e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

MALICIOUS URL DETECTION USING MACHINE LEARNING AND DEEP LEARNING

Amogh J[*1], Rajendra KN[*2], Deekshith NG[*3], S Parikshith[*4], Asst. Prof. Suma L[*5]

[*1,2,3,4,5]Department Of Computer Science & Engineering Bangalore Institute Of Technology Bengaluru, India.

DOI : https://www.doi.org/10.56726/IRJMETS29486

ABSTRACT

The risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a result, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviours and attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviours. In short, the proposed detection system consists of a new set of URLs features and behaviours, a machine learning algorithm, and a bigdata technology. The experimental results show that the proposed URL attributes and behaviour can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.

Keywords: URL; Malicious URL Detection; Phishing; Machine Learning.

I. INTRODUCTION

Uniform Resource Locator (URL) is used to refer to resources on the Internet. In [1], presented about the characteristics and two basic components of the URL as: protocol identifier, which indicates what protocol to use, and resource name, which specifies the IP address or the domain name where the resource is located. It can be seen that each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or other phishing site, or malware download. Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs

Include: Drive-by Download, Phishing, Spam, and Defacement.

Malicious URLs are a dangerous threat to cyber security, these types of attacks can lead to scams, where people lose money, their information and accounts. It is important to be able to detect and act against these threats, the most conventional way is the use of blacklists, but this technique has many difficulties in acting against new URLs, so we are increasingly focused on machine learning algorithms, and that is precisely the focus of this project. The risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities.

These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. URL sharing is a core attraction of existing social media systems like Twitter and Facebook. Recent studies find that around 25% of all status messages in these systems contain URLs, amounting to millions of URLs shared per day. With this opportunity comes a challenge, however, from malicious users who seek to promote phishing, malware, and

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

other low-quality content. Indeed, several recent efforts have identified the problem of spam URLs in social media ultimately degrading the quality of information available in these systems. The Covid 19 has a great impact on the growth of on-line businesses such as e-banking, e-commerce and social networking. Unfortunately, the technological advancements accompany state of the art techniques to exploit users. Such attacks generally include malicious websites that steal all kinds of private information that a hacker can exploit.

APPLICABILITY

There are many applications for malicious URL detection in a variety of industries, including Protecting end users' financial information, Making the Internet Safer and More Secure, Browsers alerting users to fraudulent websites, Specialized spam filters can reduce the number of fraud emails that reach their addresses inboxes.

We show the Model which classifies benign and malicious URLs and display the evaluation result considering accuracy metrics and compare the obtained result for trained models and specify which is better.

II. LITERATURE SURVEY

In this study, our innovations and contributions are as follows: (1) This paper proposes a malicious URL detection model based on a DCNN. The dynamic convolution algorithm adds a new folding layer to the original multilayer convolution structure. It replaces the pooling layer with the k-max-pooling layer. In the dynamic convolution algorithm, the width of feature mapping in the middle layer depends on the vector input dimension. Moreover, the pooling layer parameters are dynamically adjusted according to the length of the URL input and the depth of the current convolution layer, which helps extract more in-depth features in a wider range. (2) In the stage of feature extraction and representation, the features are extracted from the URL sequence. We extracted features are integrated into a vector, and the vector is processed directly by the convolutional neural network to learn the classification model. This method not only simplifies the process of feature extraction, it does not depend on extracting features manually, but also combines the advantages of character embedding and word embedding. Word embedding can obtain word sequence information, which cannot be obtained by character embedding. Character embedding can process special characters and unfamiliar words in the URL. The dictionary and vector dimension are also not too big. The combination can save memory space and express the URL more effectively, which will help extract information from the URL. (3) To prove the feasibility of the model proposed in this paper, we did a lot of comparative experiments. As for the embedding method, we conduct three contrast experiments to verify that word embedding based on character embedding achieves higher accuracy than word embedding and character embedding. We also perform three contrast experiments and prove that leveraging the network structure consisting of a DCNN and different fields extracted from the URL can achieve a better effect.

At present, the methods of detecting malicious URL can be roughly divided into traditional detection methods based on blacklist and detection methods based on machine learning. Literatures introduce the detection method based on a blacklist. Although this method is simple and efficient, it cannot detect the newly generated malicious URL, which has severe limitations. Literature points out that attackers can generate various malicious domain names through a random seed to effectively evade the traditional detection method based on a blacklist. In literatures, researchers have applied machine learning technology to detect malicious URL. Machine learning learns the prediction model based on statistical properties and classifies a URL as a malicious URL or a benign URL. This method attempts to analyze URL and their relevant websites or web page information to extract the features. The features extracted by this method are often divided into two types, static features and dynamic features. It obtains lexical information in URL strings, information about hosts, and sometimes HTML and JavaScript content. It extracts a series of network traffic-related features from URL, and the support vector machine (SVM) is adopted for detection. Literature proposes three methods of feature processing to optimize the classification effect. While the above methods have shown good performance, there are still some limitations. Traditional detection methods based on machine learning often require extract features manually. Attackers can avoid being detected by these detection methods by designing these features, making it very difficult to maintain the detection system based on traditional machine learning. Additionally, in large-scale malicious URL detection, a trained model may lose some useful information from URL. Referring to the idea of text classification, many

