# SMART MALICIOUS URL'S DETECTION SYSTEM TO PREVENT PHISHING USING DEEP LEARNING APPROACH

Submitted in partial fulfillment of the requirements

of the degree of

(Bachelor of Engineering)

by

Jainam Soni BE-4 24

Palak Nisar BE-3 66

Shamika Dumbre BE-3 29

Siddhi Sheth BE-4 15

Guide:
Prof. Deepti Nikumbh


**Department of Computer Engineering**

**Shah and Anchor Kutchhi Engineering College, Mumbai**

**University of Mumbai, Mumbai**

**Year 2018-2019**

www.shahandanchor.com Tel: 022 2558 0854

Mahavir Education Trust's

SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE

Mahavir Education Trust Chowk, W.T. Patil Marg, Chembur, Mumbai 400 088

Affiliated to University of Mumbai, Approved by D.T.E. & A.I.C.T.E. ISO 9001:2008 Certified

Awarded provisional accreditation for Computer & Electronics Engineering by NBA

(for 2 years from 06-08-2014)

# Certificate

This is to certify that the report of the project entitled

**SMART MALICIOUS URL'S DETECTION SYSTEM TO PREVENT PHISHING USING DEEP LEARNING APPROACH**

is a bonafide work of

| | |
|---|---|
| Jainam Soni | BE4-24 |
| Palak Nisar | BE3-66 |
| Shamika Dumbre | BE3-29 |
| Siddhi Sheth | BE4-15 |

submitted to the

**UNIVERSITY OF MUMBAI**

during semester VII in partial fulfilment of the requirement for the award of the degree of

**BACHELOR OF ENGINEERING**

in

**COMPUTER ENGINEERING.**

---

(Prof.Deepti Nikumbh)
Guide

---

(Prof. Uday Bhave)
I/c Head of Department

---

(Dr. Bhavesh Patel)
Principal

# Project Report Approval for B. E. Semester VII

This project report entitled ***Smart Malicious Url's Detection System To Prevent Phishing Using Deep Learning Approach*** by ***Jainam Soni, Palak Nisar, Shamika Dumbre, Siddhi Sheth*** is approved for **Semester VII** in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineeering**.

Name and Signature of the Examiner

1.----------

2.----------

Guide

1.----------

2.----------

Date:

Place:

# Declaration

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

| Name of Student | Roll No | Signature |
|---|---|---|
| Jainam Soni | BE4-24 | |
| Palak Nisar | BE3-66 | |
| Shamika Dumbre | BE3-29 | |
| Siddhi Sheth | BE4-15 | |

Date:

## Attendance Certificate (from college)

**Date**

**To,**
**The Principal**
**Shah and Anchor Kutchhi Engineering College,**
**Chembur, Mumbai-88**

**Subject: Confirmation of Attendance**

**Respected Sir,**

**This is to certify that Final year (BE) students Jainam Soni, Palak Nisar, Shamika Dumbre and Siddhi Sheth have duly attended the sessions on the day allotted to them during the period from _____to_____ for performing the Project titled _____.**

**They were punctual and regular in their attendance. Following is the detailed record of the student's attendance.**

**Attendance Record:**

| Date | Jainam Soni | Palak Nisar | Shamika Dumbre | Siddhi Sheth |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

**Signature and Name of Internal Guide**

# 4. Abstract

Over the past years, there has been an increase in the amount of phishing attacks and security threats. Phishing is a malicious practice in which the attacker fraudulently acquires confidential information like bank details, credit card details, or passwords from legitimate users. In phishing, users are tricked with an phished website containing malicious url rather than legitimate one. Here we propose an anti-phishing technique to safeguard our web experiences. Our approach uses an automatic feature extraction of a website to detect any suspicious or phishing website. These features are passed to Long Short Term Memory (LSTM) to predict whether the url is malicious or benign. The results obtained from our experiment shows that our proposed methodology is very effectual for preventing such attacks as it has better accuracy than other traditional algorithm and Recurrent Neural Network (RNN).

Keywords: Social Engineering, URL, Phishing, RNN, ANN, LSTM.

