Synopsis Report on

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

Submitted in partial fulfilment of the requirements

of the degree of Bachelor in Engineering

by

Name	Contact No.	Email Id
Jainam Soni	9773699390	jainam.soni@sakec.ac.in
Palak Nisar	9167267216	palak.nisar@sakec.ac.in
Shamika Dumbre	8976507132	shamika.dumbre@sakec.ac.in
Siddhi Sheth	9167056231	siddhi.sheth@sakec.ac.in

Under the Guidance of

Prof. Deepti Nikumbh

DEPARTMENT OF COMPUTER ENGINEERING

SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE

CHEMBUR, MUMBAI-400088.

2018-2019

Synopsis Report Content

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.

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. Related Work

Sr. No.	Author/Year	Work done/Algorithm/Concept/Idea presented in the paper	Remarks
1	Mohammed Al-Janabi, Ed de Quincey, Peter Andras / 2017	“Using supervised machine learning algorithms to detect suspicious URLs in online social networks”	It helps to detect all malicious URLs on OSN using Twiiter api as dataset and virus total for labelling it. It is then implemented using Random Forest which gives an accuracy of around 88% without parameter tuning.
2	Lonfei WC / 2016	“A lightweight anti-phishing scheme for mobile phones”	It is an detection system implemented for low memory mobile phones.
3	Frank Vanhoenshoven, Gonzalo Napoles, Koen Vanhoof / 2016	‘’Detecting Malicious URLs using machine learning techniques’’	In these paper all lexical and host based features are extracted and then applied on traditional ML classifiers.
4	Ankit Kumar Jain, B.B Gupta / 2016	“A novel approach to protect against phishing attacks at client side using auto-updated white-list”	It contains a list of malicious and benign url and can detect only those urls which are present in the list.

3. **Problem Definition and Objectives**

- Problem Definition

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.

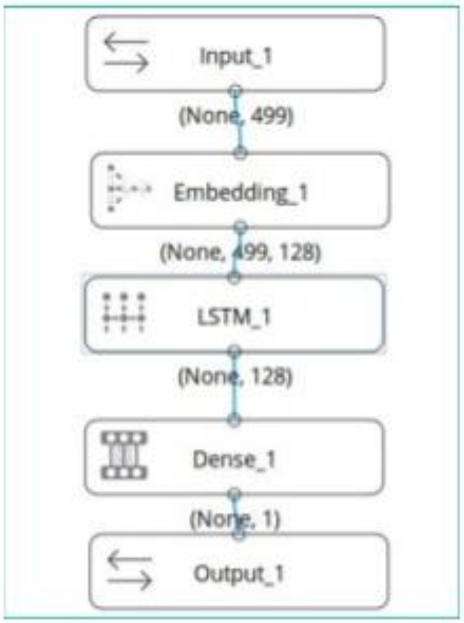
- Objectives

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.

4. Methodology

- ## Block Diagram
  - **Layers used in Model.**



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

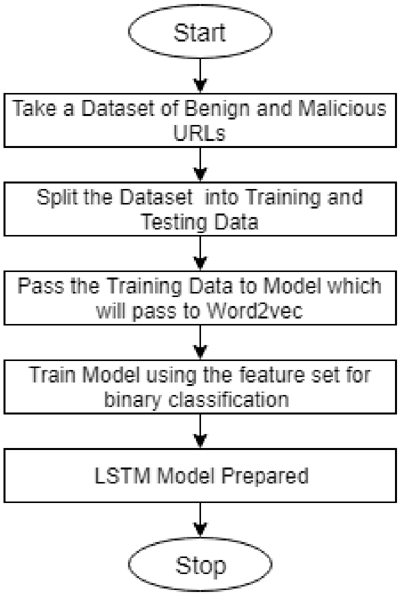
Embedding_1: Turn positive integers (indexes) into dense vectors of fixed size, with dropout rate **0.2**.

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

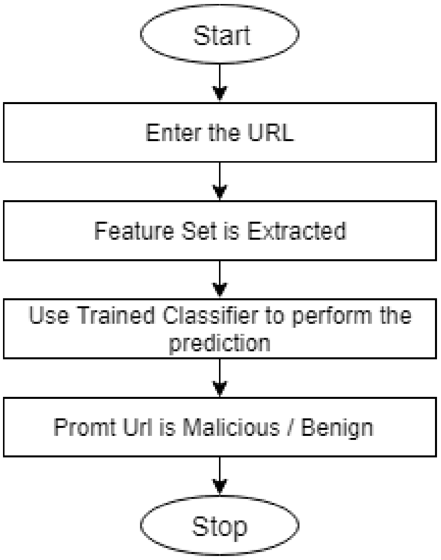
Dense_1: Just your regular densely-connected nn layer. Activation function **sigmoid** is used.

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

- ## Flowchart


Start
Take a Dataset of Benign and Malicious URLs
Split the Dataset into Training and Testing Data
Pass the Training Data to Model which will pass to Word2vec
Train Model using the feature set for binary classification
LSTM Model Prepared
Stop

Training Flowchart


Start
Enter the URL
Feature Set is Extracted
Use Trained Classifier to perform the prediction
Promt Url is Malicious / Benign
Stop

Smart Url Detection System

5. 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.

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.