software Requirements

❖ **Server Side**

❑ XAMPP Package

❑ Windows 7/8/10 or Linux

❑ Any Latest Web Browser such as chrome or firefox

❖ **Client Side**

❑ Any Latest Web Browser

❑ Windows 7/8/10 or linux

2.3 Hardware requirements

❖ **Server Side**

❑ 1.75 GHz or Above Processor

❑ 32-bit / 64-bit Operating System

❑ x86 or x64 based processor

❑ RAM :2 GB or above

❖ **Client Side**

❑ x86 or x64 based processor

❑ RAM: 1 GB or above

2.4 User characteristics

1. First of all user need to register into our website.Then login.

2. Then user can view all the products along with its description,highlighted features,price and reviews.

3. User can buy the product by clicking buy now then it will redirect you to some famous online shopping websites.

4. User can review the product of your choice and be sure that the reviews you are writing is not fake.

5. A user can review on a particular product only once.

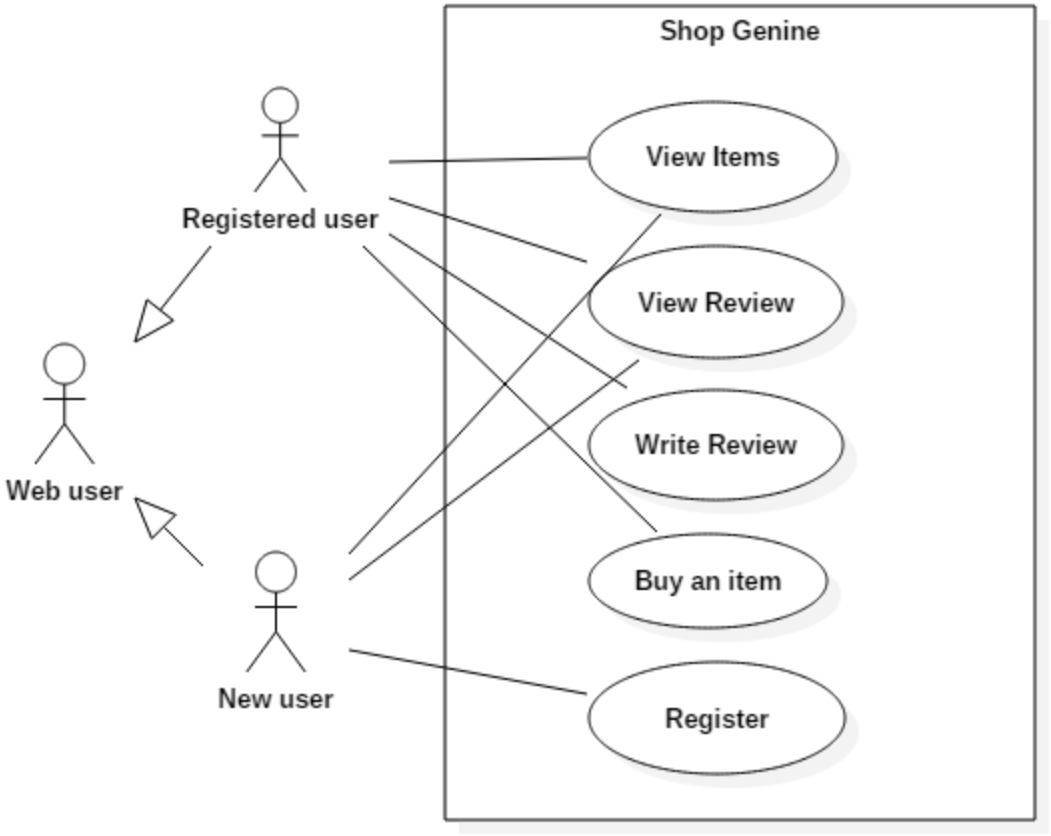
6. A profile is maintained for every user. User can update his/her profile or change password if he/she desires by clicking on 'Settings'.

7. User can view all the reviews given by him at various products at a single place by clicking on 'My Reviews' button.

2.5 Use Case Diagram

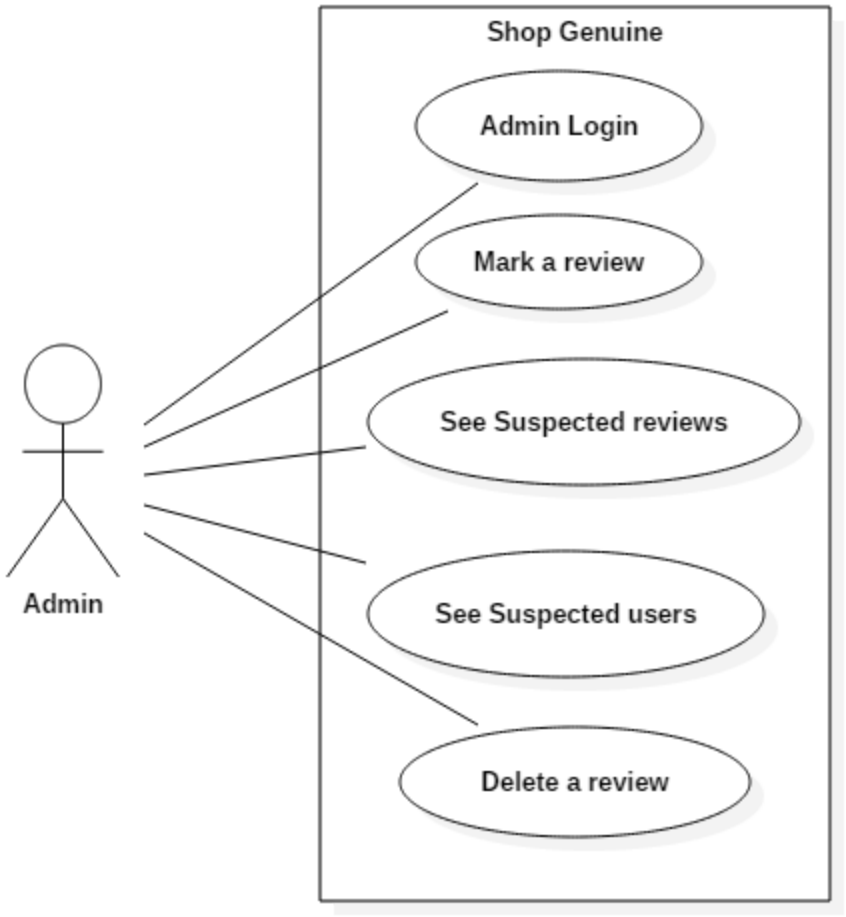
Two use case diagrams are implemented as

a) Top level Use case



Top Level Use case for Shop genuine

b) Admin use case



Use case for Admin

2.6 Class Diagram

Class diagram shows following classes

a) Web User This class is for all the users on the web. It has two operations - view any item and view reviews on those items.