# Chapter 1

# 1. Introduction

An identity theft that occurs when a malicious web site masquerades a legitimate one is called Phishing. Such a theft occurs in order to procure sensitive information such as passwords, bank account details, or credit card numbers. Phishing makes use of spoofed emails which look exactly like an authentic email. These emails are send to a bulk of users and appear to be coming from legitimate sources like banks, e-commerce sites, payment gateways etc. The makers of such illegitimate website made them exactly look like a legitimate one so that no user can identify the difference easily. The phishing attackers use different kind of social engineering tactics to lure users for example: giving attractive offers to just visit the site.

Malicious URL is a URL created with malicious purposes, among them, to download any type of malware to the affected computer, which can be contained in spam or phishing messages, or even improve its position in search engines using Blackhat SEO techniques.

**Machine learning** is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to learn. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; example applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank and computer vision.

Smart Malicious Urls Detection System is an anti-phishing technique to safeguard our web experiences Our approach uses an automatic feature extraction of a website to detect any suspicious or phishing website. These features are passed to Long Short Term Memory (LSTM) to predict whether the url is malicious or benign. The results obtained from our experiment shows that our proposed methodology is very effectual for preventing such attacks and the performance was measured by using Confusion Matrix for the classifier.

## 1.1 Objective:

To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years. So we here try to develop a system with various machine learning techniques or deep learning approach that gives the best results.

To develop a Smart Malicious URL phishing detection system for end user with the following characteristics:

- Malicious URL Detection System to curb the Phishing attacks.
- The GUI of our system will engage end users and provide user friendly experience.
- System will be based on discerning Urls by their patterns.
- No manual Feature Extraction.
- Maximise the prediction result of the system than the others which are developed.
- Create awareness about phishing attack and some other cyber security threats.

## 1.2 Problem Statement:

The main purpose of the system is to not only protect the social network management system, but also protect users from being exposed to malicious content and phishing attacks i.e to minimise the cyber threats that is spread over internet rapidly and is growing each day with increase in newer technologies and with the large exposure of Internet.

## 1.3 Methodology used:

The purpose of this project is to build a classifier that can detect malicious URLs. This is accomplished using a Featureless Deep Learning approach. The more traditional approach requires deriving hand-crafted features prior to training the Machine Learning classifier. This can not only be the most tedious part, but also require advanced domain expertise and data wrangling skills. There are 4 main "URL features families":

1. BlackList Features
2. Lexical Features
3. Host-based Features

4. Content-based Features

An alternative approach can be Featureless Deep Learning, where an embedding layer takes care of deriving feature vectors from the raw data.

**Dataset Collection:**

The dataset (containing both malicious and benign URLs to train the Deep Learning binary classifier) was custom build from various open source data sources. Note that for training better Deep Learning classifiers much more data is needed.

**Dataset Preprocessing:**

Limited pre-processing of the raw URLs is still necessary. For “One Hot Encoding”, Each character has to be expressed as unique integer whereas, For “Word2vec”, Each character has to be expressed as unique integer as well as all URLs have to be of the same length. This results in cropping or padding with zeros (Only if it is using Word2vec). Word2vec is trained first, then applied to each URL to embed the URL and thus derive it's "features", so that the actual binary classifier can be trained thereafter. Using the CBOW method the characters that surround a target character for which the embedding vector is predicted will be used as context. Here we can define our own vocabulary size and embedding dimension. Each character sequence exhibits correlations, that is, nearby characters in a URL are likely to be related to each other. These sequential patterns are important because they can be exploited to improve the performance of the predictors.

**LSTM Model Unit:**

Recurrent Neural Networks (RNN) are a type of neural network that is able to model sequential patterns. It allows them to process sequential data one element at a time and learn about there sequential data elements. But RNN’s limitation is that they cannot correlate elements that are 5 or 10 times apart. This can be overcome by LSTM, as it can correlate elements that are more than 1000 steps apart without loss of short time lag capabilities. In these each neuron is replaced by a memory cell that uses multiple units as gates to control the flow of Information.

