Project Batch No: 2

Student USN	Student Name	Mobile No	Email ID
1PE17CS023	Ankita Aditya	9801306182	ankita.aditya20@gmail.com
1PE17CS029	Atul K Uchil	9632691907	atulkuchil@gmail.com
1PE17CS182	Mitali Singh	9740569216	mitali281099@gmail.com
1PE18CS419	Prescilla Angel	8495917052	sombathiniprescilla8@gmail.com

Analysis of malicious URLs using Machine Learning

Abstract: Nowadays, the WEB has become the highest priority issue in the field of cybersecurity, giving platform for various online criminal activities. Massive online social networks like Facebook, Twitter, etc. with hundreds of millions of active users are increasingly being used by Cybercriminals to spread malicious URLs, which exploit vulnerabilities on the user's machine for personal gain. URLs are used as the main vehicle in this domain. To tackle this major issue, the community is mainly focused on techniques for blacklisting the malicious URLs. The detection of malicious URLs is one of the highest priority issues for cybersecurity practitioners. There are plenty of machine learning techniques to address this issue, the most used approach remains blacklisting. The main obstacle to using machine learning is the difficulties in data collection. We are building a model that can overcome the obstacles and create a proactive system for the effective and efficient detection of malicious URLs.

Keywords: malicious URLs, machine learning techniques, lexical-based features, cybersecurity

Introduction: Malicious Web sites are a cornerstone of Internet criminal activities. As a result, there has been broad interest in developing systems to prevent the end-user from visiting such sites. The main notion of our project is to identify malicious URLs based on certain lexical features. The project model analyzes the lexical-based feature of malicious URLs. It shows that lexical analysis is effective and efficient for the proactive detection of malicious URLs. We often click some URLs and after visiting the page, we got to know that it was a malicious URL. To avoid this

issue, we are incorporating the static model as a chrome extension, wherein users can add the extension to their system. Once the extension is added, and if the user clicks on the malicious URL, the extension throws an alert message that the URL is malicious. This helps the user to know prior that whether the URL, he/she is visiting is malicious or safe to proceed. The complete model provides the set of sufficient features necessary for optimal & accurate categorization and evaluates the accuracy of the technique over thousands of URLs.

Literature Survey (Minimum 5 papers –Published between 2016 and 2020):

1) *Detection of Malicious URLs using Machine Learning Techniques* - Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma – IJITEE - **March 2019**
   - The proposed methodology in this paper, not only takes care of the syntactical nature of URL, but also the semantic and lexical meaning of dynamically changing URLs.
   - The source was taken from Phishtank data (an opensource that allow the registered users to add new malicious URLs that are not in the existing one).
   - The classification model used was Convolutional Neural Network (CNN) which provided better results because of the effective learning rate and quite suitable for feature extraction. The classifying method was based on the TF - IDF word association.
   - Future Scope: For the handling of a huge number of URLs whose feature set will evolve over time, certain efforts have to be done so as to come up with a robust feature set which can change with respect to evolving changes.

2) *Empirical Study on Malicious URL Detection using Machine Learning* – Ripon Patgiri(B) , Hemanth Katari(B) , Ronit Kumar(B) , and Dheeraj Sharma - ICDCIT - **January 2019**

   In this paper, the malicious URLs detection is treated as a binary classification problem and performance of several well-known classifiers are tested with test data. The algorithms Random Forests and support Vector Machine (SVM) are studied in particular which attain a high accuracy. These algorithms are used for training the dataset for classification of good and bad URLs. The dataset of URLs is divided into training and test data in 60:40, 70:30 and 80:20 ratios. The detection of malicious URLs is a binary classification problem and several Machine Learning Algorithms, namely Random Forests, SVMs and Naive Bayes are implemented on training dataset. Also, it has been seen that the Random Forest classifier performs better for the particular problem than the SVM classifier.

3) *A Machine Learning Approach for Detecting Malicious Websites using URL Features* – Akshay Sushena Manjeri, Kaushik R, Ajay MNV, Priyanka C. Nair – IEEE - **June 2019**
   - This paper proposes a method to classify an URL as either malicious or benign by handling class imbalance.
   - The data source was obtained by a process that included different sources of benign and malicious URL.
   - In pre-processing, Mode technique was used to handle the missing values. SMOTE was adopted to handle the class imbalance.
   - Classification included five models namely, Random Forest, Decision Tree, KNN, Logistic Regression, and SVM.
   - This paper also used a feature selection named Recursive Feature Elimination based on which features were ranked and Association Rule Mining algorithms such as Apriori, FPGrowth, and Decision Tree Rules are used to generate rules which helps to establish the relationship among the features.
   - Future Scope: This paper can pave way for a plug-in to be developed for a web browser. In this way, the user can be warned about malicious website URLs in real-time while browsing the internet.

4) *Chrome Extension For Malicious URLs detection in Social Media Applications Using Artificial Neural Networks And Long Short Term Memory Networks* - Shivangi S, Pratyush Debnath, Sajeevan K, D. Annapurna – IEEE - **September 2018**

