International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 30 Issue 2 – MARCH 2023.

MACHINE LEARNING METHODS OF SLEUTHING MALEVOLENT WEB CHANNELS

JOHN WILLIAM P[#1] and SIVAPRAKASH T[*2]

[#1]*Assistant Professor, Department of Computer Science and Engineering, CK College of Engineering and Technology, Cuddalore, India*

[*2]*Assistant Professor, Department of Computer Science and Engineering, CK College of Engineering and Technology, Cuddalore, India*

Abstract— ***The* Internet has become an inevitable part of humans living in 21 st century. It has eased our lives in such a way that one could do almost anything from paying bills to placing orders or to sneak onto social media or to take part in an e-auction sitting relaxed on the living room's couch. As the saying goes on "Too much of anything is good for Nothing", The extravagant usage of the internet cast light on its lurking darker side posing serious perils to the users. The technology grows exponentially paving a way to collapse oneself bereaved of money even without his knowledge. Identity theft and fraudsters are two major traps that one has to be cautioned of in this transcendental world wide web. Malevolent web links are the malicious channels that may host unsolicited elements like a virus, malware, spyware, phishing traps. Recent Pegasus has the potential to tap the users that don't require even a click or download from the user's end. The user's complete Message history, call history, App usages, and the most sensitive financial transactions are revealed to the perpetrator. In this paper, we study various machine learning approaches towards the detection of malevolent URLs that have been proved to be effective when compared to the non-machine learning approaches that have been existing conventionally.**

***Index Terms— Cyber Security, Machine learning, classification, logistic regression, support vector machine, TF-IDF Classification, Deep Learning, Neural Network, Perceptron, Convolutional neural networks*.**

1. INTRODUCTION

The World Wide Web with its paramount benefits finds its implication in almost all the fields like Science, Engineering, Medicine, etc owing to its steadfast development. Accessibility of sensible and confidential data is no more a big deal for intruders. They employ many techniques like hosting malware and using phishing strategies to steal the user credentials even without their knowledge. They victimize the users with attractive advertisements wherein the moment the users click the link, the entire system functionality will be tapped by the perpetrators. The security breach is the most dominant prevailing issue with the advent of the internet. Also, the recent trends that make use of cloud provisioning in the way of IAAS, PAAS, and SAAS have eased the miscreants to exploit the users' data even without investing for their resources to instigate the online theft.

For the past year, Kaspersky's detection systems have found lakhs of malicious files-nearly about 3,80,000/day which was just 20,000/day for the past year. Also, 91% of threats were assimilated via Windows PE. In this current era, Cybercriminals have started infecting LINUX OS leading to 57% growth in Linux malware and other unwanted software. Subsequently, TROJAN DROPPERS grew by 2%. Also, it has been found that 54% of the threats detected by Kaspersky's systems were unspecified Trojans. While the most type of threats contributes to a decreased volume in 2021 as that of 2020, Trojan Droppers grew by 2.24% when compared to 2020. Trojans deploy sophisticated malware that proves to be a serious threat to the victim.

Worms were also grown by 117% depicting a significant sharp rise in its graph. Worms tend to replicate themselves after breaching a system and they propagate independently. Viruses have also grown by 27%.

2. URL

A URL (Uniform Resource Locator)is the address of a web page that has the components like protocol identifier, hostname, and pathname. A Protocol could be HTTP, HTTPS, FTP, etc which is followed by hostname and path that leads to the resource. The hostname could also be an IP address.

The protocol identifier and the resource identifier are separated by a colon and two forward slashes as given in Figure 2.1. This paper analyses various methodologies for classifying URLs as malevolent or not using machine learning approaches.

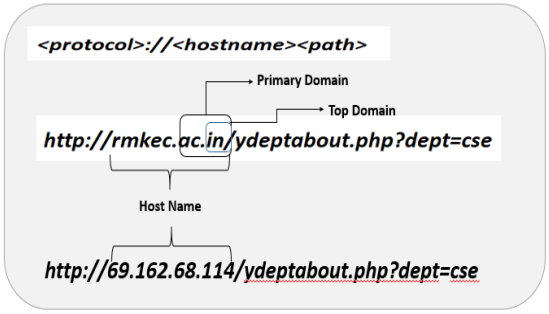

Figure 2.1 Components of a URL

MalevolentURLs are the gateway to malicious websites that pose serious threats to users in form of viruses like MorrisWorm, Nimda, Stuxnet. ILOVEYOU, Conficker, SQL, Slammer, CryptoLocker, Tinba, Trojan virus, adware, and spyware. Users tend to fall into those traps when they visit such websites and may incur monetary loss and identity theft wherein the complete personal details of a person like Aadhar, PAN, Voter ID details along with the Credit card information are stolen. The miscreants imitate the users and involve in various identity fraud.

Ransomware that gets installed in the user's PC leaves the system encrypted demanding lump some money to get back the data intact. Google's security team through its service called 'Safe Browsing' blacklists thousands of malware and phishing sites every month. Billions of websites hosted over the net are human-readable. The intruders play skilfully by making the malicious URL look like a legitimate one thereby tending to lure the victims.

With the advancement of social networking platforms, intruders can swiftly propagate unauthorized URLs thereby reaching multiple audiences at a time. Many of the URLs we encounter every day seem to be genuinely related to the promotion of business and self-advertisement, but some of these are unprecedented resource locators that can pose a vulnerable threat to naive users.

The naive users who click the malicious URLs, face serious security threats initiated by the adversary. A verification mechanism should exist for the user to perpetuate in a safe browsing environment.

Mechanisms that detect the malicious URLs should precisely block the suspicious links that may not be benign and allow otherwise. The user should be notified of the underlying threat and caution is exercised. This is achieved by taking semantic and lexical properties of the URL rather than relying on syntactical properties.

3. CONVENTIONAL NON-MACHINE LEARNING APPROACHES

URLs can be classified as Malicious or benign using Machine learning and Non-Machine learning approaches.

Conventionally URLs have been detected using the following mechanisms to name a few,

1. Blacklists
2. Regular Expressions
3. Signature Matching approaches

The above traditional techniques rely extensively on keyword matching and URL syntax matching and hence they are not effective due to the ever-changing nature of the URL patterns and the underlying complicate correlations among their features.

Eventually modern URLs such as short URLs, dark web URLs are failed to be trapped by conventional detection mechanisms. Blacklisting mechanisms fail as the perpetrator can easily deceive the system by changing the URL components.

This leads to a scenario where numerous malicious links bypass the blacklist validation. Millions of URLs are being updated and maintained as an exhaustive database by web giants like Google, Microsoft and Meta platforms Inc.

But this doesn't prove to be reliable or otherwise feasible when time and accuracy are significant constraints.

4. MACHINE LEARNING APPROACHES

Machine learning techniques for malicious URL detection depend on features of URL, web content, and activities in the Network. With selected features and statistical properties of the URL in the training dataset that contain 1000s of URLs, Machine Learning systems learn a prediction function to classify the URL as malevolent or benign. This gives the system ability to predict any kind of new URLs, unlike the conventional systems.
Machine learning can broadly be classified as

1. Supervised
2. Unsupervised
3. Semi-supervised.

The supervised Learning algorithm is provided with the training labels in pairs such as (x,y) which is a tuple comprising of input and the respective output whereas in an unsupervised learning algorithm labels are not provided and the training data is analysed by the ML model for similarities among the data.

Unsupervised learning models are termed unguided methodologies as the model analyses the data for similarities among them to group them as clusters.
The semi-supervised learning algorithm is midway between supervised and unsupervised learning methods. The input comprises labels for only a fraction of training data with the majority being unlabelled data. Similarities between the data are used to cluster it given unsupervised methods and the groups are then analysed with help of partially labelled data we have in the training set and it aids in labelling the rest of the tuples.
In this regard, features are extracted that describe the genuineness of the URL. Also, machine learning models that use mathematical interpretations are employed to classify the URL as malevolent or benign. Simply predicting with the help of the strings of URLs may alter the prediction model to a normal blacklist system which is inefficient due to the large accumulation of novice URLs every minute. Hence

articulating specific features that pinpoint the URL as malicious is becoming the need of the hour. It can be based on some heuristics or some principles.
In this paper, we study various machine learning techniques that have been existing for the detection and categorization of malicious URLs.

4.1 FEATURE ENGINEERING

The Feature engineering approach focuses on selecting the most potential features of URL that kindles the precise classification of URL as malicious or benign. Feature engineering includes the steps such as Indicator Variables, Interaction Features, Feature Representation, External Data, and Error Analysis.

W..Zhang [1] proposed a genetic algorithm that splits the URL's features as critical and non-critical and utilizes them as input for the detection model. He also used a 2 stage PP algorithm for extracting features on non-critical features. His focus was on phishing web pages detection which he validated with a dataset sourced from PhishTank and observed that LR, KNN, and NB outperformed with feature engineering.

Tie Li a, Gang Kou b, Yi Peng [2] proposed a linear transformation method using a 2 stage distance metric learning approach. Here an orthogonal space is attained by performing a singular value decomposition. The optimal distance metric is solved using linear programming. They also proposed a non-linear transformation wherein Kernel approximation is done with the Nystrom method and for radial basis function revised distance metric is used.

Features were selected based on Domain like TLD of the domain name. Register Days, Expire Days, Based on Hosts like CountrySponsor of Host and Technologies, Based on Reputation like Google Page Rank, Alexa Rank, Baidu Inverted Index, etc, and, based on lexical like URL Characteristics and Keyword.

This amalgamation of linear and non-linear transformation has resulted in superior results for classifiers like KNN, SVM, and MLP.

4.2 UE MODEL

Xiaodan Yan, Yang Xue, Baojiang Cui, Shuhan Zhang, Taibiao Guo, and Chaoliang Li [4] proposed an URL Embedding model called as UE Model. Among different domains, the correlations are analyzed and the coefficients of the URLs are calculated. The concern here is to choose a distributed representation of the URLs thereby attaining a low dimensional vector. Initially, a mapping between the URL and their distributed representation is maintained. Then Huffman coding is used to convert the URL into a Huffman Tree. With this tree, the Huffman code is generated for various domains. Finally based on Huffman code a distributed representation of the domain is achieved which is termed URL Embedding.

Positive samples	Negative samples
Google.com	1yb3mkw1vipc2qt1mv4qr3xcqf.org
Sina.com	1f3yeryza1ulk1vuyrdw1nek6dd.com
qq.com	egkcoc1oay5hij4j78qgo8fbk.net
Youtube.com	1cj5mni164n5xqx2hvjiuyzpvf.com
Baidu.com	1qoqlc84ov1ax11dyg3h1y5y2xt.com
Facebook.com	1rlyqqjmg96163qi2bcn6hzkx.org
Sohu.com	1wakafb1qxf5jpl3mhp510bghi2.com
Yahoo.com	1nj3ubrxjm9p317bdrm3dcul8x.org
Sina.com	1f3yeryza1ulk1vuyrdw1nek6dd.com
Amazon.com	14t5kg6184p31fpjzi8yss8dfq.org
Taobao.com	9vw9k51jl2kgdk5y69k1bj6121.org

Table 4.1.1 List of Samples

The positive and negative samples listed in table 1.1 are fetched from open datasets like 360NetLab and Alexa and it was found that URL Embedding used to extract the feature vectors proved to be efficient when compared to Feature Engineering classifier which can be inferred in Fig 4.2.1, 4.2.2, 4.2.3 for PrecisionRate, F1 score, and Recall.

Precision is the ratio of True Positive predictions to the summation of True Positive and False Positive predictions.

$$P = \frac{Tp}{Tp + Fp} \quad (1)$$

The recall is the ratio of True Positive predictions to the summation of True Positive and False Negative predictions.

$$R = \frac{Tp}{Tp + Fn} \quad (2)$$

The F1 score is the weighted average of precision and is given by

$$F_1 = 2 . \frac{P.R}{P+R} \quad (3)$$

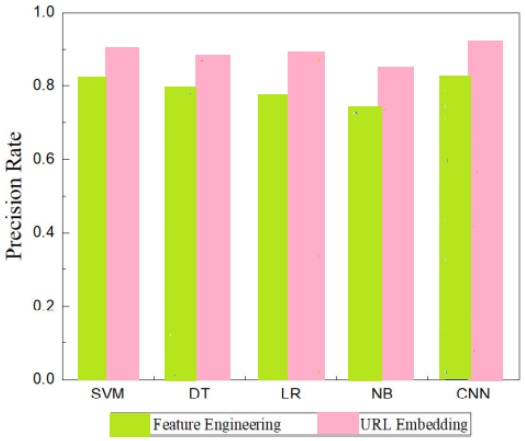

Figure 4.2.1 Precision Rate Comparison

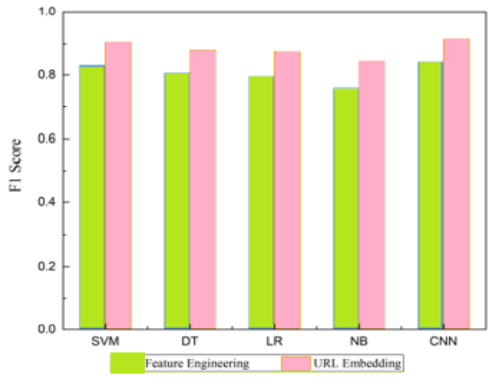

Figure 4.2.2 F1 Score Comparison

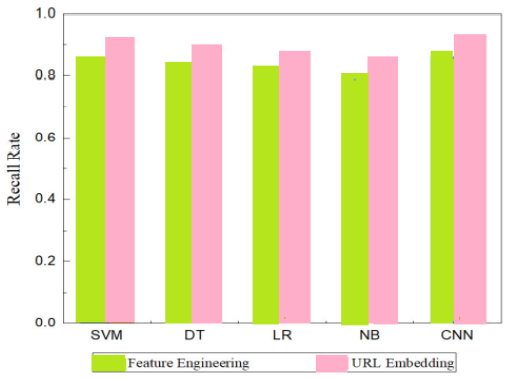

Figure 4.2.3 Recall Rate Comparison

4.3 CLASSIFICATION AND REGRESSION TECHNIQUES

The given data that is image/text can be grouped into distinct classes using various Machine learning algorithms. When only 2 classes are involved we term it as binomial classification and when it involves more than 2 classes we name it as Multi-class classification. The following section describes various forms of classifying URLs as malicious or benign based on various algorithms.

4.3.1 Logistic regression:

Logistic Regression learns the target variable and gives an inference on how the dependent variable is related to the independent variables. Here the sigmoid function is used to find the probability of the possible classes which is malicious or benign.

4.3.2 Perceptron:

Perceptron is a collection of McCulloch and Pitt's neurons and is known as a linear classifier. The yellow circles are input nodes, blue circles are neurons and green boxes are the threshold for comparison. The Perceptron fires when the multiplied product of weigh and input node exceeds the threshold.

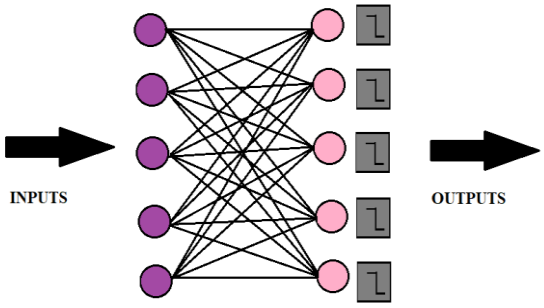

Figure 4.3.2.1 Perceptron

It works on Stochastic Gradient Descent optimization for adjusting the parameters on misclassification as given in the following formula.

$$w_{ij} \leftarrow w_{ij} + \eta(t_j - y_j) \cdot x_i \quad (4)$$

Where wij is the weight tj is the target variable and yj is the output and xi is the input. The variable i refers to the input node and j refers to the neuron. The model is made to fire on detection of malicious URLs. Whenever an URL is misclassified then the parameters are tuned by adjusting the weight through backward and relevant forward propagation.

4.3.3 Support Vector Machine:

In A Malicious Web Page Detection Model based on SVM Algorithm proposed by Jingbing Chen,Jie, Yuan,Yuewei Li,Yiqi Zhang, Yufan Yang,Ruiqi Feng [5] malicious URLs are detected with SVM - a kernel based method used for binary classification where the output is 0/1, Yes/No

4.4 MULTI LAYER RECURRENT CONVOLUTIONAL NEURAL NETWORK

Zuguo Chen, Yang long Liu, Chaoyang Chen, Ming Lu, and Xuzhou Zhang [3] detected Malicious URLs based on Improved MultiLayer Recurrent Convolutional Neural Network.

In order to process the character vector of the URL, a one-dimensional convolutional neural network is employed. In the training process of the model, Word embedding is used and the resulting word embedding space is used for the input URL to vectorize at a character level. An URL is transformed into a 2D tensor and hence based on the multilayer convolutional neural network model the feature extraction is carried out and this is improved by the YOLO algorithm. At last, the feature tensor extracted is fed as input into the bidirectional LSTM neural network. This detects the malicious URLs correctly. The experimental results will give an inference that the embedding layer used in the training

process of the entire model, the embedding space obtained has a relatively close relationship between the character vectors and at the same time has good representation ability.

The YOLO algorithm can thus effectively extract the features of the URL two-dimensional numerical tensor. Also employing a multilayer convolutional neural network, the dimensionality of the URL numerical tensor is reduced and hence the complexity of the model is also reduced successively.

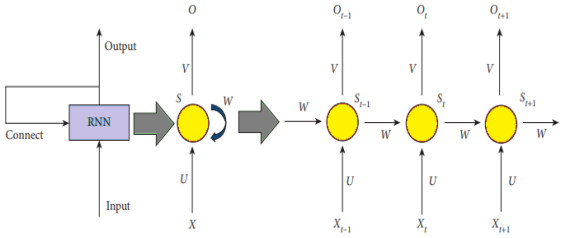

Figure 4.4.1 Recurrent Neural Network

The existence of the Memory function in the hidden layer is highlighting the feature of RNN. The hidden layers not only include the input of the current time step but also the output of the previous time step hidden layer just like relating to a memory. The RNN-specific network structure is shown in Figure 4.4.1.

5. CONCLUSION

As security is the elixir of internet prone world, gauging the URL as Malevolent or not help us in not getting victimized in the miscreant's trap. This paper gives the various methodologies on classifying the URL. As efficiency is concerned rather than dwelling on simple black list techniques complex techniques ranging from SVM to Convolutional neural networks with gated recurrent units proved to provide beneficial results in blocking higher percentage of malevolent links. In this work, we have depicted how a machine can able to judge the URLs based upon the given feature set. When conventional strategy drop brief in recognizing the ne w malicious URLs on its claim, our proposed strategy can be expanded with it and is anticipated to provide improved results. The future work is to fine tuning the machine learning calculation that will create the way for better result by utilizing the given include set.

REFERENCES

[1] Wei ZHANG Wei, Huan REN, Oingshan -Application of Feature Engineering for Phishing Detection

[2] Tie Li a, Gang Kou b, Yi Peng a - Improving malicious URLs detection via feature : Linear engineering and nonlinear space transformation methods.

[3] Zuguo Chen , Yanglong Liu ,Chaoyang Chen, Ming Lu , and Xuzhuo Zhang - Malicious URL Detection Based on Improved Multilayer Recurrent Convolutional Neural Network Model -

[4] Xiaodan Yan, Yang XuB, Member, IEEE, Baojiang Cui, Shuhan Zhang, Taibiao Guo, and Chaoliang Li - Learning URL Embedding for Malicious Website Detection

[5] Jingbing Chen,Jie Yuan,Yuewei Li,Yiqi Zhang, Yufan Yang,Ruiqi Feng - A Malicious Web page detection model based on SVM Algorithm