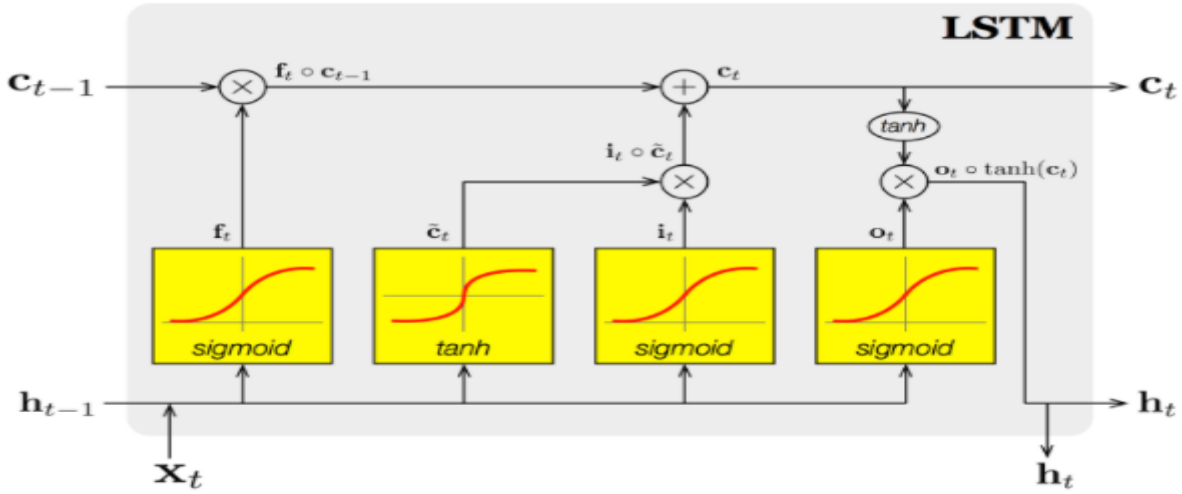



Figure 1.3.1

The first sigmoid activation function is the **forget gate**. Which information should be forgotten from the previous cell state (Ct-1). The second sigmoid and first tanh activation function is our **input gate**. The last sigmoid is the **output gate** and highlights which information should be going to the next **hidden state**.

**LSTM Predictive Classifier Model:**

Thus LSTM model is build that receives an input of Url as character sequence and predicts whether the Url is benign or malicious or not.

## 1.4 Organization of report:

Chapter 1 of the report introduces the concepts of Phishing and ML and Deep Learning, along with the issues faced in the traditional detection of Url's. Chapter 2 presents the review of the existing literature concerned with the system. Chapter 3 gives a detailed description about the proposed system and all the requirements related to it. Chapter 4 specifies the design specifications system architecture along with the details of the block diagram of LSTM design. Chapter 5 enumerates the list of references and Chapter 6 gives the Acknowledgement of the project.

# Chapter 2

# Literature Survey

Typically phishing attacks can be exploited by sending spoofed link to the trusted user and then redirected them to bogus website. Whenever user will enter the important information to that fake website then immediately this information will store to the system of the hacker and then finally the user will be redirected to any other websites that will not be related to the user .There were many research that were conducted to detect phishing attacks.

Phishtank is a website which identifies if a website is legitimate on the basis of the data stored in it. It uses the black list based approach which is fast but has disadvantage of very low detection rate.

- "Using supervised machine learning algorithms to detect suspicious URLs in online social networks" [1].

The paper describes a supervised machine learning classification model that has been built to detect the distribution of malicious content in online social networks (ONSs). For the data collection stage, the Twitter streaming application programming interface (API) was used and VirusTotal was used for labelling the dataset. It uses a Random Forest Classification Model with the combination of different features. This led to a recall value of 0.89 and after parameter tuning and feature selection method, it could improve the performance to 0.92.

- "Deep Sentiment Representation Based on CNN and LSTM, IEEE 2017" [2]

- "Detecting Malicious URLs using machine learning techniques" [6].

An approach was proposed which uses the lexical, host based and other features extracted from the web page rather than executing full page matching and then passing them to different types of classifiers such as SVM, KNN, Logistic-Regression, Gradient Boosting, Tree-Bagging Classifier and Decision- Tree Classifier.

- "A novel approach to protect against phishing attacks at client side using auto-updated white-list" [4].

An approach was proposed to prevent against the Phishing attacks by the use of auto- updating white lists. In this approach, when user tries to open a website, the browser warns him/her to

not access the website if that website is not available in the white-list. This technique also examine the legitimacy of given website using hyperlink features. Hyperlinks are extracted from the source code of given webpage and are passed as an input to the proposed phishing detection algorithm. This proposed approach is effective as it has true positive rate of 86.02 % while false negative rate lesser 1.48%.