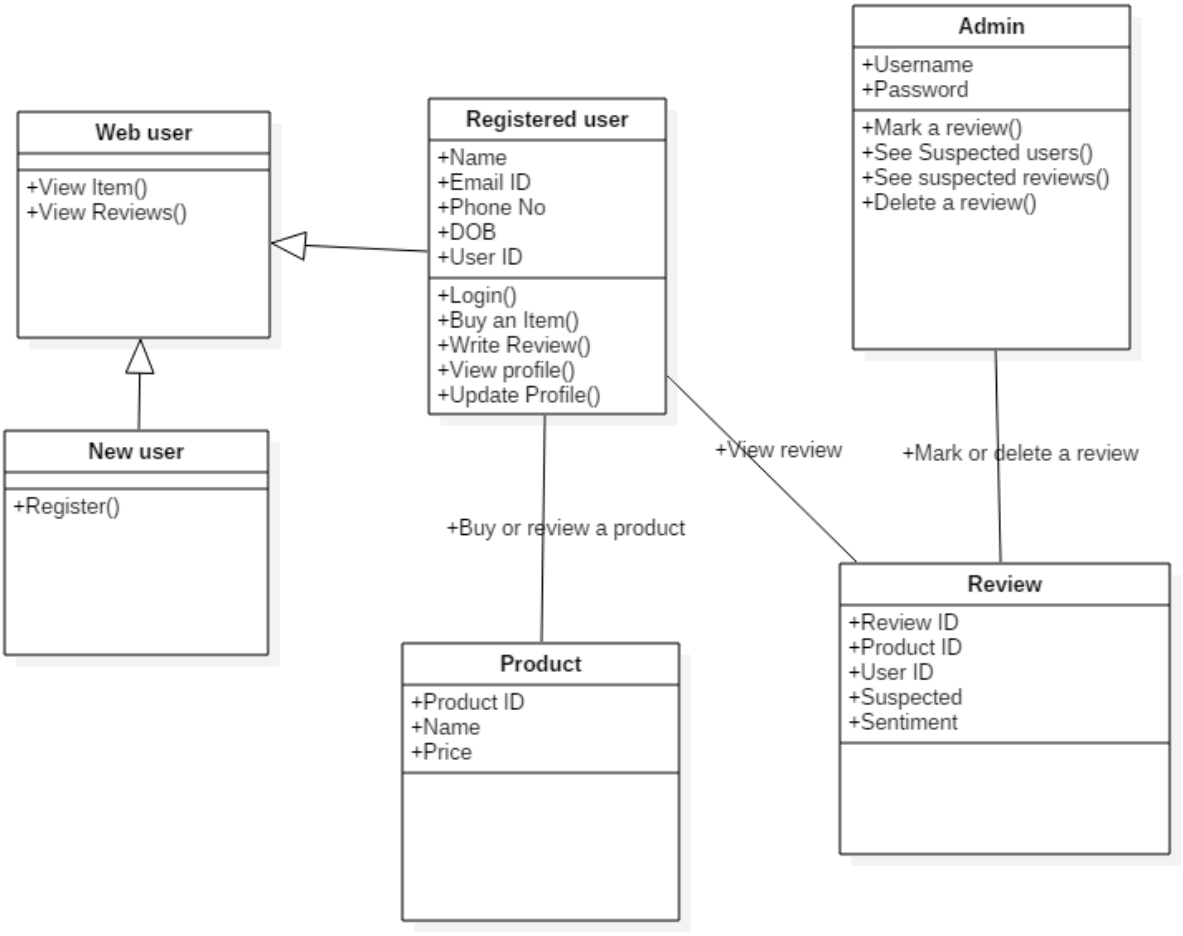
b) Registered user This class is inherited from the Web user class. It has many attributes as well as operations. Any registered user can also buy and review products and it is associated to review and product class.

c) New User This class is also inherited from the web user class and it has one operation - Register().

d) Admin In this class we have operations such as mar or delete a review. Admin class is associated with Review class.

e) Product This class has attributes such as product id, name and price. It is associated to registered user class.

f) Review class This class takes care of all the reviews of our website with attributes such as review id, sentiment and suspected value (0 or 1).



Class Diagram for Our Website

2.7 Functionality of our system

We have used two techniques to track the fake reviews-

A) Detection through duplicate IP address-

If we are getting reviews on a single product from different users but from the same IP address then after a threshold such as 5, we would declare those reviews fake and inform the admin. We have used PHP code at back end in the "mark.php" file where we could identify IP address of the users(user may be behind a proxy server).

As soon as admin click on the "**Mark Review**" Button, the PHP code will run at back end and all the reviews will have 1 in the "suspected" field of the database. This "suspected" field is 0 by default.

B) Detection through Sentiment Analysis

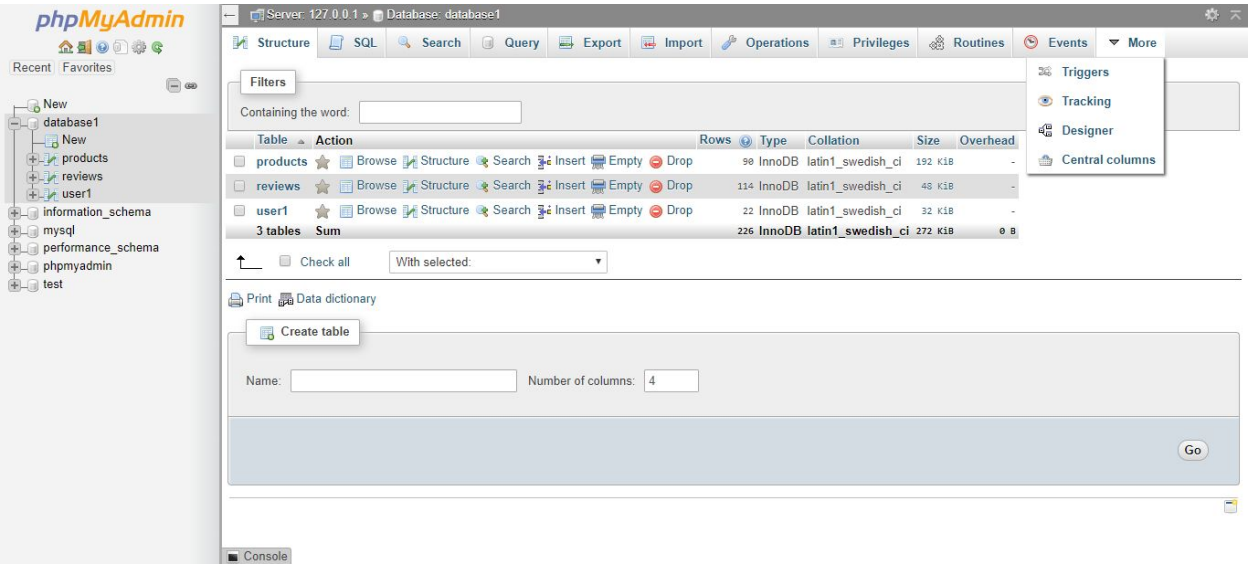
Fake users can be smart and can hide their identity by using some application which allows them to have different IP address so here we could take help of sentiment analysis and get to know the intent behind that review. Here we have set a threshold polarity value(inclination towards positive or negative) for every product(Which is the average of polarity value/sentiment of all reviews for that product). If we get a review whose confidence level varies much from the threshold polarity value of the product,then we would identify that review as fake.

