Detecting Malicious URLs via a Keyword Based Convolutional Gated Recurrent Unit Neural Network

Anitha C

PG Scholar, Department of Computer Science & Engineering
Sri Shanmugha College of Engineering and Technology, Sankari,TN
Email : canithabe@gmail.com

Abstract -- The world wide web is an inevitable part of today's fast moving world as a means of swift communication as well as for knowledge gathering and sharing. Though it is bestowed with those interesting features it also comes with the worst face being a ground for many illegal and serious threats like identity theft and illegal money launder in various forms. Hackers find innumerable ways day by day to steal the personal data of the legitimate users and use it deceptively for daring underground activities. The users are even unaware of those suspicious activities many a times and fall as a prey for many cyber crimes. Malicious URLs host unsolicited contents like spam, phishing, drive-by exploits, and attract unsuspecting users to become victims of scams causing monetary loss in a huge magnitude. To detect such crimes systems should be vigorous in detecting the new malicious contents. Traditionally, this detection is done mostly through various methods like Black list, SVM(support Vector Machine),Random Forest(RF) and LR(Logistic Regression). But they lack the ability to detect newly generated malicious URLs. We propose an advanced Convolutional Gated Recurrent Neural Network which detects any kind of URL. We add a character-level embedding which acts as a hidden layer before the convolution layer.With various features selected in the query string like length of the URL, Presence of redirection symbol or any special characters, subdomain existence, age of domain and other statistical report the model accurately predicts whether the URL is benign or suspicious.The test results demonstrate that our model has higher performance metrics like Accuracy, Precision, Recall compared with previously existing detection techniques

Keywords -- Cyber crimes, malicious URL , Convolutional Gated Recurrent Neural Network.

I. INTRODUCTION

With the undeniable prominence of the World Wide Web as the paramount platform supporting knowledge dissemination and increased economic activity, the security aspect continues to be at the forefront of many companies and governments' research efforts. The web has been used as a hub for a variety of malicious activities from malware hosting and propagation to phishing websites' tricking users to provide their personal user information. URL is the global address of documents and other resources on the World Wide Web. A URL has two main components as follows.

(i) Protocol identifier (indicates what protocol to use)
(ii) Resource name (specifies the IP address or the domain name where the resource is located).

The protocol identifier and the resource name are separated by a colon and two forward slashes as shown in the below figure:

Malicious URL or malicious website, is a common and serious threat to cyber security. They act as a gateway for the unsolicited activities hosting a variety of unsolicited content in the form of spam, phishing in order to launch attacks. Unsuspecting users visit such web sites and become victims of various types of scams, including monetary loss, theft of private information (identity, credit-cards, etc.) and ransomware installation on the user devices resulting huge loss globally. According to the latest Google Safe browsing report, Google search blacklisted over 50,000 malware sites and over 90,000 phishing sites monthly. The human understandable URLs are used to identify billions of websites hosted over the present day internet. Adversaries who try to get unauthorized access to the confidential data may use malicious URLs and present it as a legitimate URL to naive user. Such are called as malicious URLs. With the advancement of social networking platforms, many allow its users to publish the unauthorized URLs. Many of these URLs are related to the promotion of business and self-advertisement, but some of these unprecedented resource locators can pose a vulnerable threat to the naive users. The naive users who use the malicious URLs, are going to face serious security threats initiated by the adversary. The verification of URLs is very essential in order to ensure that user should be prevented from visiting malicious websites. Many mechanisms have been proposed to detect the malicious URLs. One of the basic feature that a mechanism should possess is to allow the benign URLs that are requested by the client and prevent the malicious URLs before reaching the user. This is achieved by notifying the user that it was a malicious website and a caution should be exercised. To achieve this, a system should take semantic and lexical properties of every URL rather than relying on syntactic properties of the URLs. There are many filtering mechanisms to detect the malicious URLs. These are broadly classified into 2 groups:

Non-Machine Learning approach:

Techniques such as Black – Listing, Heuristic Classification etc. comes under Non-Machine Learning approach. These traditional mechanisms rely on keyword matching and URL syntax matching. Therefore, these conventional mechanisms cannot effectively deal with the ever evolving technologies and web access techniques. Furthermore, these approaches also

fall short in detecting the modern URLs such as short URLs, dark web URLs. While URL blacklisting has been effective to some extent, it is rather easy for an attacker to deceive the system by slightly modifying one or more components of the URL string. Inevitably, many malicious sites are not blacklisted either because they are too recent or were never or incorrectly evaluated. Many of these web-based companies like Google,Facebook use exhaustive data bases which can store as many as millions of URLs, and refine these URL sets regularly. But this is not the feasible solution to all the problems. Despite having the greater accuracy, the need for human intervention to update and maintain the URL list is one of the major limiting factors in this method.