- “Malicious Url Detection using Google Safe Browsing Api” .

One of the service which the Google provides for safe browsing allows to check the URL against a list of malicious domains that is constantly updated by Google. The Safe Browsing Lookup API permits to pass the suspicious website’s url to Safe Browsing service which tells if the URL sent by the client is benign or malicious.The client URLs are verified using the malicious and phishing lists maintained by Google. However,this approach has the following shortcomings: (i) Before sending the URL hashing is not performed and (ii) There isn’t any constraint over the response time taken server to lookup.

# Chapter 3

# System Requirements

## 3.1 Functional Requirements

The system should be able to convert text to feature, which can take the necessary part and obtain a feature vector. The system should provide a well trained Autoencoder to generate better inputs for classifier. The system needs a classifier which is well trained to detect the correctness of urls.

## 3.2 Hardware Requirements:

- Computer with minimum configuration of processor 1.33 GHz.
- 512mb RAM.
- 80 GB hard disk.

## 3.3 Software Requirements:

- Python 2.7 or Higher.
- It can be used on all platforms with help of PyInstaller (Windows, Linux, Mac OS X, etc.)

# Chapter 4

# Design

## 4.1 Proposed Architecture Design.

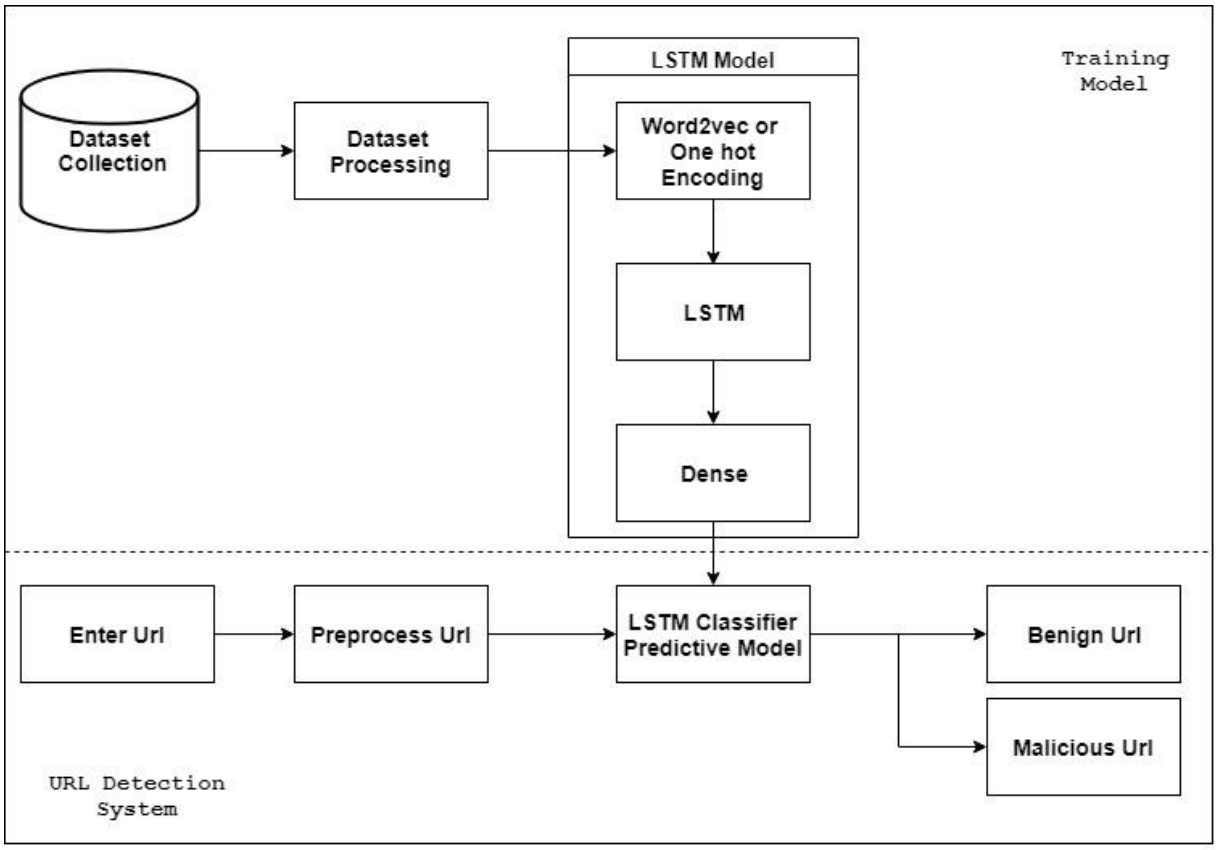
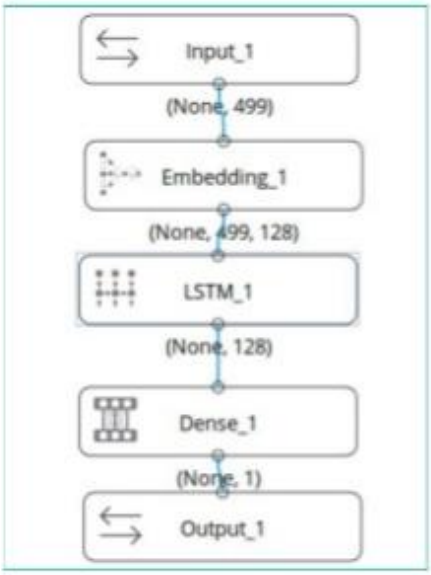