B.1) Python code for sentiment detection

```
#!C:/Program Files/Python36/python.exe
from textblob import TextBlob
sent={}
rid={}
f=open('rev.txt','r')
data=f.readlines()
i=0
for line in data:
        i=i+1
        l=line.split(',',1)
        li=list(l)
        s=TextBlob(li[1])
        r=s.sentiment
        sent[i]=r[0]
        polarity[i]=r[1]
        rid[i]=li[0]
        with open("Output.txt", "a+") as f1:
                print("{} {}".format(rid[i],sent[i]),file=f1)
```

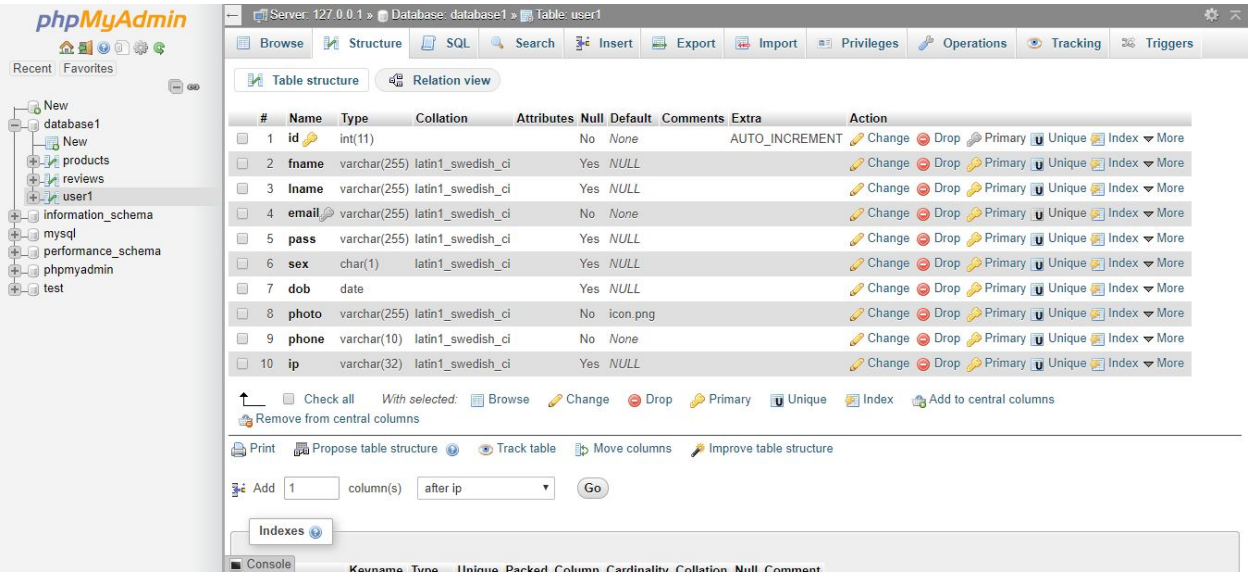
2.8 Database

We have one database named “database1” for our website. Our database has three tables-



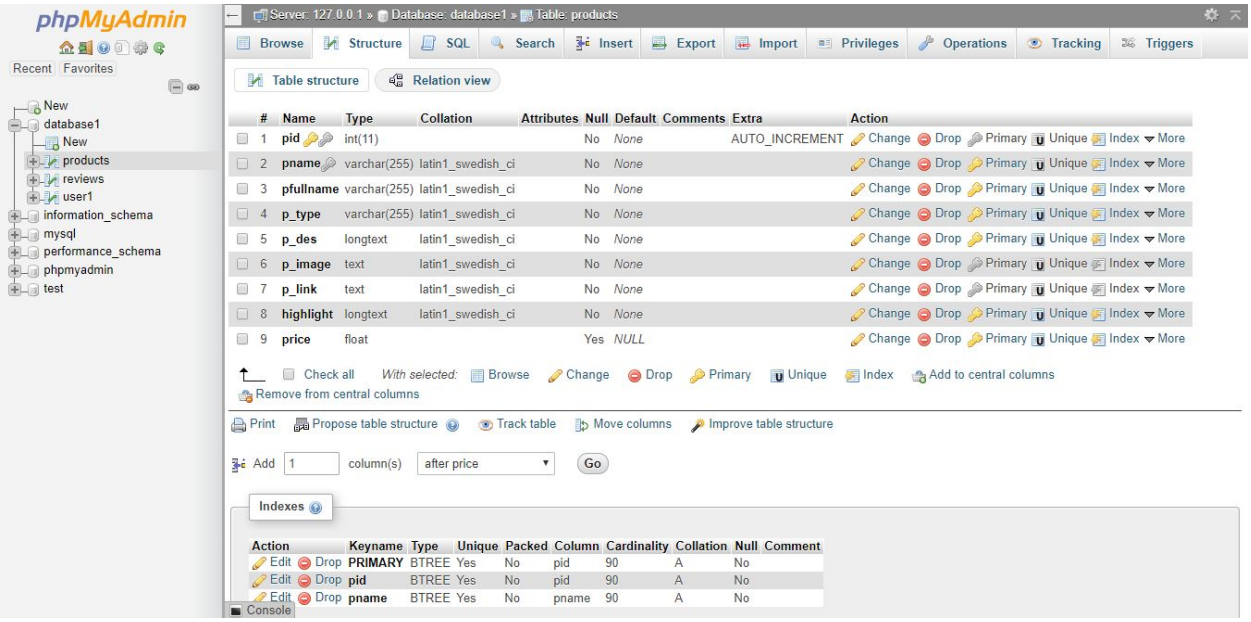
A) User1

This table stores all the information related to users. It has 10 columns and ID is the primary key for this table which auto increments and a new id is assigned to a new user automatically.



