International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 08, August 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

# FAKE WEBSITE DETECTION USING REGRESSION

**G kumari[1],M Naveen kumar[2],A Mary Sowjanya[3]**

[1,3]*Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam.*

[2]*Department of Computer Science and Engineering, TKRCET, Hyderabad.*

## ABSTRACT

*There are number of users who purchase products online and make payment through various websites .There are multiple websites who ask the user to provide sensitive data such as username, password or credit card details etc. often for authentication. But there exist some phishing websites. Which use that information for malicious reasons. In order to detect and predict phishing websites, we implemented a flexible and effective system that is based on data mining algorithm. We implemented Logistic Regression algorithm and techniques to classify their legitimacy. The phishing website can be detected based on some important characteristics like URL, domain identity, and security in the final phishing detection rate.When the user wants to check whether the website is legitimate or not, our system uses data mining algorithm to check its legitimacy. This application can be used by many internet users in order to save themselves from an ocean of phishing sites. The data mining algorithm used in this system provides better performance. With the help of this system user can also purchase products online without any hesitation. Adminscan also add phishing website URL's or fake website URL's into system where system could access and scan the phishing websites. New suspicious URL's can be added when a user submits it.*

***Keywords: Phishing, Phishing characteristics, Regression, Data mining.***

## I. INTRODUCTION

### 1.1 Phishing

Phishing is an attempt to obtain sensitive information such as usernames, passwords, and credit card details (and indirectly money) often for malicious reasons, by disguising as a trustworthy entity in an electronic communication. This word is a neologism created as a homophone of fishing due to the similarity of using bait in an attempt to catch a victim. Phishing is typically carried out by email spoofing or instant messaging, and it often directs users to enter personal information at a fake website, the look and feel of which are almost identical to the legitimate one. Communications purporting to be from social web sites, auction sites, banks, online payment processors or IT administrators are often used to lure victims. Phishing emails may contain links to websites that are infected with malware. Phishing is an example of social engineering techniques used to deceive users, and exploits weaknesses in current web security. Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures. Many websites have now created secondary tools for applications, like maps for games, but they should be clearly marked as to who wrote them, and users should not use the same passwords anywhere on the internet.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

### 1.1.1 Link manipulation

Most methods of phishing use some form of technical deception designed to make a link in an email (and the spoofed website it leads to) appear to belong to the spoofed organization. Misspelled URLs or the use of sub domainsare the common tricks used by phishers. In the following example URLhttp://www.yourbank.example.com/, it appears as though the URL will take you to the example section of the bank website; Actually this URL points to the "your bank" (i.e. phishing) section of the example website. Many email clients or web browsers will show previews of where a link will take the user in the bottom left of the screen, while hovering the mouse cursor over a link. This behavior, however, may in some circumstances be overridden by the phisher.

A further problem with URLs has been found in the handling of internationalized domain names (IDN) in web browsers that might allow visually identical web addresses to lead to different, possibly malicious, websites. Despite the publicity surrounding the flaw, known as IDN spoofing or homograph attack, phishers have taken advantage of a similar risk, using open URL redirectors on the websites of trusted organizations to disguise malicious URLs with a trusted domain. Even digital certificates do not solve this problem because it is quite possible for a phisher to purchase a valid certificate and subsequently change content to spoof a genuine website, or, to host the phish site without SSL at all.

## II. RELATED WORK

### 2.1 Python

Python is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and was first released in 1991.An Interpreted Language, Python has a design philosophy which emphasizes code readability and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale.Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming and procedural styles. It has large and comprehensive standard library.Python interpreters are available for many operating systems, allowing python code to run on a wide variety of systems.

### 2.1.1 Features:

Python is a multi-paradigm programming language. Python uses dynamic typing and a mix of reference counting and a cycle-detecting garbage collector for memory management. An important feature of Python is dynamic name resolution, which binds methods and variable names during execution.Universally, Python has gained a reputation because of its easy to learn. Python has significant popularity in Scientific Computing.Nowdays we are working on bulk amount of data, popularly known as big data. The more data you have to process, the more important it becomes to manage the memory you use. The Python works efficiently in this case. Python is an interpreted language, but some developments has been done over past years to improve Python's performance. For high Performance Python is a viable best option today.

