Mahavir Education Trust's **Shah & Anchor Kutchhi Engineering College,** Chembur, Mumbai 400 088
UG Program in Computer Engineering

Smart Malicious Url's Detection System To Prevent Phishing Using Deep learning Approach

Academic Year: 2018 – 2019

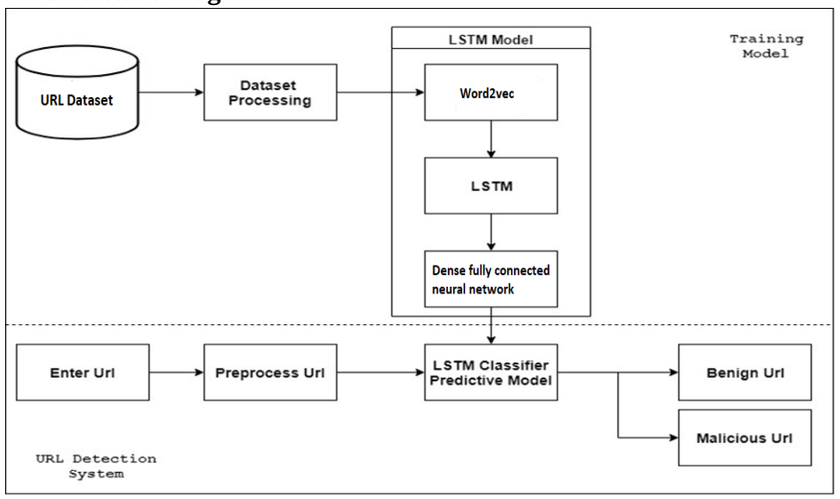
Abstract

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).

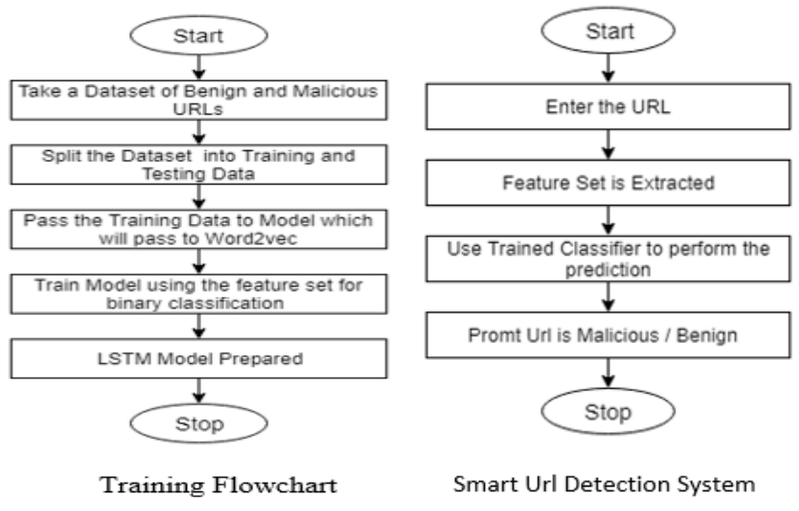
Introduction

Smart Malicious Urls Detection System is an anti-phishing technique to safeguard our web experiences. One 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%. We can use 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 95%. Future scope for this could be creating an chrome extension or API service for this system.

Architecture Diagram:



Activity Flow:



References

1) Qiongxia Huang, Xianghan Zheng, Riqing Chen and Zhenxin Dong, 'Deep Sentiment Representation Based on CNN and LSTM', IEEE 2017.
2) Mohammed Al-Janabi, Ed de Quincey, Peter Andras, "Using supervised machine learning algorithms to detect suspicious URLs in online social networks", ACM 2017.

LSTM Algorithm

1) Data collection of 80% and 20% benign and malicious url's.
2) Preprocessing raw Url's.
   i) Building a dictionary and converting into integer.
   ii) Url's cropping and padding with zero.
3) Input Layer where you define the initial input shape (here initial " n " characters of the URL).
4) Embedding Layer, where the input will be max vocabulary length i.e " n " as input dimension and output will be a embedded dimension.
5) Lstm using keras with Tensorflow as backend.
6) The Dropout layer is a trick used in deep learning to prevent overtraining and it precedes a Dense layer (fully connected layer) of size 0.5 or 1.
7) The sigmoid activation function to squash the output of this layer between 0 and 1.
8) Use Adam optimizer and metrics accuracy to see the efficiency of the model.

Conclusion

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

i) The GUI of our system will engage end users and provide user friendly experience with a (.exe) setup file.
ii) System will be based on discerning Urls by their patterns.
iii) Automatic Feature Extraction using Neural Network.
iv) Future scope is to create an Google Chrome extension.

Project Group

1. Jainam Soni **2.** Palak Nisar
3. Shamika Dumbre **4.** Siddhi Sheth

Under the Guidance of
Ms. Deepti Nikumbh