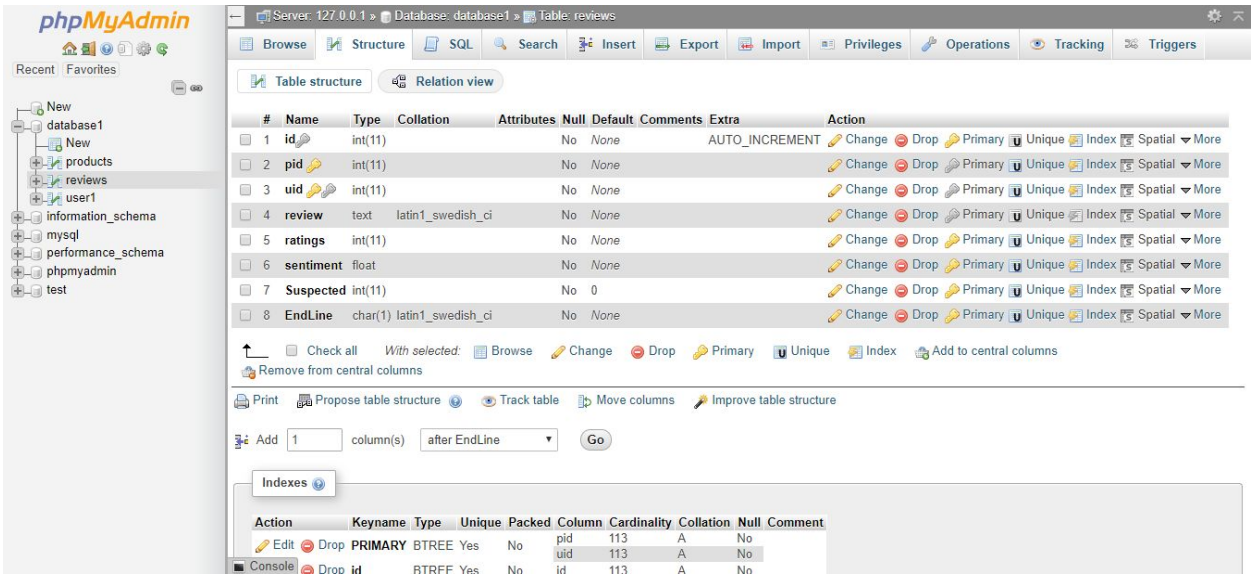
B) Products

This table contains all the data related to products in our website. We have used only three product categories as of now- Mobile phones, laptops and tablets. We may add more categories as needed. “Pid” is used to identify each row uniquely. We have 9 columns in our table.



C) Reviews

This is the most important table for our project as we store all product reviews given by users and pid uid as a combined primary key for uniquely identify each review. This table has 8 columns including “sentiment” which stores the sentiment of each review and suspected which has a value of either 0 or 1 indicating genuine or fake review.


phpMyAdmin
Recent Favorites
New
database1
New
products
reviews
user1
information_schema
mysql
performance_schema
phpmyadmin
test
Server: 127.0.0.1 » Database: database1 » Table: reviews
Browse Structure SQL Search Insert Export Import Privileges Operations Tracking Triggers
Table structure Relation view
Name Type Collation Attributes Null Default Comments Extra Action
1 id int(11) No None AUTO_INCREMENT Change Drop Primary Unique Index Spatial More
2 pid int(11) No None Change Drop Primary Unique Index Spatial More
3 uid int(11) No None Change Drop Primary Unique Index Spatial More
4 review text latin1_swedish_ci No None Change Drop Primary Unique Index Spatial More
5 ratings int(11) No None Change Drop Primary Unique Index Spatial More
6 sentiment float No None Change Drop Primary Unique Index Spatial More
7 Suspected int(11) No 0 Change Drop Primary Unique Index Spatial More
8 EndLine char(1) latin1_swedish_ci No None Change Drop Primary Unique Index Spatial More
Check all With selected: Browse Change Drop Primary Unique Index Add to central columns
Remove from central columns
Print Propose table structure Track table Move columns Improve table structure
Add 1 column(s) after EndLine Go
Indexes
Action Keyname Type Unique Packed Column Cardinality Collation Null Comment
Edit Drop PRIMARY BTREE Yes No pid 113 A No
uid 113 A No
Console
Drop id BTREE Yes No id 113 A No