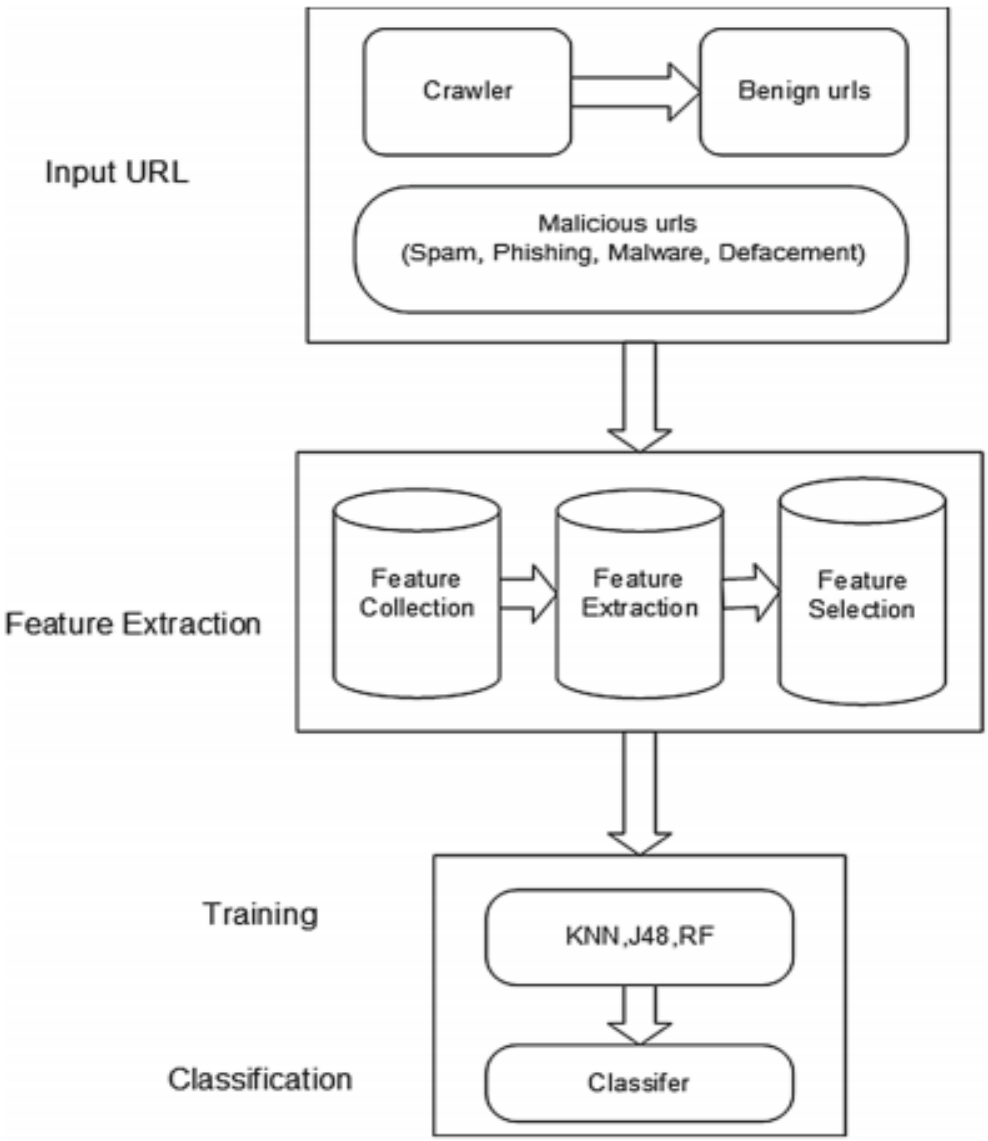
   In this paper, they propose a method to overcome the problem of malicious URLs victimizing users. They propose a tool deployed as a chrome extension. This tool, analyses URLs and classifies them using two different neural networks, Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks which is a specific type of Recurrent Neural Network (RNN). The major objective of the proposed model is to aid the users to avoid becoming a victim of malicious and fraudulent activities like malicious URLs, phishing, and social engineering that favor social media as their target medium by detecting them accurately.

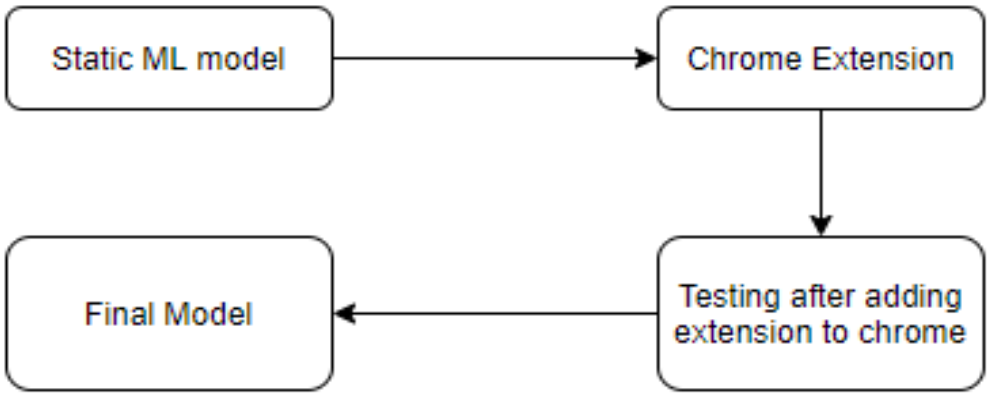
5) *Detecting Malicious URLs using Lexical Analysis* - Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker - NSS - **September 2016**

   Malicious Web sites are a cornerstone of Internet criminal activities. As a result, there has been broad interest in developing systems to prevent the end-user from visiting such sites. In this paper, they have described an approach to this problem based on automated URL classification, using statistical methods to discover the tell-tale lexical and host-based properties of malicious Web site URLs. These methods are able to learn highly predictive models by extracting and automatically analyzing tens of thousands of features potentially indicative of suspicious URLs. The resulting classifiers obtain 95-99% accuracy, detecting

large numbers of malicious Web sites from their URLs, with only modest false positives.

Proposed Methodology/Architecture diagram:





Guide Name: Prof. Sudeepa Roy Dey

Date: 24-10-2020