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

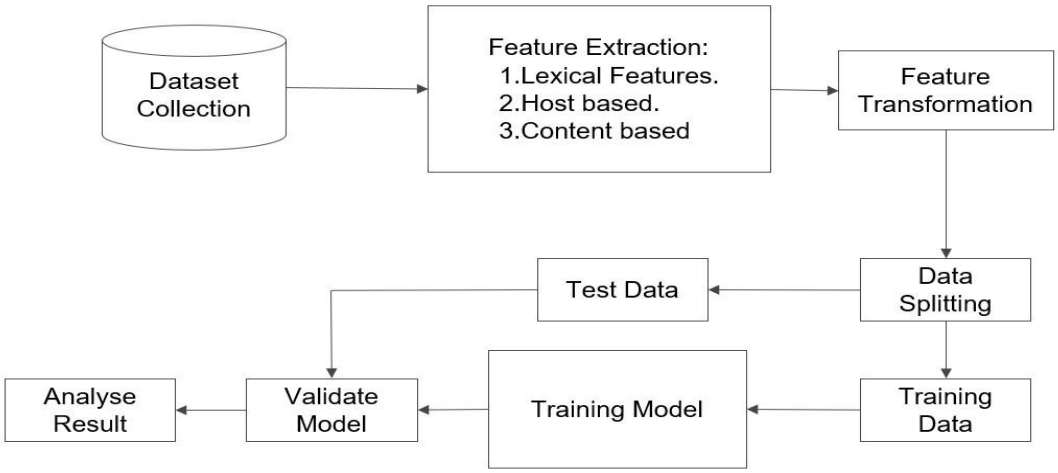
researchers have proposed a variety of methods based on deep learning models to detect malicious URL and judge whether a URL is malicious only according to the strings contained in the URL. These methods can automatically extract valid information in the URL. For example, literature uses the cyclic neural network model at the character level to classify URL generated by DGA. Literature proposes the method of extreme machine learning to detect malicious URL. Combining n-gram model with deep learning, literature takes the advantages of character level semantic features to detect whether DGA generates the URL. A variety of deep learning architectures for malicious URL detection are listed in the literature, including the structure of single-layer long short term memory(LSTM), the structure of bidirectional LSTM, the combined structure of CNN and LSTM, and the deep convolution structure.

III. PROPOSED METHDOLOGY

In order to complete this task, a CNN and an RNN are combined in the neural network architecture. The diagram for this is as follows: A sequence generator architecture like RNN or LSTM can start by converting an image into a feature vector of fixed length, which can then be used to generate a series of words or captions for the image. ResNet50 is the encoder that we have employed for the benefit of this project. The ImageNet Dataset's million images were divided into a thousand categories using a pretrained model. Since its weights are tuned to identify a lot of things that commonly occur in nature, we can use this net effectively by removing the top layer of 1000 neurons (meant for ImageNet classification) and instead adding a linear layer with the number of neurons same as number of neurons that you're LSTM is going to output. The RNN consists of a series of LSTM (Long Short-Term Memory) Cells which are used to recursively generate captions given an input image. These cells utilize the concept of recurrence and gates in order to remember information in past time steps. You can watch this or read this to understand more about the same. Eventually, the output from both encoder and decoder is merged and passed to a Dense Layer and finally an output layer which predicts the next word given our image and current sequence.

The proposed system is as follows:

- Our first option is the graphical user interface (GUI). The user engages with the system at this point.
- If a user is a first-time visitor, he or she must login or register
- After completing this, the user will have the choice to upload an image and receive a description of it.
- After the user inputs the link or provides the text, we'll use CNN to extract features from the image and turn it into a feature vector with a fixed length.
- Following the extraction, we pre-process the images by modifying their size, orientation, colour, brightness, and perspective. We also remove a lot of noise from the caption, such as punctuation, during this process.
- The feature vector would then be fed to the RNN, which would then recursively generate a caption for the provided image.
- The users would be shown the description produced by the model as the last step.