### 2.2 Python packages for Data Mining:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

**1**.NumPy:NumPy is the fundamental package for scientific computing with Python. NumPy is an extension to python programming language, adding support for large, multi-dimensional array and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

**2.SciPy:** SciPy is open-source software for mathematics, science, and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation.

**3.PANDAS:**Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labelled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

**4.SciKit-Learn:** The Scikit-learn project started as SciKits. The original code base was later extensively rewritten by other developers. This Consists of many methods for data mining tasks.

### 2.3 Django Framework:

Django follows the MVC pattern closely; however it does use its own logic in the implementation. Because the "C" is handled by the framework itself and most of the excitement in Django happens in models, templates and views, Django is often referred to as an MTV framework.

In the MTV development pattern:M stands for "Model," the data access layer. This layer contains anything and everything about the data: how to access it, how to validate it, which behaviors it has, and the relationships between the data. T stands for "Template," the presentation layer. This layer contains presentation-related decisions: how something should be displayed on a Web page or other type of document.V stands for "View," the business logic layer. This layer contains the logic that accesses the model and defers to the appropriate template(s). You can think of it as the bridge between models and templates.

A Django template is a string of text that is intended to separate the presentation of a document from its data. A template defines placeholders and various bits of basic logic (template tags) that regulate how the document should be displayed. Usually, templates are used for producing HTML, but Django templates are equally capable of generating any text-based format.

## III. LITERATURE REVIEW

World Wide Web Consortium (W3C) is the international standards organization for the World Wide Web (www). It develops standards, specifications and recommendations to enhance the interoperability and maximize consensus about the content of the web and define major parts of what makes the World Wide Web work. Phishing is a type of internet scams that seeks to get a user's credentials by fraud websites, such as passwords, credit card numbers, bank account details and other sensitive information. There are some characteristics in webpage source code that distinguish phishing websites from legitimate websites and violate the w3c standards, so we can detect the phishing attacks by check the webpage and search for these characteristics in the source code file if it exists or not.

(Mona Ghotaish Alkhozae, Omar Abdullah Batarfi, et al., 2011)proposed a phishing detection approach based on checking the webpage source code, they extracted some phishing characteristics out of the W3C standards to evaluate the security of the websites, and check each character in the webpage source code, if they find a phishing character, they would decrease from the initial secure weight [1].

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

There are many phishing detection techniques available, but a central problem is that web browsers rely on a black list of known phishing website, but some phishing website has a lifespan as short as a few hours. These website with a shorter lifespan are known as zero day phishing website. Thus, a faster recognition system needs to be developed for the web browser to identify zero day phishing website.(Chandan, Chheda, Gosar, R. Shah, Bhave, et al., 2013)[2].

In existing Online Phishing Detection systems, usually the reference to the database is taken for making any conclusion about the degree of phishiness of the website. Concentrating on getting the necessary attributes in real time environment using Hadoop, Map Reduce, increases both speed & efficiency of the system. (Kaustubh A. Hiwarekar, Dr. R. C. Thool, et al., 2013)[3].

A heuristic algorithm is to distinguish phishing sites from others based on user'sexperience that is heuristics checks if a site seems to be phishing site. A heuristic based solution employs several heuristics and converts each heuristics into a vector (Miyamoto, Hazeyama, Kadobayashi, et al., 2008) [4].

## IV. PHISHING WEBSITE DETECTION METHODOLOGY

### 4.1 Gathering Data

The firstimportant task is gathering data. We found some websitesoffering fake links. The next task was finding clear URLs. There was a data set available [5]. We found around 400,000 URLs out of which around 80,000 were fake and others were clean.

### 4.2 Analysis

Step 1: Tokenizing the URL's.Some of the tokens identified one 'virus','exe','php','wp','dat' etc.

Step2: Load data into a list and to store it.

Step 3: We have vectorized URLs. We used tf-idf scores instead of using bag of words classification since there are words in urls that are more important than other words e.g. 'virus', '.exe' ,'.dat' etc.

Step 4: Aftervectorizing the URL's, they are converted it into test data and training data.

Step 5: The logistic regression is performed.

The proposed System focuses on predicting the URL whether is fake or not, based upon the previous trends of dataset available. The dataset collected consists of above 4 lakh URLs .Which consist of both fake (phishing and fake) and legitimate URLs obtained from various sources. They can be updated by admin just by updating the file which consists of the URLs. Python in Django Framework has been used because of its vast set of libraries. We used Logistic Regression technique, which is a binary classifier, to plot the available dataset and t predict the URL

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

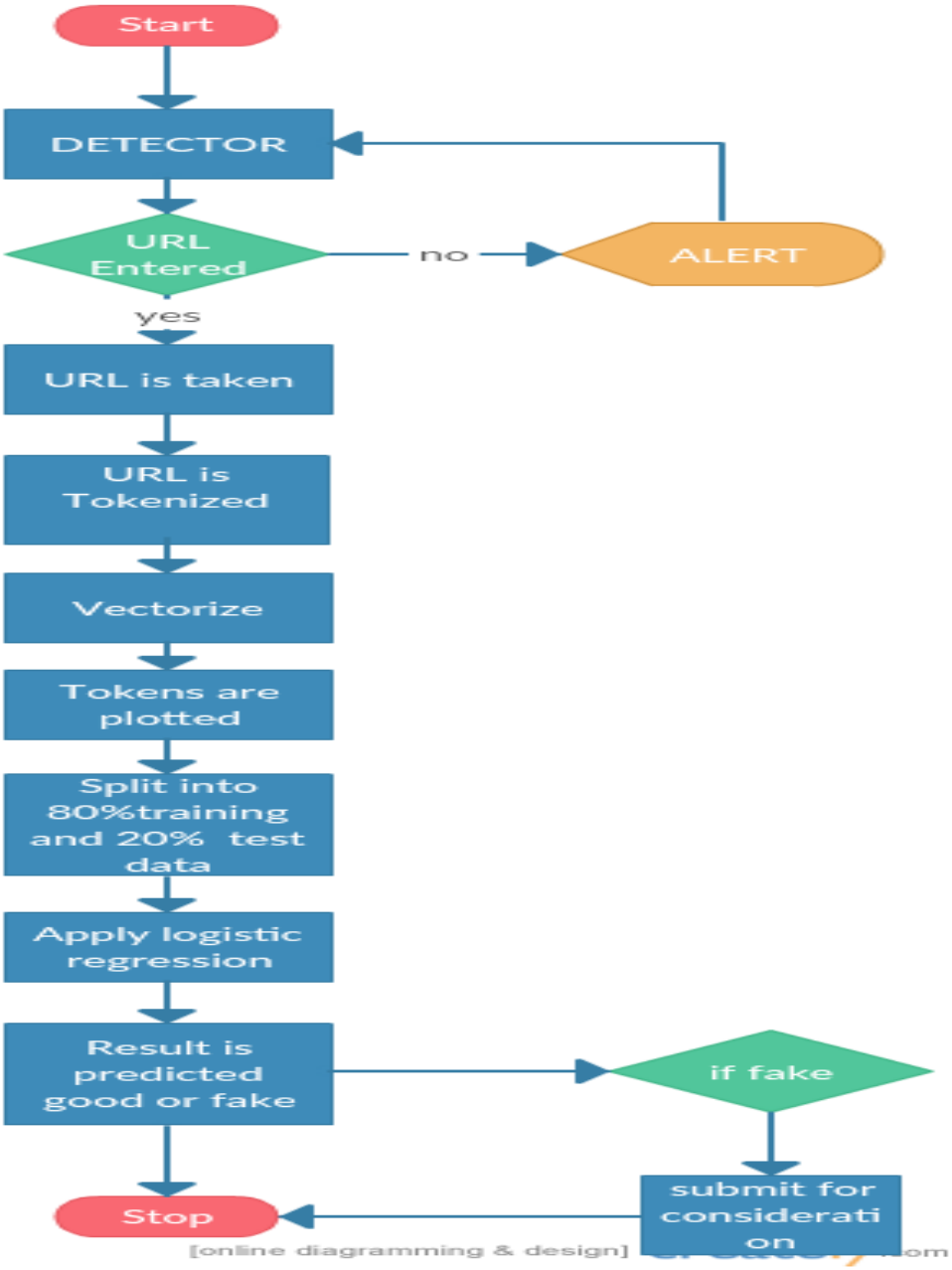


**Fig 1 Flow chart for the proposed system**

### 4.3 Implementation

1. The detector page is selected from the tab to the left of the page.
2. Then the URL is entered into the form page. If no URL is entered then an alert pops up asking to enter a URL.
3. After URL is entered it passes through URLS.py and then to views.py.
4. The URLS in the data set are then tokenized first using the predefined tokenizer function.
5. The tokenized URLs are then again vectorized and are plotted by splitting data into 80% training set and 20% testing data.
6. Now we apply logistic regression to detect whether our URL which is taken from the Webpage is fake or good.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

7. If the webpage is good the page gets redirected and message displays it as good.
8. If the webpage is fake the page redirects to show that it's bad and that website would be considered to be included in the further considerations by the admin.

### 4.3.1 Admin

Admin plays a crucial role in this type of system. The list of functions of an admin is given below.

- The admin should update the data set periodically.
- The submitted URLS can be seen in text.txt file which will in the path of the file.

### 4.3.2 Logistic regression

Logistic regression is a regression model where the dependent variable (DV) is categorical. This Logistic regression covers the case of a binary dependent variable--that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analyzed in multinomial logistic regression orif the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.

### 4.3.3 Detector Module

The detectormodule will consists of a form field through which the URL is taken. The URL is then sent to test its legitimacy and the result is displayed in the output page. If no URL is entered then the page will display an alert to show to enter a URL. The form field is accessible in the views.py file which consists of the whole logic of the code.

### 4.3.4 Submit Module

The submit module also contains only a single field. The field will take suggested URL from the user and writes it to the notepad directly which can be accessed by the admin. The admin further checks for any requirements and can update it to the master list of URL's directly. If no URL is entered an alert pops up asking to enter a URL. If URL is written to the file after submission then it displays a success message that the URL has been successfully added to the list.

## V. SYSTEM DESIGN

### 5.1 MVT ACHITECTURE

Model View Template or MVT as is popularly called, is a software design pattern for developing web applications. A Model View Template pattern is made up of three parts:

- Model - The lowest level of pattern which is responsible for maintaining data.
- View - This is responsible for displaying all or a portion of data to user.
- Template – This layer deals about the presentation of data to the user.

MVT is popular as it isolates the application logic from the user interface layer and supports separation of concerns. Here the urls.py will take care of all the calls and maps them to the necessary views. From views the required views functions are called. The views may contain both logic and code necessary for the Functionality. In DjangoFramework the database declarations and its related operations are done in Model. Hence it is

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

responsible for maintaining data. Before running the server the model has to be applied for migrations so that changes to the database made are committed.

### 5.1.1 The model

This layer contains anything and everything about the data: how to access it, how to validate it.

### 5.1.2 The view

This layer acts as a bridge between Model and templates. This layer consists of the business logic.

### 5.1.3 The Template

This layer deals with how something should be displayed on a webpage or other document.

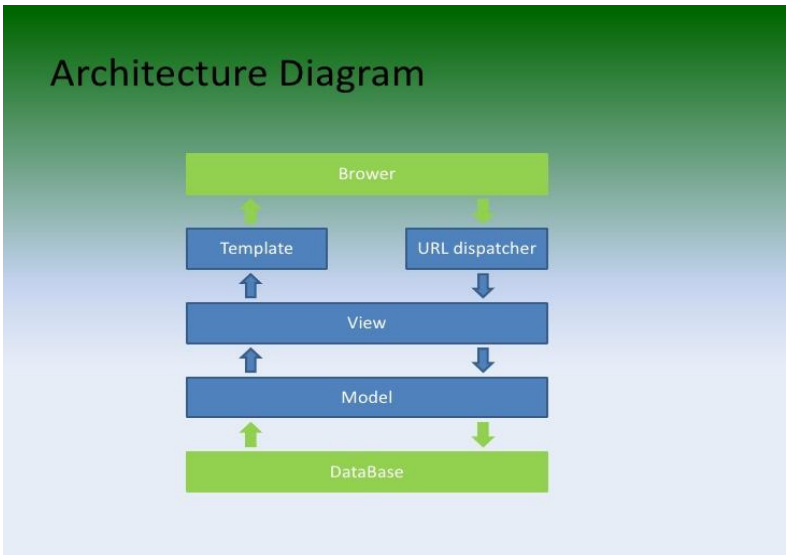


**Fig2 Architecture Diagram for MVT**

## VI..RESULTS

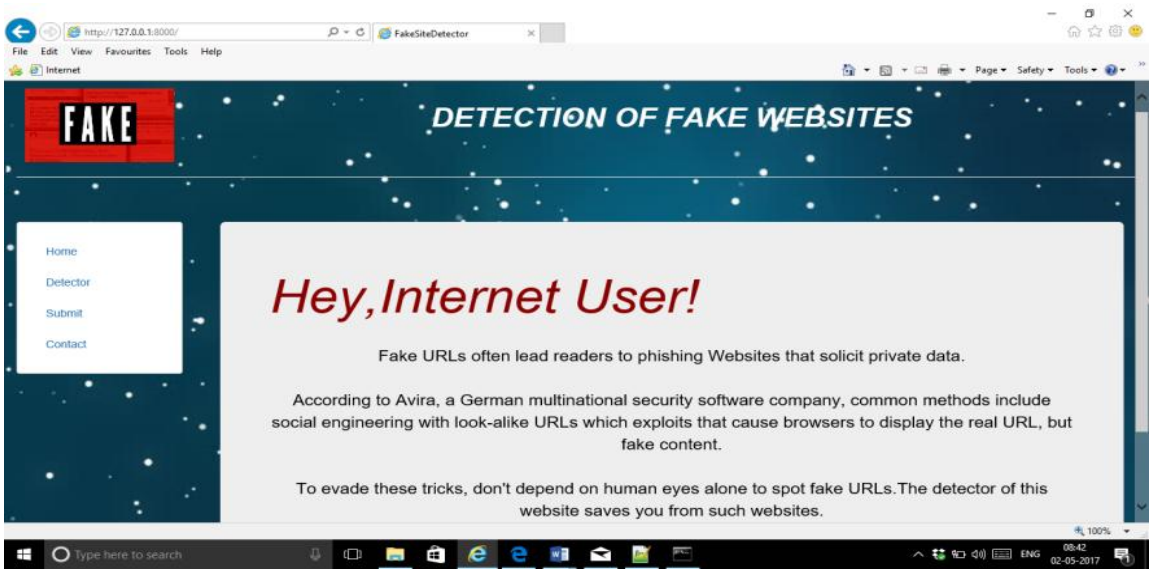


**Fig 3 Homepage**

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 08, August 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

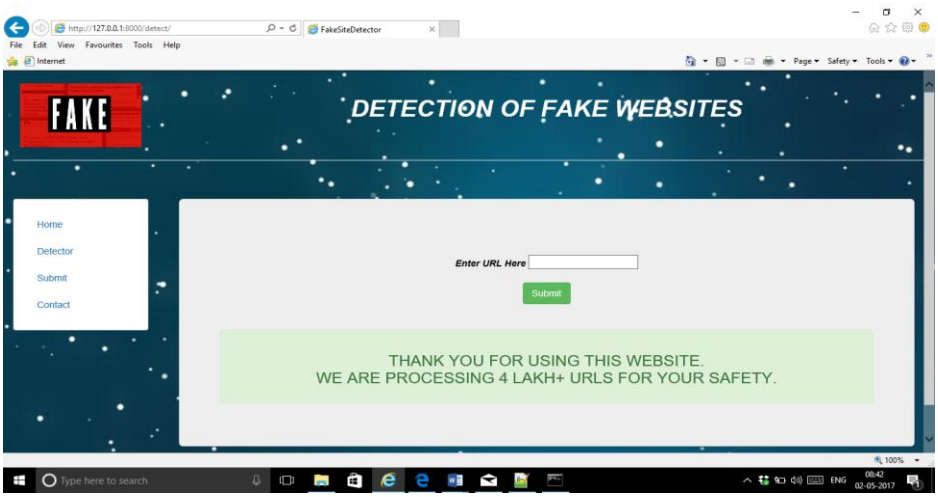


**Fig 4 Detector Page**

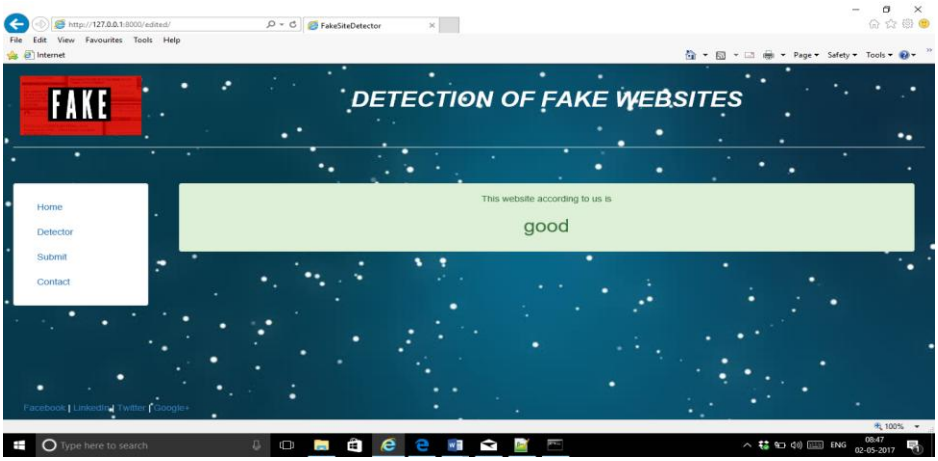


**Fig 5 Output for good URL**

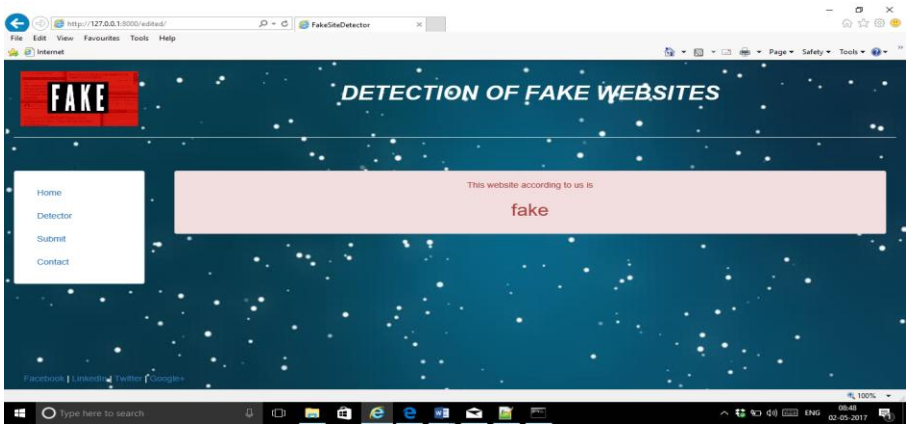


**Fig 6 Output for fake URL**

## VIII. CONCLUSION AND FUTURE WORK

As the usage of internet increases, the number of users accessing the websites available on the net has also increased. Multiple websites ask the users to provide sensitive information like email ids, phone numbers, etc. for registration purposes. Some banking and utility websites require more personal information like bank

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

account details and credit card numbers. Phishing websites deceive users and trick them into believing that they are using the original websites.This paper implemented an approach using data mining to detect and predict phishing websites based on characteristics like URL and domain identity. For this purpose regression has been used which predicts whether the entered URL points to a fake or good website.In future the master list of the URLs could be updated dynamically for more accurate and up to date prediction.

## REFERENCES

[1] Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code byMona Ghotaish Alkhozae, Omar Abdullah Batarfi.

[2] A Machine Learning Approach for Detection of Phished Websites Using Neural Networks by Charmi J. Chandan, Hiral P. Chheda, Disha M. Gosar, Hetal R. Shah.

[3] Phishing Detection System Using Machine Learning and Hadoop-MapReduce Kaustubh A. Hiwarekar, Dr. R. C. Thool.

[4] An Evaluation of Machine Learning-based Methods for Detection of Phishing Sites by Daisuke Miyamoto, Hiroaki Hazeyama, and Youki Kadobayashi.

[5]https://archive.ics.uci.edu/ml/datasets/URL+Reputation (for dataset)

[6]http://sysnet.ucsd.edu/projects/url/ (for dataset)