A Report On

# PHISHING WEBSITE DETECTION

Submitted by:-

1. Mr. Lanchenba Ngangom
2. Miss. Shraddha Devkar
3. Miss. Ankita Mankar
4. Miss. Vandana Pal
5. Miss. Dikshika Yeul
6. Mr. Animesh Roy
7. Mr. Abhishek Revaskar

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **Table of Contents** | **Page No.** |
| **01** | Problem Statement | **05** |
| **02** | Introduction | **05** |
| **03** | Existing System | **06** |
| **04** | Objective | **06** |
| **05** | Development Technologies and Tool | **06** |
| **06** | Methodology | **07** |
| **07** | Flow of the System | **11** |
| **08** | Result | **13** |

## ProblemStatement

Tocreateanextensionforchromewhichwillactasmiddlewarebetweentheusersandthephishingwebsites.

## Introduction

### What is a phishingwebsite?

* + 1. Phishingisanactofattemptingtoacquireinformationsuchasusername,password,creditcarddetails,etcofaperson ororganisationillegallyinelectroniccommunication.
    2. Inphishing,criminalscreateafakewebsitewhoselooksandfeelareidenticaltothelegitimateone,inwhichthevictims aretoldtoentertheirconfidentialdetailslikeusername,passwordandaccountdetails.

### Overview

* + 1. AtoolisneededwhichexaminestheURLenteredbytheuserandchecksifthewebsiteismaliciousornot.
    2. Basically,thesystemisagooglechromeextensionasitisa verygoodwayofensuringeasyaccesstothetoolfortheuser.
    3. WhentheuserentersaURL,theextensiontakestheURLandpredictswhetherthewebsiteisphishingornot.

## Existing System

* 1. Traditionalapproachtodetectwhetherthewebsiteisphishingorlegitimateistocheckwhetherthewebsiteispresentinafakewebsitedatasetornot.
  2. Thedrawbackoftheabovemethodisthatthedatasetisincreasingdaybyday.
  3. Blockthephishingemailsbyvariousspamfiltersoftware.

## Objective

* 1. ToimplementclassificationtechniquesforphishingURLdetection.
  2. Todevelopanextensionusingscriptinglanguages.
  3. Togenerateanalertmessageforadetectedphishingwebsite.
  4. Tomaintaintherecordofdetectedphishingwebsites.

### Development Technologies and Tools

* 1. **Backend Technology- Machine LearningAlgorithms**
     1. LogisticRegression
     2. RandomForest
     3. Support VectorMachine

### ScriptingLanguages

* + 1. HTML5
    2. CSS3
    3. Javascript
    4. PHP

### DevelopmentTool

* + 1. Visual StudioCode(1.45.1)

## Methodology

* 1. Phishingattacksareonarise,toprovideprotectionagainstnewattacks,machinelearningprovidesthebestsolution.
  2. WeusedtheUCIDatasetofPhishingWebsitetotraintheclassifier.
  3. Later,wheneverauserenterstheURL,thefeaturesareextracted andtheURListestedonthetrainedclassifiertoobtaintheresult.
  4. Wehaveconsideredthebelowfeaturestofindoutifthewebsiteisphishingorlegitimate.
     1. **Presence of IP address in URL:**If IP addresspresentinURLthenthefeatureissetto1elsesetto0.

MostofthebenignsitesdonotuseIPaddressasanURLto downloada webpage.Use of IP address inURL indicates that attacker is trying to steal sensitive information.

* + 1. **Presenceof@symbolinURL:If**@symbolpresent

inURLthenthefeatureissetto1elsesetto0.Phishers addspecialsymbol@intheURLleadsthebrowserto

ignoreeverythingprecedingthe“@”symbolandtherealaddressoftenfollowsthe“@”symbol.

* + 1. **Number of dots in Hostname:**Phishing URLs have many dots in the URL. For example[http://shop.fun.amazon.phishing.com,](http://shop.fun.amazon.phishing.com/)inthisURLphishing.comisanactualdomainname,whereasuseof“amazon”wordisto trickusersto clickonit.

Average number of dots in benign URLs is 3. If the

numberof dotsinURLsismorethan3thenthefeatureisset to 1

else to 0.