Figure 4.1.1

## 4.2 LSTM Model Block Design.




**Input_1:** Layer that represents a particular input port in the network.

**Embedding_1:** Turn positive integers (indexes) into dense vectors of fixed size, with dropout rate **0.2**.

**LSTM_1:** LSTM Layer. Dropout rates for gate and itself is **0.2**. Activation function **tanh** is used.

**Dense_1:** Just your regular densely-connected nn layer. Activation function **sigmoid** is used.

**Output_1:** Layer that represents a particular output port in the network.

Figure 4.2.1

## 4.3 User Interface Design

### 4.3.1 System Flowchart

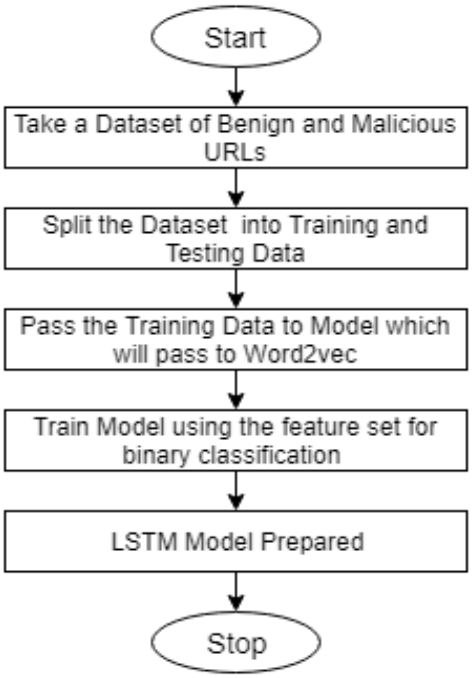



Figure 4.3.1.1

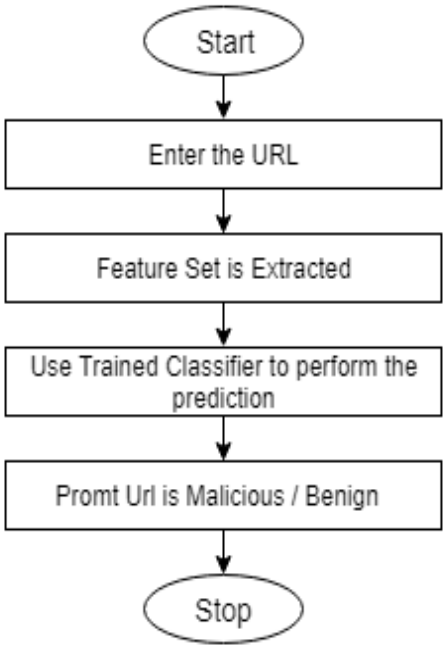



Figure 4.3.1.2

## 4.3.2 Input Output Forms

The prototype for GUI is as shown below:

