Literature Survey on

# SMART MALICIOUS URLS DETECTION SYSTEM TO PREVENT PHISHING USING DEEP LEARNING APPROACH

Fourth Year of Engineering
In
Computer Engineering
By


| Name | Class | Roll No. |
|---|---|---|
| Jainam Soni | BE-4 | 24 |
| Palak Nisar | BE-3 | 66 |
| Shamika Dumbre | BE-3 | 29 |
| Siddhi Sheth | BE-4 | 15 |


**Under the Guidance of**

Prof. Deepti Nikumbh



DEPARTMENT OF COMPUTER ENGINEERING

**SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE**

CHEMBUR, MUMBAI-400088.


2018-2019

# 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.

# Table of Contents

# 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.

# 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.

- "A lightweight anti-phishing scheme for mobile phones" [5].

There is an another technique that is developed which focuses on detecting phishing attacks but on mobile phones using anti-phishing scheme on mobile based platforms. It was a challenging task to apply this technique because mobile phone and users have more limitation with them.

- "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 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.

# 3. Summary

As we know day by day the use of internet is growing rapidly and so it's effects. Present day all the work is done over internet may it be booking reservations or shopping or bank transactions. Due to lack of knowledge of security and privacy, huge amount of cyber crime and security attacks are prone to this. Phishing is considered as the most easiest way to attack user and steal it's data. To avoid this many Url detection systems are built using blacklist and whitelist data, using feature extraction and training them on traditional machine learning algorithm. But both of them are not good enough as first method identifies malicious urls which are only blacklisted and entered earlier in the list whereas second method extracts specific lexical and host based features and are classified using KNN, SVM, linear regression and so on, but all these classifiers accuracy is around 90%.

But we can overcome these using neural networks to built a detection system which can give us an accuracy above 95% and removes manual feature selection process by using RNN i.e specifically Long Short Term Memory (LSTM) by embedding Url using word2vec and creating a feature set for it. This approach uses a character sequence pattern to create the feature set. By tuning certain parameters we can even achieve an accuracy of almost 98%. Future scope for this could be creating an chrome extension or API service for this system.

# 4. 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] 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

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

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

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

(IEEE: 2016)-MobiFish:

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

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

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

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

[10] Common Crawl: www.commoncrawl.org

[11] Phishtank: www.phishtank.com: 2017.