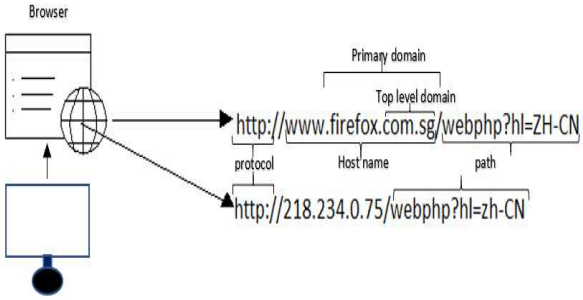

Fig. 1 Components of a URL

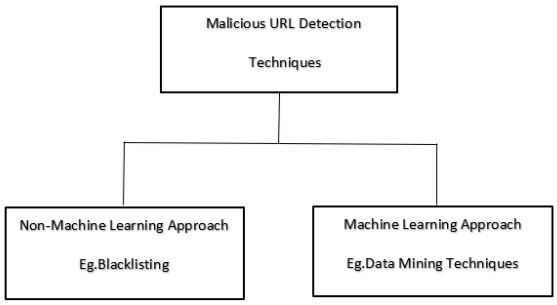

Fig. 2 Types of Malicious URLs Detection

Machine Learning approach:

Machine learning techniques are used to classify malicious websites through features taken from URLs, web content and network activity. Machine Learning approaches, use a set of URLs as training data, and based on the statistical properties, learn a prediction function to classify a URL as malicious or benign. This gives them the ability to generalize to new URLs unlike blacklisting methods. The primary requirement for training a machine learning model is the presence of training data. In the context of malicious URL detection, this would correspond to a set of large number of URLs.

Machine learning can broadly be classified into supervised, unsupervised, and semi-supervised, which correspond to having the labels for the training data, not having the labels, and having labels for limited fraction of training data, respectively. Labels correspond to the knowledge that a URL is malicious or benign. After the training data is collected, the next step is to extract informative features such that they sufficiently describe the URL and at the same time, they can be interpreted mathematically by machine learning models. For example, simply using the URL string directly may not allow us to learn a good prediction model (which in some extreme cases may reduce the prediction model to a blacklist method).Thus, one would need to extract suitable features based on some principles or heuristics to obtain a good feature representation of the URL. This may include lexical features (statistical properties of the URL string, bag of words, n-gram, etc.), host-based features (WHOIS info, geo – location properties of the host, etc.), etc. The detection methods and tools which adopt the approach of patrolling web content may consume more computation time and resource. Therefore, URL based detection techniques for malicious URL detection are largely limited to classification of URLs in general or any specific attack i.e. spam [3, 6, 20].

Meanwhile research shows that the characteristics of malicious URLs differ with the type of technique used for exploitation (e.g., spam, adware, phishing, drive-by-downloads etc.) The features after being extracted have to be processed into a suitable format (e.g. a numerical vector), such that they can be plugged into an off – the – shelf machine learning method for model training. The ability of these features to provide relevant information is critical to subsequent machine learning, as the underlying assumption of machine learning (classification) models is that feature representations of the malicious and benign URLs have different distributions. Therefore, the quality of feature representation of the URLs is critical to the quality of the resulting malicious URL predictive model learned by machine learning. In this paper we adapted machine learning techniques to the detection and categorization of the malicious URLs. We will use CNN in order to detect whether the given URL is malicious or benign. Identification of attack types is also useful since the knowledge of the nature of a potential threat allows us to take a proper reaction as well as a pertinent and effective countermeasure against the threat. For example, we may conveniently ignore spamming but should respond immediately to malware infection.