* + 1. **Prefix or Suffix separated by(-)to domain:**Ifdomainnameseparatedbydash(-)symbolthenfeatureissetto1elseto0.Thedashsymbolisrarelyusedin legitimateURLs.Phishersadddashsymbol(-)tothedomainnamesothatusersfeelthattheyaredealingwithalegitimatewebsite.Forexample:Actualsiteis[http://www.onlineamazon.com](http://www.onlineamazon.com/)butphishercancreate

another fake website like[http://www.online-amazon.com](http://www.online-amazon.com/)to confuse the innocent users.

* + 1. **URLredirection:If**“//”presentinURLpaththenfeatureissetto1elseto0.Theexistenceof“//”within theURLpathmeansthattheuserwillberedirectedtoanotherwebsite.
    2. **HTTPS token in URL:**If HTTPS token present inURLthenthefeatureissetto1elseto0.Phishersmay addthe“HTTPS”tokentothedomainpartofaURLin ordertotrickusers.Forexample,[http://https-www-](http://https-www-/)paypal-it-mpp-home.soft-hair.com.
    3. **Information submission to Email:**Phisher mightuse“mail()”or“mailto:”functionstoredirecttheuser’s

informationtohispersonalemail[4].IfsuchfunctionsarepresentintheURLthenfeatureissetto1elseto0.

### URL Shortening Services “TinyURL”:TinyURL

service allows phisher to hide long phishing URLby makingitshort.Thegoalistoredirectusertophishingwebsites.IftheURLiscraftedusingshorteningservices(likebit.ly)thenfeatureissetto1else0

* + 1. **LengthofHostname:Average**lengthofthebenignURLsisfoundtobea25,IfURL’slengthisgreaterthan25thenthefeatureissetto1elseto0
    2. **PresenceofsensitivewordsinURL:Phishing**sites

usesensitivewordsinitsURLsothatusersfeelthattheyaredealingwithalegitimatewebsite.BelowarethewordsthatfoundinmanyphishingURLs:-'confirm','account','banking','secure','ebyisapi','webscr','signin',

'mail','install', 'toolbar', 'backup', 'paypal', 'password','username',etc;

* + 1. **Number of slash in URL:**The number of slashes in benignURLsisfoundtobea5;ifnumberofslashesinURLisgreaterthan5thenthefeatureissetto1elseto0.
    2. **Presence of Unicode in URL:**Phishers can make auseofUnicodecharactersinURLtotrickuserstoclickonit.Forexamplethedomain“xn--80ak6aa92e.com”isequivalent to "аррӏе.com". Visible URL to user is"аррӏе.com"butafterclickingonthisURL,userwillvisitto“xn--80ak6aa92e.com”whichisaphishingsite.
    3. **Age of SSL Certificate:**The existence of HTTPS isveryimportantingivingtheimpressionofwebsitelegitimacy.ButminimumageoftheSSLcertificateof

a benign website is between 1 year to 2 year.

* + 1. **URL of Anchor:**We have extracted this feature bycrawlingthesourcecodeohtheURL.URLoftheanchorisdefinedby<a>tag.Ifthe<a>taghasamaximum numberofhyperlinkswhicharefromtheotherdomainthenthefeatureissetto1elseto0.
  1. ForclassifyingtheURLentered,aseitherphishingorlegitimate,weconsideredthefollowingthreealgorithms:

### LogisticRegression

Logistic regression is basically a supervised classificationalgorithm. Logistic regression becomes a classification techniqueonlywhenadecisionthresholdisbroughtintothepicture.ThesettingofthethresholdvalueisaveryimportantaspectofLogisticregressionandisdependentonthe classification problem itself. Fitting logistic regression andcreatingaconfusionmatrixofpredictedvaluesandrealvaluesitwasabletogetgoodaccuracy.

### RandomForest

Thisalgorithmmakesuseofdecisiontreesforclassification.Itcreatesmultipledecisiontreesatthetimeoftraining.

RandomForestsareacombinationoftreepredictors.Themainprincipleisthatagroupof“weaklearners”can combineandforma“stronglearner”.

### Support VectorMachine(SVM)

TheSVMalgorithmusesadatasetwheretheinputsamplesaredividedintotwoclasseswithlabelseither0or1.The

methodologyincludesfindingaline(intwodimensionspace)oraplane(inmultidimensionalspace)alsocalleda hyperplane which will most efficiently separate the twoclasses.Byplugginginthesedatapointsintothelineequation,wecancalculatewhetheranewpointisonwhichside of the line and the class will be predicted.