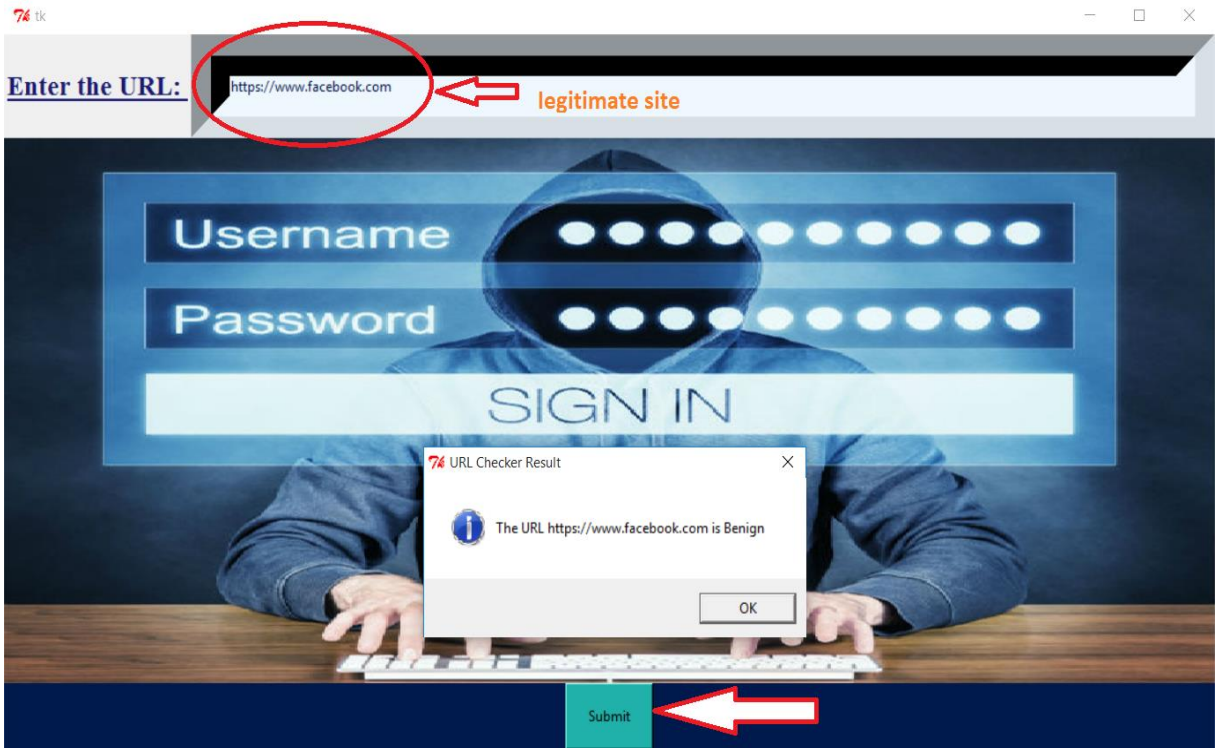



Figure 4.3.2.1

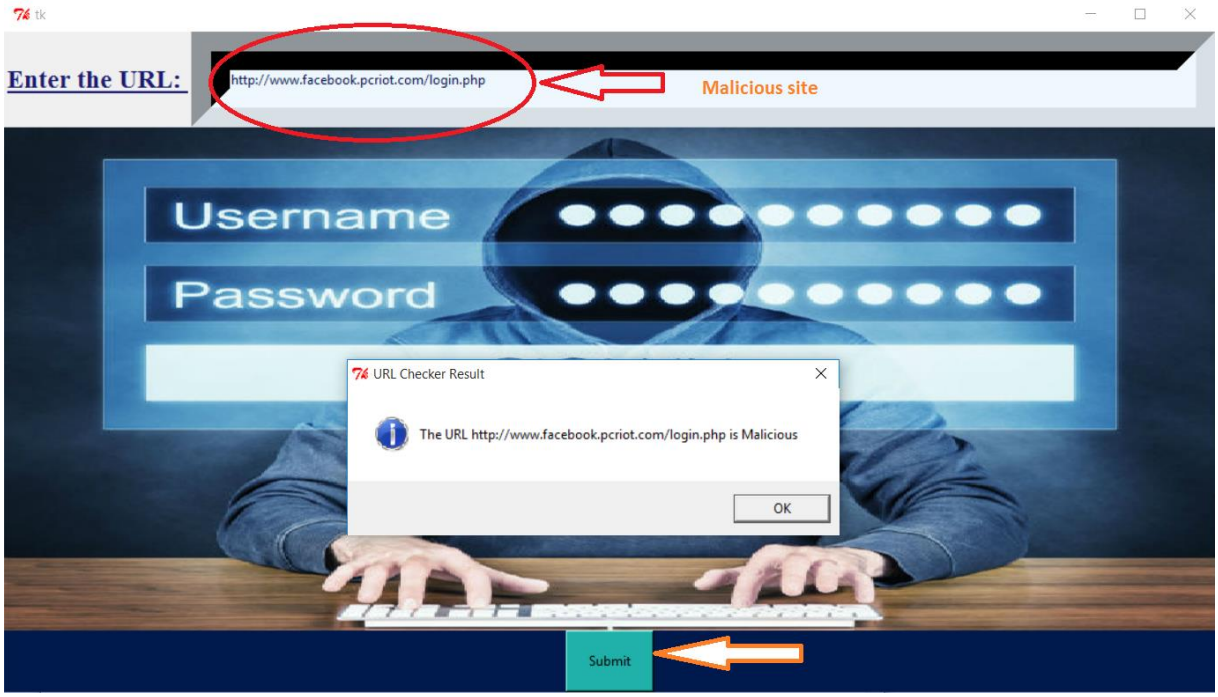



Figure 4.3.2.2

# Chapter 5

# References


[1] Mohammed Al-Janabi, Ed de Quincey, Peter Andras, Using supervised machine learning algorithms to detect suspicious URLs in online social networks, ACM 2017.

[2] Qiongxia Huang, Xianghan Zheng, Riqing Chen and Zhenxin Dong, Deep Sentiment Representation Based on CNN and LSTM, IEEE 2017.

[3] A.Bavani, D.Aarthi, V.C, Detecting phishing websites on real time using anti-phishing framework, Department of Information Technology (UG) 1, 2, 3, Assistant Professor4 Kingston Engineering College, India, 2017.

[4] Adulghani Ali Ahmed, N. A. A.: 2016, Real time detection of phishing websites, IEEE

[5] A novel approach to protect against phishing attacks at client side using auto-updated white-list, IEEE 2016.

[6] Longfei WC, J. W.: 2016, A lightweight anti-phishing scheme for mobile phones,

(IEEE: 2016)-MobiFish.

[7] Frank Vanhoenshoven, Gonzalo Napoles, Koen Vanhoof , Detecting Malicious URLs using machine learning techniques, IEEE, 2016.

[8] A. K. Shrivas, R. S.: 2015, Decision tree classifier for classification of phishing website with info gain feature selection, IJRASET.

[9] Afroz, S. and Greenstadt, R.: 2015, Detecting phishing websites by looking at them, IEEE Communications Society.

[10] Zhang, J. and Wang, Y.: 2012, A real-time automatic detection of phishing urls.

[11] Common Crawl: www.commoncrawl.org.

[12] Phishtank: www.phishtank.com.

[13] Kaggle: www.kaggle.com.

# List of Figures



# List of Abbreviations

| Abbr. | Description |
|---|---|
| Url | Uniform Resource Locator |
| ANN | Artifical Neural Network |
| RNN | Recurrent Neural Network |
| LSTM | Long Short Term Memory |
| CNN | Convolution Neural Network |
| SVM | Support Vector Machines |
| KNN | K-Nearest Neighbors |

.

# Chapter 6

# Acknowledgement

We sincerely express our deep gratitude to our Principal Dr. Bhavesh Patel, our Head of Department Mr. Uday Bhave and our guide Prof. Deepti Nikumbh for providing us with their invaluable guidance, advice and suggestions. We were fortunate to have met such supervisors. We would like to thank our guides for providing us with the opportunity to do this project Smart Malicious Url's Detection System to Prevent Phishing Using Deep Learning Approach, which helped us in doing lot of research and learning new things. We acknowledge with a deep sense of gratitude, the encouragement and inspiration received from our faculty members and colleagues.