II. RELATED WORKS

Ripper Algorithm

Sonika Thakur, Er. Meenakshi and Akansha Priya showed that the rule based classifier model of RIPPER algorithm can identify URLs with an accuracy of 82%. They predicted 1050 URLs of testing dataset with rule based classifier model which is generated by training dataset of 12000 URLs, out of which 561 URLs are detected as malicious and 320 URLs are detected as legitimate. After this they calculated the accuracy of the generated rule based classifier model. The result showed that the rule based classifier model of RIPPER algorithm can identify URLs with an accuracy of 83%. They also stated that if a training dataset is of large dataset then, optimized result can be obtained.

Machine Learning Techniques

Immadisetti Naga Venkata Durga Naveen, Manamohana K and Rohit Verma presented that how a machine can able to judge the URLs based upon the given feature set. They described the feature sets and an approach for classifying the given feature set for malicious URL detection. When traditional method fall short in detecting the new malicious URLs on its own, their proposed method can be augmented with it and is expected to provide improved results. They also proposed the feature set which can able to classify the URLs

A Survey On Various Machine Learning Techniques

Doyen Sahoo, Chenghao Liu and Steven C.H. Hoi they conducted a comprehensive and systematic survey on Malicious URL Detection using machine learning techniques. In particular, they gave a systematic formulation of Malicious URL detection from a machine learning perspective, and then detailed the discussions of existing studies for malicious URL detection, particularly in the forms of developing new feature representations, and designing new learning algorithms for resolving the malicious URL detection tasks.

In this survey, they categorized most of the existing contributions for malicious URL detection in literature, and also identified the requirements and challenges for developing Malicious URL Detection as a service for real-world cyber security applications. They also highlighted some practical issues for the application domain and indicated some important open problems for further research investigation.

They mentioned that despite the extensive studies and the tremendous progress achieved in the past few years, automated detection of malicious URLs using machine learning remains a very challenging open problem. Future directions include more effective feature extraction and representation learning (e.g., via deep learning approaches), more effective machine learning algorithms for training the predictive models particularly for dealing with concept drifts (e.g., more effective online learning) and other emerging challenges (e.g., domain adaption when applying a model to a new domain), and finally a smart design of closed-loop system of acquiring labelled data and user feedback (e.g., integrating an online active learning approach in a real system).

Lexical Analysis

Mohammad Saiful Islam Mamun, Mohammad Ahmad Rathore, Arash Habibi Lashkari, Natalia Stakhanova and Ali A. Ghorbani explored an approach for classifying This technique acts as an addon for the blacklist techniques, in which new malicious URLs cannot be identified and will be efficient for analysing large number of URLs. Selected feature sets applied on supervised classification on a dataset yields a classification accuracy of 97 % with a low false positive rate.

Their experiment can also be helpful to improve classifier accuracy. In addition, it can be extended to calculate risk rating of a malicious URL after parameter adjustment and learning with huge training data. Despite random forest classification accuracy is able to identify approx. 97% of the malicious or benign URL, by using proper SD filter they could reach up to around 99 % accuracy.

Drawbacks Of The Existing System

- High volume and high velocity
- Difficulty in acquiring labels
- Difficulty in collecting features
- Feature Representation
- Concept drifting and emerging challenges
- Interpretability of Models
- Adversarial Attacks

III. GENERALIZED MODEL

In the generalized model, prediction of the URL as malign or benign is done based on our proposed Convolutional Gated Recurrent Unit Neural Network. In this system, when the URL is fed to the browser, we are left with two cases:

Case 1: When the URL already exists in our blacklist, it will be qualified as malicious.
Case 2:If a new URL is encountered,then the following procedures are carried out sequentially
1.Data Preprocessing
2.Feature Extraction
3.Feature Selection
4.Prediction of URL

The predicted URLs are trained to the model as a basis of supervised learning.

The system architecture of the model is as displayed in the following figure.