### Chrome extension

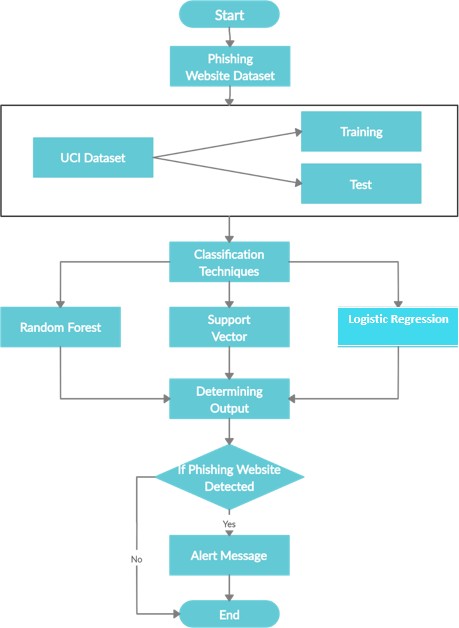
Extensionsareadd-onstothebrowserwhichhelpinaddingmore featuresandmakingbrowserusageeasierfortheuser.Wehaveimplemented a Browser Action type of extension.

### Localhostserver

Localhostisusedforcommunicationbetweenclientside(i.echromeextension)andserverside(pythoncode).Wehaveusedan XAMPPapplication.

### Flow of thesystem

* 1. WhenauserenterstheURL,theextensiontakestheURLwiththe helpofjavascriptfunctionsandpassesthisURLtothePHPscriptrunningontheserver.
  2. ThisURLispassedtothepythoncode.Insidepythoncode,the functionsarecreatedforextractingfeaturesfromURL.Everyfunctionreturnseither**1:PhishingWebsiteor-1:Legitimate Website.**
  3. Combiningallthereturnvaluesformsanarray,thenwetestthisonatrainedclassifierofrandomforest.
  4. Ifthewebsiteissafe,thenoutputwillbedisplayedontheextension.
  5. Ifthewebsiteisphishing,thenitwilldisplayontheextension alongwiththealertnotificationtotheuser.Thevisitedphishing sitewillbeaddedtothefiletokeeptherecordsofthepreviouslyvisited phishingwebsites.

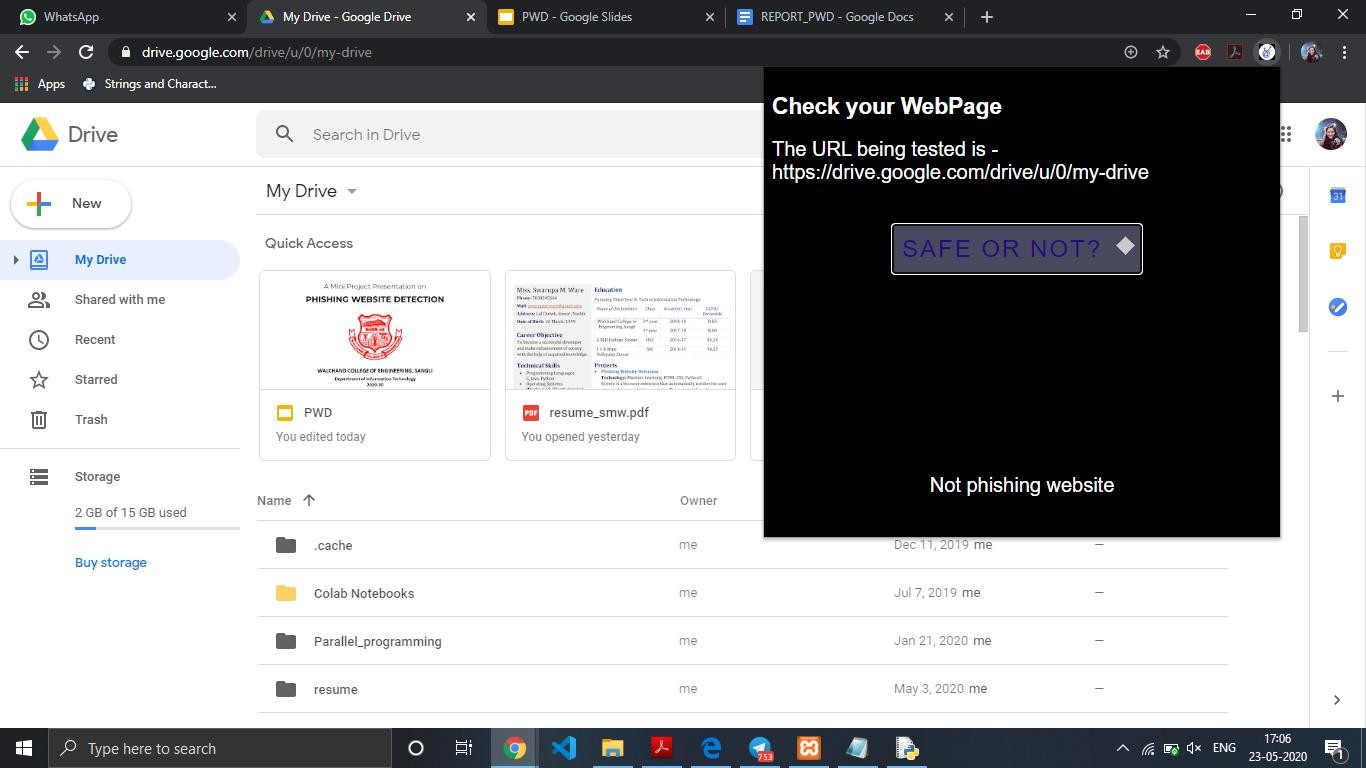


Flow Chart

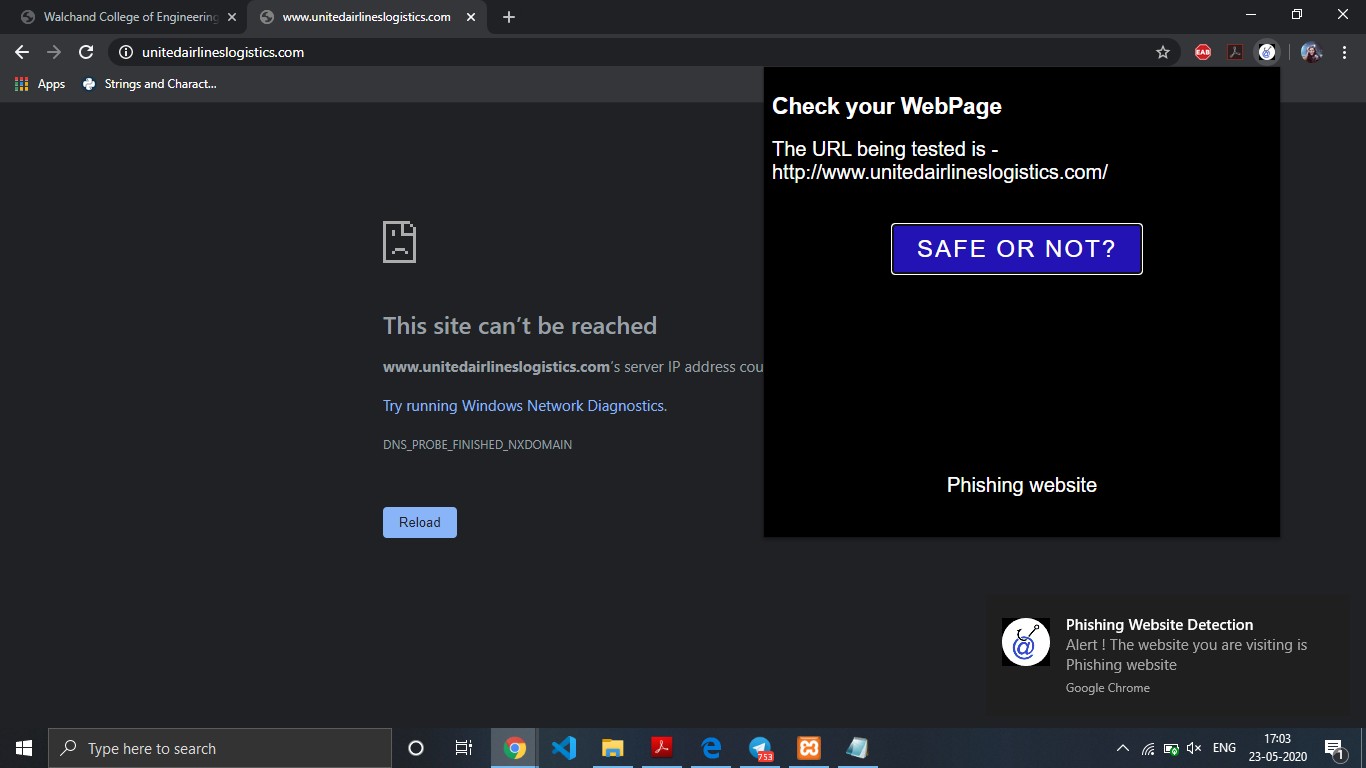
### Result

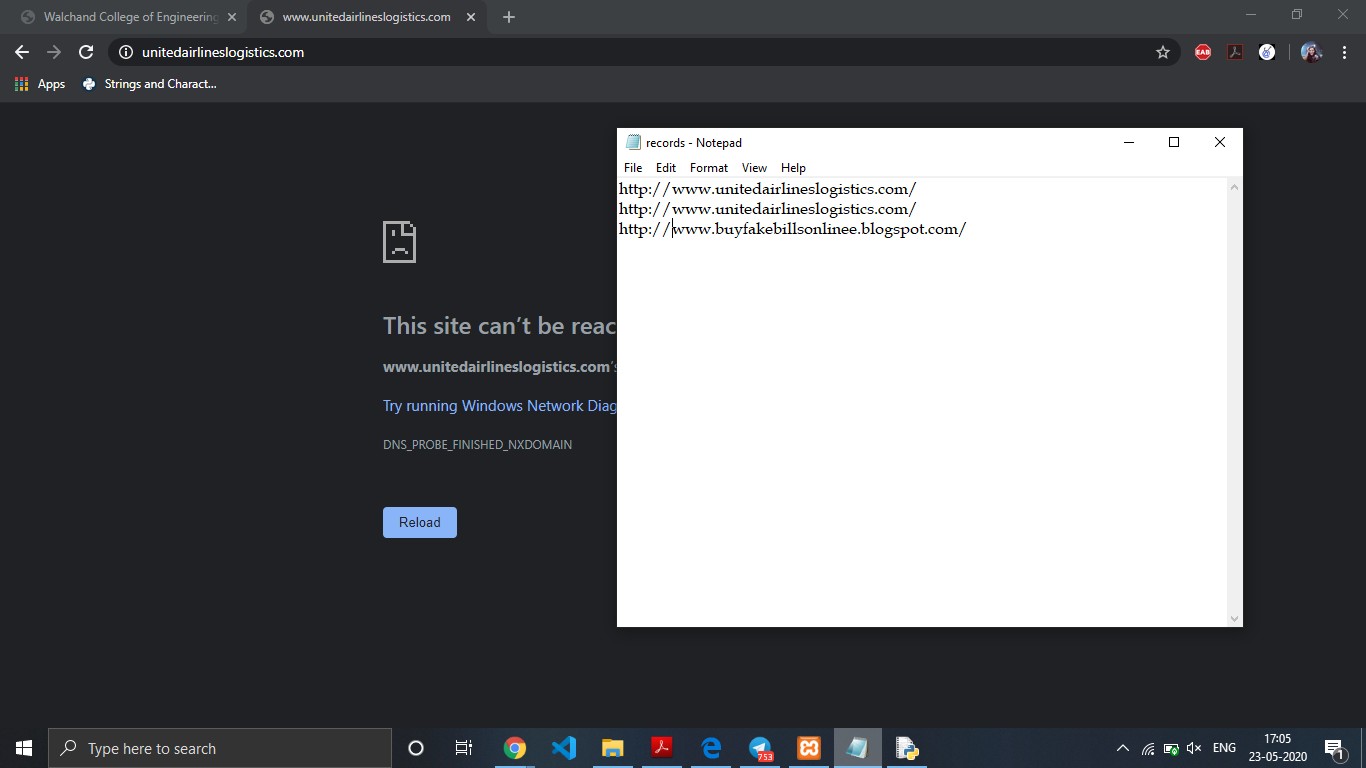
Phishingisawaytoobtainauser'sprivateinformationviaemailor website.Asusageoftheinternetisveryvast,almostallthingsareavailable online now it is either about shopping clothes, electronicgadgets,crockeryortopaymentofmobile,TV&electricitybills.Ratherthanstandingoutinlineforhours,peoplearebeingawareofusing onlinemethods.Duetothisphisherhaswidescopetoimplementphishingscam.Asthereisalotofresearchworkdoneinthisarea,thereisnotanysingletechnique,whichisenoughtodetectalltypesofphishing attack.Astechnologyincreases,attackersusenewmethodsdaybyday.Thisenablesustofindeffectiveclassifiertodetectionofphishing.

The following are some screenshots of the system:



Checking Google Drive Website



Checking Phishing Website (Taken from FakeWebsite Dataset)along with notification

Visited phishing website added to records file