2.9 Screenshots

A) Web user

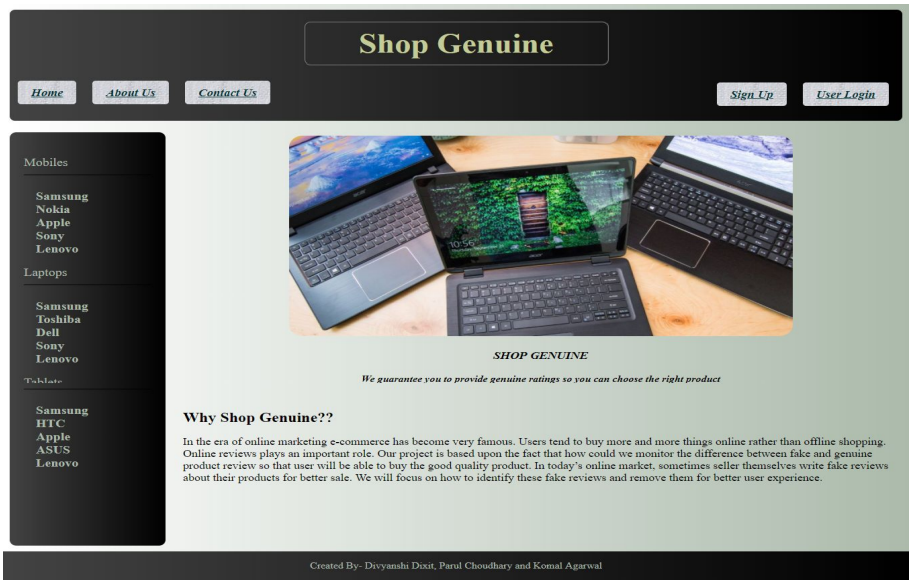


Fig.1 Home Page

Fig.2 About Us

Fig.3 Mobiles

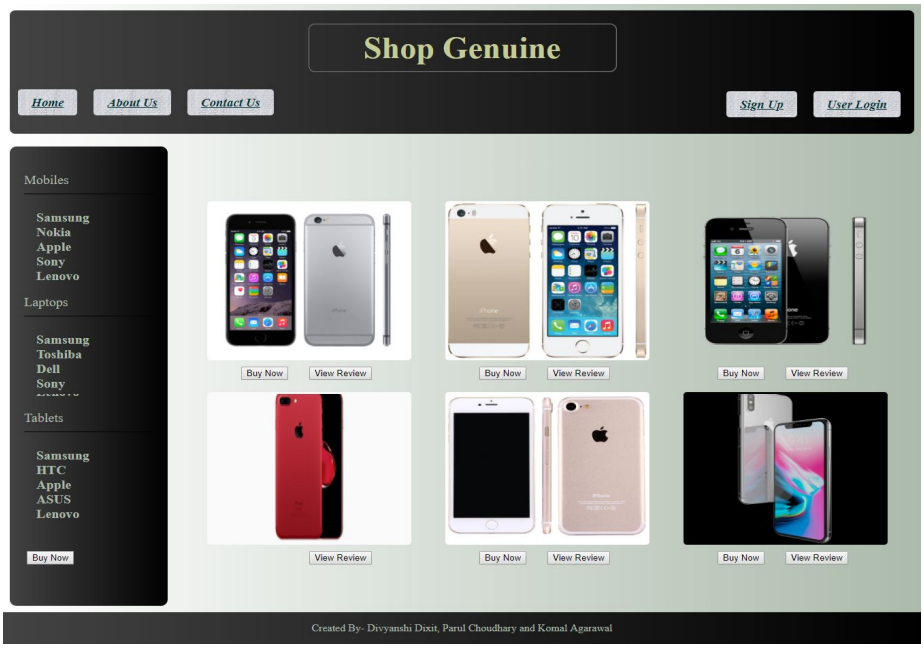


Fig.4 Apple Mobiles

Fig. 5 Laptops

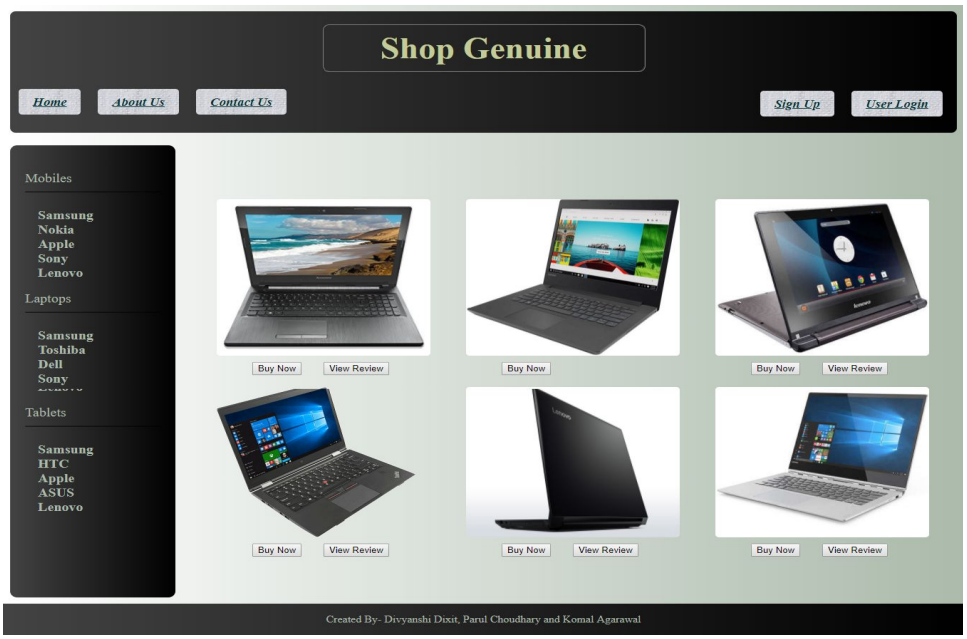


Fig. 6 Laptops Lenovo

Fig. 7 Tablets

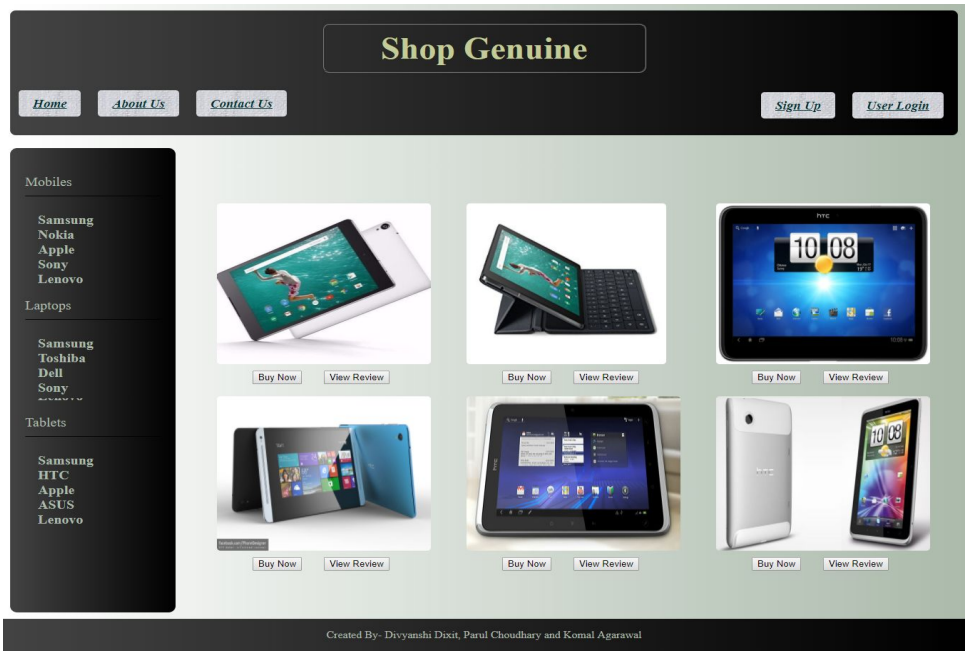


Fig. 8 Tablet HTC

B) New User

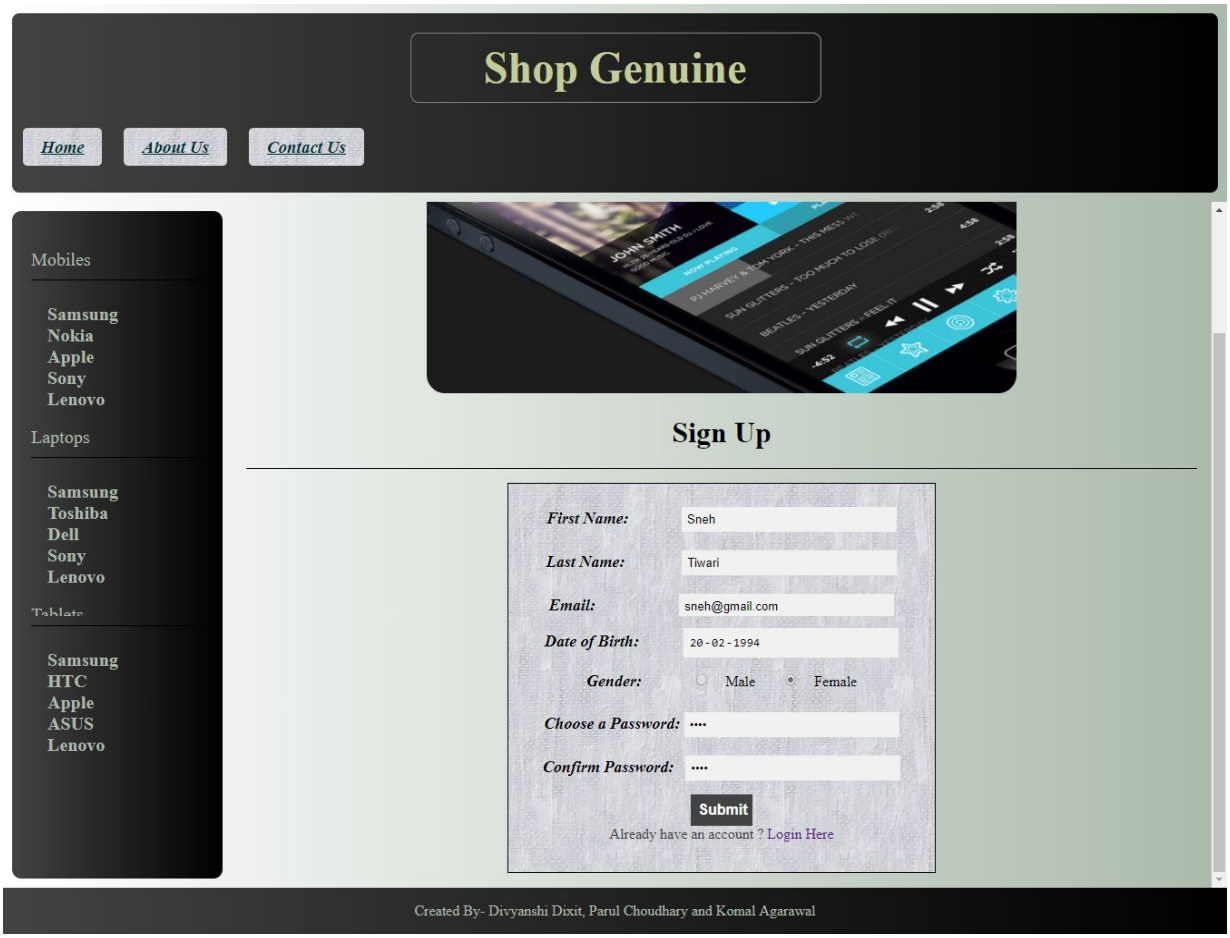


Fig . 9 New User Registration

C) Registered User

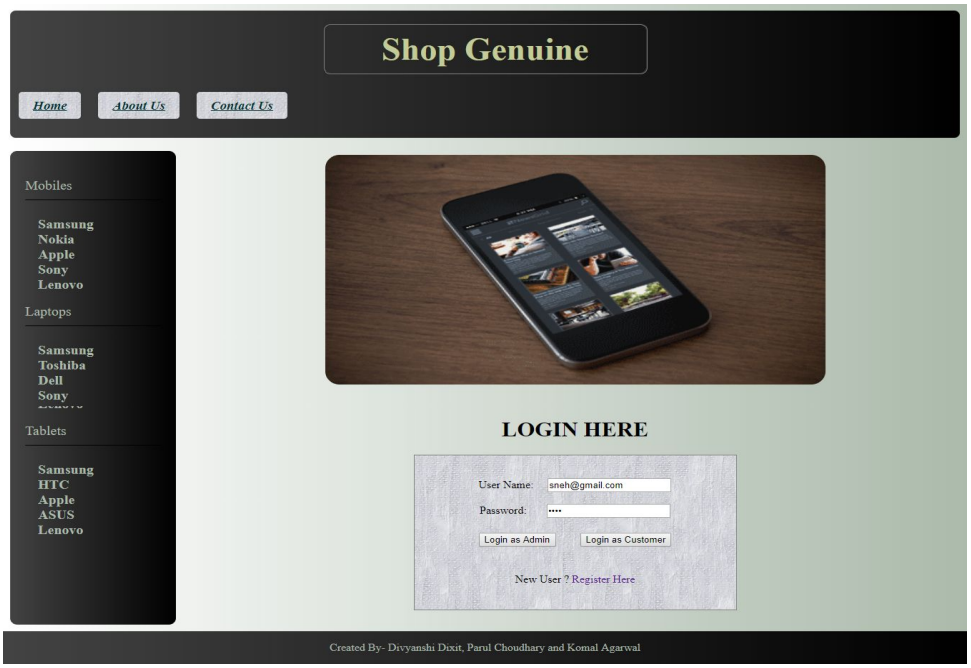


Fig. 10 Registered User Login

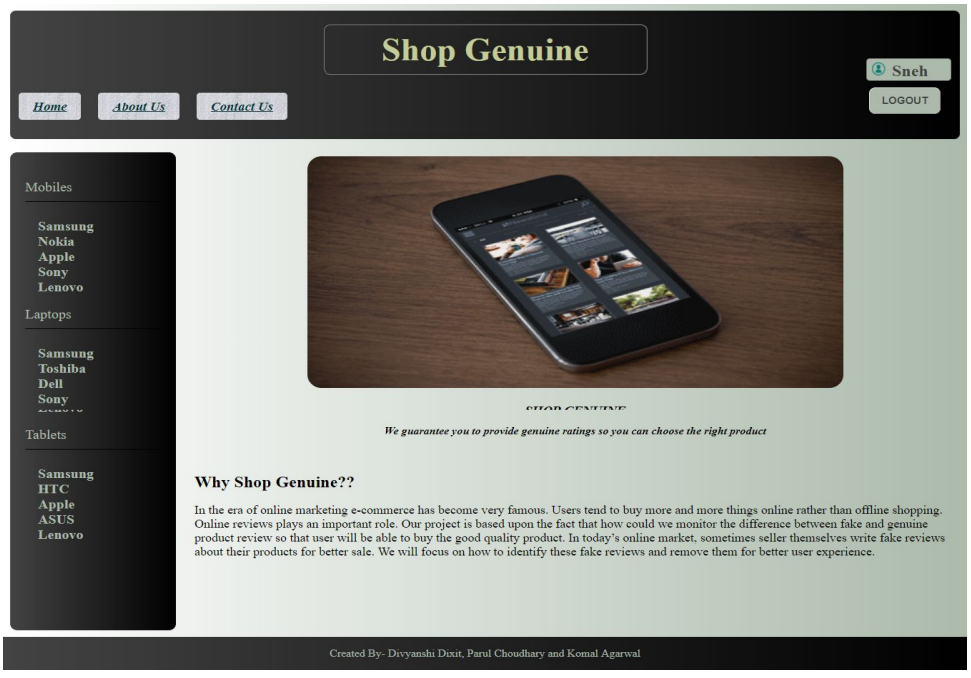


Fig. 11 Redirect to Home page

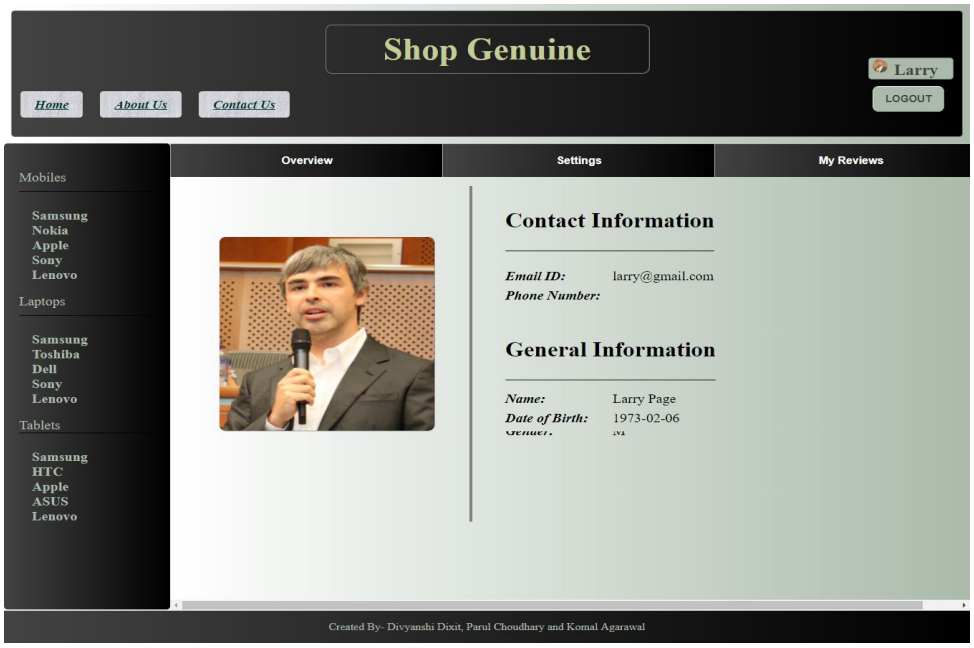


Fig. 12 User Profile

Fig. 12 User Profile Update

Fig. 13 User Posted Reviews

Fig. 14 Viewing Product Details

Fig. 15 Viewing a Products reviews

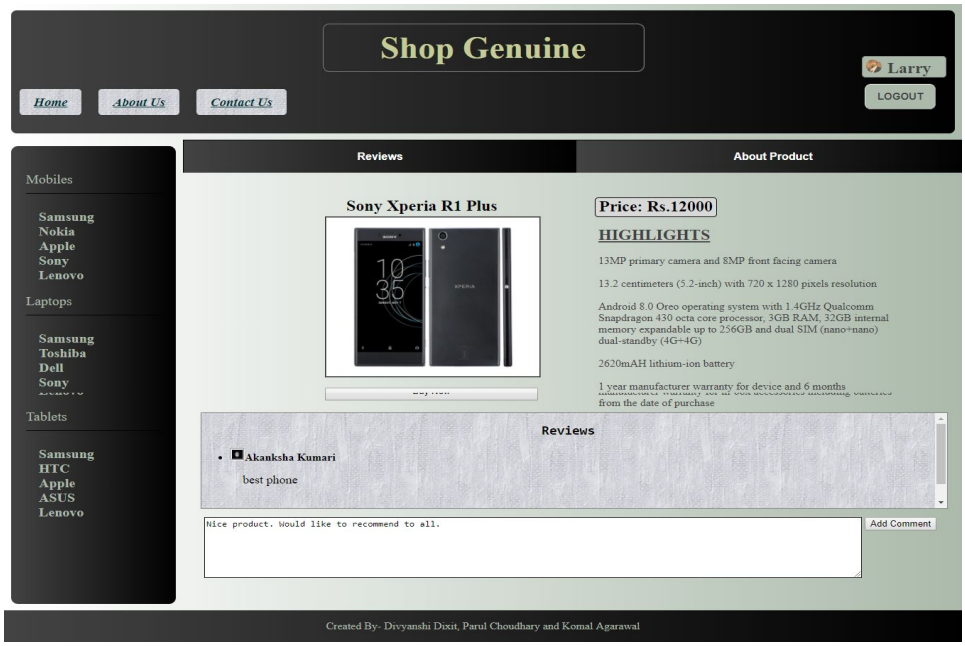


Fig. 16 Writing a Review

Fig. 17 Review Posted

D) Admin Panel

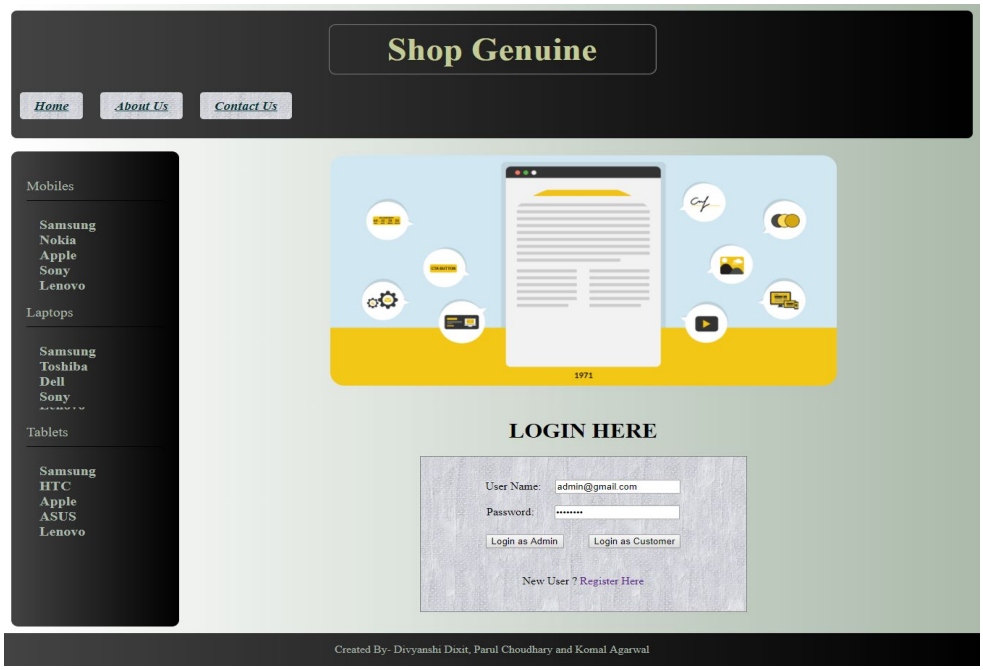


Fig. 18 Admin Login

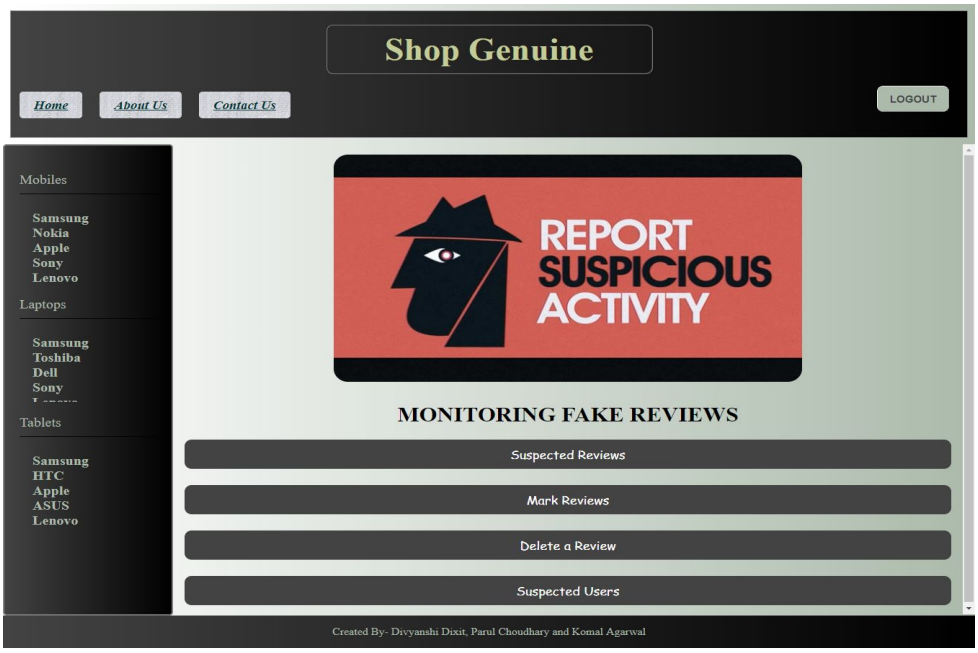


Fig. 19 Admin Panel

Fig. 20 List of Suspected Reviews

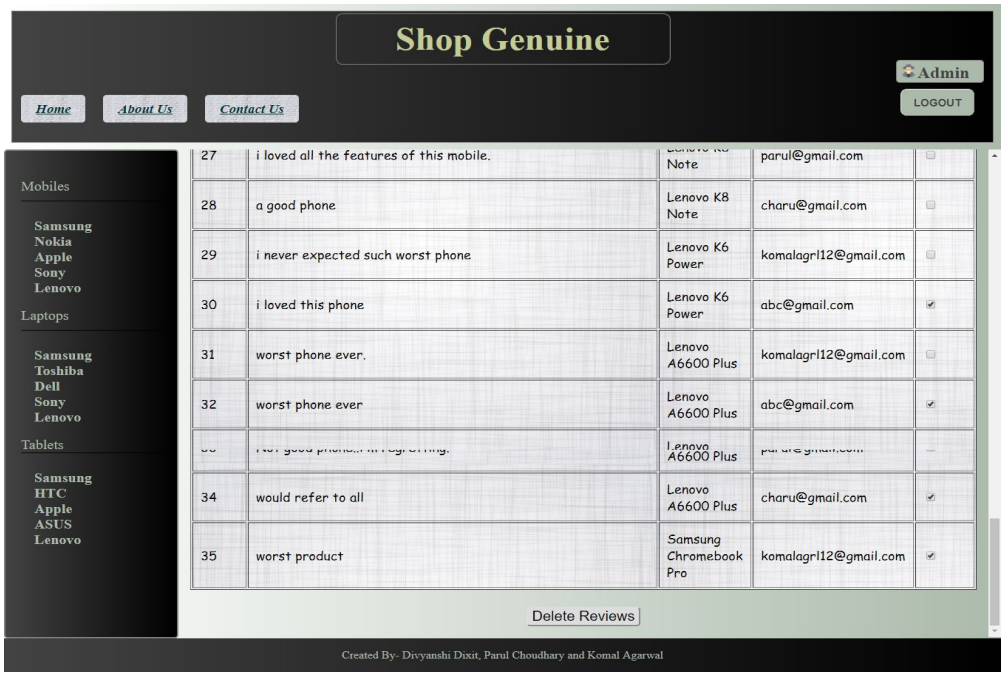


Fig. 21 Deleting a Review

Fig. 22 List of Suspected Users

2.10 Constraints

Different hardware, software and memory constraints would be there in our website.