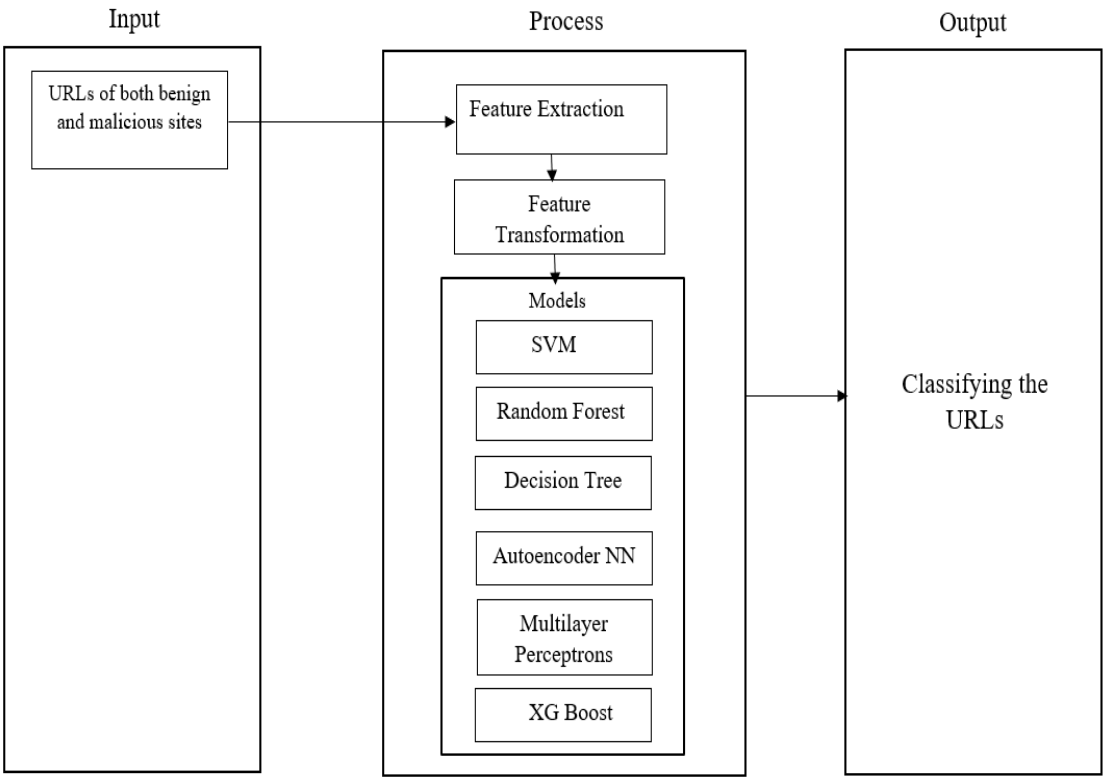
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

IV. MODULE DECOMPOSITION

The Module includes:

- Importing libraries and collecting the dataset: We are using .csv format for the dataset-URL Dataset. We then import various libraries required for other modules, Csv file converted into pandas data frame. URLs are given for Feature Extraction.
- Feature Extraction: IP address- provides the capabilities to create, manipulate and operate on IPv4 and IPv6 addresses and networks. re - regular expressions are used for feature extraction purpose. Whois: Create a simple importable Python module which will produce parsed WHOIS data for a given domain. Urllib: urllib is a package that collects several modules for working with URLs.urllib.request for opening and reading URLs.
- Feature Transformation: Feature values are assigned as 0's (Legitimate) and 1's (Malicious) based on the condition.
- Combining all the Features: Features extracted from different sources are combined after the feature transformation step for further process.
- Dividing feature vector dataset: Divide the dataset into training dataset and test dataset.
- Building Multiple Model: We are building six models which includes both Machine Learning and Deep Learning techniques.
- Evaluating and comparing the model: The models are evaluated using Accuracy_score metric. It is the ratio of number of correct predictions to the total number of input samples. We are comparing all the models based on their train and test accuracy.

V. BLOCK DIAGRAM



VI. CONCLUSION

A malicious website is a common social engineering method that mimics trustful uniform resource locators (URLs) and web pages. The objective of this project is to train machine learning models and deep neural networks on the dataset created to predict malicious websites. Both malicious and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The

performance level of each model is measured and compared. This project aims to provide a better prediction for Malicious URL's by using Machine learning and Deep learning techniques.

VII. REFERENCES

[1] D. J. Lemay, R. B. Basnet, and T. Doleck, "Examining the relationship between threat and coping appraisal in phishing detection among college students," Journal of Internet Services and Information Security, vol. 10, no. 1, pp. 38–49, 2020.

[2] H. Kim, "5G core network security issues and attack classification from network protocol perspective," Journal of Internet Services and Information Security, vol. 10, no. 2, pp. 1–15, 2020.

[3] K. Aram and J. O. SoK, "A systematic review of insider threat detection," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 10, no. 4, pp. 46–67, 2019.

[4] R. B. Basnet and R. Shash, "Towards detecting and classifying network intrusion traffic using deep learning frameworks," Journal of Internet Services and Information Security, vol. 9, no. 4, pp. 1–17, 2019.

[5] F. Valenza and M. Cheminod, "An optimized firewall anomaly resolution," Journal of Internet Services and Information Security, vol. 10, pp. 22–37, 2020.

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96.

[8] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.

[9] Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).

[10] D. Sahoo, C. Liu, S.C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey". CoRR, abs/1701.07179, 2017.