The proposed CGRU model detects the malicious URL by combining the characteristics of URLs in the web attacks at the character level and based on statistical differences in historical data of legitimate domains and malicious domains, domains lifetime, changes of who is information, who is information integrity, IP changes, domains that share same IP, TTL value, etc. As main parameters and concrete representations of features for classification were given and on this basis

the proposed method constructed SVM classifier for detecting anomaly domains. Features analysis and experimental results show that the algorithm obtains high detection accuracy to unknown domains, especially suitable for detecting long lived malicious domains. The most of the existing approaches are feature based and cannot detect dynamic attacks. Mostly the attacker uses the input form, active content and embeds @ symbol in URL for malicious attack. To detect this attack a Behaviour based Malicious URL Finder (BMUF) algorithm is proposed. It analyses the behaviour of the URL. The FSM based state transition diagram is used to model the URL behaviour into various states. The state transition from initial to final state is used for classification. This approach tests the genuine and malicious behaviour of the URL based on the responses to the user. It accurately detects the nature of the URL.

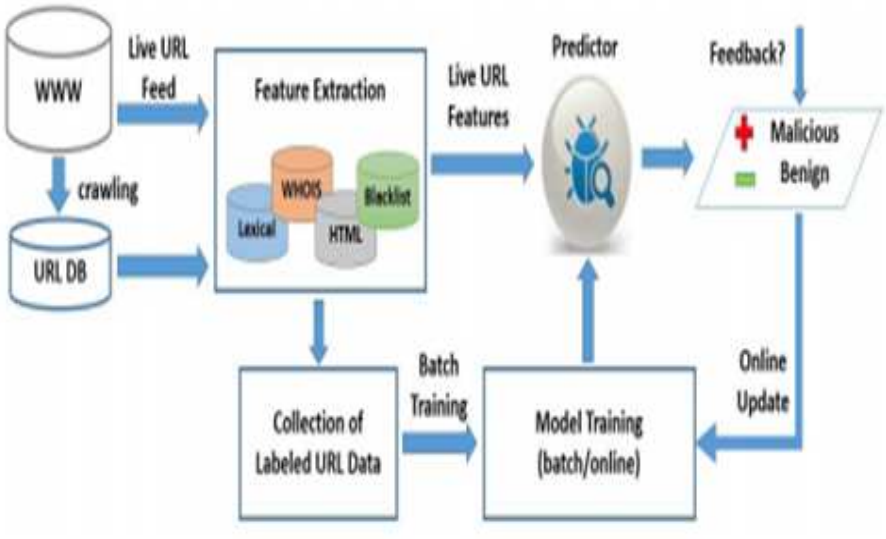

Fig. 3 The system Architecture

V. CONCLUSION

Malicious URL detection plays a pivotal role in many cyber security applications, and hence deep learning approaches proves to be a promising direction. In this paper the advanced CNN algorithm which is the Convolutional Gated Recurrent Neural Network based on the word2vec feature is used to classify the URL as safe or malicious which saves the user from being a prey for many phishing sites and cyber crime deeds. This algorithm proves to have 75 percent accuracy which we infer by comparing the three aspects precision, recall, and F1 –score of CNN with the already existing SVM and logical regression. The TPR (True Positive Results) of this advanced CNN model is 1 and FPR(False Positive Results) is just 0.5 and precision is 0.6 The later two methods fall behind in its performance metrics hence this advanced CGRU neural network model proves to perform better.

REFERENCES

[1] Abdi, F. D., & Wenjuan, L. Malicious URL detection using Convolutional Neural Network.

[2] Thakur, S., Meenakshi, E., & Priya, A., "Detection of malicious URLs in big data using RIPPER algorithm," In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1296-1301, IEEE. 2017

[3] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, "Detection of Malicious URLs using Machine Learning Techniques".

[4] Sahoo, D., Liu, C., & Hoi, S. C., "Malicious URL detection using machine learning: A Survey," arXiv preprint arXiv:1701.07179, 2017.

[5] Mamun, M. S. I., Rathore, M. A., Lashkari, A. H., Stakhanova, N., & Ghorbani, A. A., "Detecting malicious urls using lexical analysis," In International Conference on Network and System Security, Springer, Cham. pp. 467-482, 2016.

[6] Vanhoenshoven, F., Nápoles, G., Falcon, R., Vanhoof, K., & Köppen, M., "Detecting malicious URLs using machine learning techniques. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-8, IEEE, 2016.