❖ Our website can work with any system that supports internet and can browse the web pages.

❖ As we are using Xampp for server connection and database requirement, we would not be able to handle large amount of users simultaneously.

❖ System should be fault tolerant and should be able to recover easily.

❖ This is a sample website, so number of products added would be just sufficient for our fake product review monitoring because that is our major objective.

3. Summary

3.1 Conclusion

To find the opinion spam from unstructured data is a big problem. And its solution is very important. It is one of the most talked issue nowadays so specialist, academics and various organizations are coming forward to solve this issue. More research and future work is needed to solve this completely so the consumers may buy the products which are actually better than the product which has fake good reviews.

It also becomes a problem when any good product is given bad reviews and people don't buy them after reading the reviews.

This proposed idea will work on analyzing the spam from various reviews and putting the reviews in the fake review category. In future we will do more contributions in this area on other different types of features related to reviews.Finding the opinion spam from huge amount of unstructured data has become an important research problem. Now business organizations,

specialists and academics are putting forward their efforts and ideas to find the best system for opinion spam analysis. Although, some of the algorithms have been used in opinion spam analysis gives good results, but still no algorithm can resolve all the challenges and difficulties faced by today's generation.

More future work and knowledge is needed on further improving the performance of the opinion spam analysis. There is a huge need in the industry, in day-to-day life for such applications because every company wants to know how consumers really feel about their products and services and those of their competitors by analyzing true reviews not spam reviews. This project proposes an opinion spam analyzer which automatically classifies input text data into either spam or non-spam category.

The proposed system will use machine learning technique. A direction for future research is to implement the system and check performance by applying proposed approach to various benchmark data sets.

3.2 Long term objective

As this is just a basic version of our website, we can enhance it in future by adding more good features such as-

- For making this review process more feasible, we will allow only authenticated buyers to write review for the product.
- We can collaborate with many e-commerce website and help them to monitor the fake products reviews in their websites.
- By monitoring reviews we can recommend our customers whether they should buy a particular product or not.
- We can develop a mobile friendly equivalent Android Application for the website.

4. References

1) Fake Product Review Monitoring and Removal for Genuine Online Product Reviews Using Opinion Mining by Manleen Kaur Kohli, Shaheen Jamil Khan, Tanvi Mirashi, Suraj Gupta,Volume 7, Issue 1, january,2017

http://ijarcsse.com/Before_August_2017/docs/papers/Volume_7/1_January2017/V7I1-0150.pdf

2) Detecting Review Spam: Challenges and Opportunities,Yingying Ma and Fengjun Li ,2012
http://cae.ittc.ku.edu/papers/li-reviewspam.pdf

3)Jindal, N., & Liu, B. 2008. Opinion Spam and Analysis. Proceedings WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining,219-230.
https://www.cs.uic.edu/~liub/FBS/opinion-spam-WSDM-08.pdf

4)Liu, B., 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers
https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf