International Journal of
Recent Innovations in Academic Research
Open Access, Peer Reviewed, Abstracted and Indexed Journal

E-ISSN: 2635-3040; P-ISSN: 2659-1561
Homepage: https://www.ijriar.com/
Volume-7, Issue-3, Mar-2023: 28-35

Research Article

Phishing Website Detection Using Deep Learning and Machine Learning

Manish Javvadi[a], Sai Suraj Mohan, M[b]., Teja Naidu, S[c]., Teja, G.V.S.S[d]. and Dr. Prem Kumar Singh[e]

[a-e]Department of Computer Science and Engineering, Gitam University, Visakhapatnam, India
[a]Email: 121910314010@gitam.in; [b]Email: 121910314045@gitam.in;
[c]Email: 121910314020@gitam.in; [d]Email: 121910314050@gitam.in; [e]Email: psingh@gitam.edu

Received: February 27, 2023 **Accepted:** March 16, 2023 **Published:** March 23, 2023

Abstract: A typical strategy for facilitating unwanted data, like spam, noxious notices, which drive-by weaknesses, are malicious or phishing Uniform Resource Locators (URLs). Perceiving rebel URLs quickly is basic. Boycotting, customary articulation, and mark matching have all been utilized in past examinations. For identifying new URLs or variants of existing malicious URLs, these methods are completely useless. Providing a solution that is based on machine learning may solve this issue. This kind of arrangement requires significant concentrate in the fields of element designing and component portrayal of safety antiquities like URLs. In addition, new URLs and variations of existing URLs must constantly be accommodated by reforming feature engineering and feature representation resources. Recently, systems considering deep learning, machine learning (ML), and artificial intelligence (AI) have shown up at human-level execution in different districts and, shockingly, beat human vision in some PC vision applications. They are able to naturally separate the best component portrayal from crude information sources. We propose different profound learning and ML estimations, including KNN, SVM, Random Forest, Decision Tree, Logistic Regression, Naive Bayes, RNN–LSTM in which rough URLs are encoded using character level embedding, to involve and redesign their display in the field of organization security.
Keywords: URL, phishing URL detection, feature extraction, feature selection, machine learning.

Introduction

Unwanted content is hosted by malicious Uniform Resource Locators (URLs), and in cyber-attacks, attackers frequently employ malicious URLs as a primary strategy. Email and informal communication locales like Facebook, Twitter, WhatsApp, and Orkut, among others, are the applications that are utilized the most often to spread vindictive URLs [1-3]. On their website, they host information that was not invited. The host may become compromised if an unwary individual accidentally accesses that website through the URL, making them victims of malware installation, data theft, and identity theft scams. Rogue URLs cause billions of dollars in damages each year [4]. In these circumstances, efficient methods for quickly identifying malicious URLs and notifying the network administrator are required. The blacklisting strategy is the foundation of the majority of commercial products currently on the market [5]. A database containing a list of harmful URLs serves as the foundation for this strategy. The anti-virus organization scans and crowd sources solutions to constantly update the blacklists. Using the blacklisting method, harmful URLs that are already in the database can be found. However, they are completely incapable of identifying completely new dangerous URLs or variants of previously known malicious URLs. Cybercriminals have recently developed a variety of new strains of malware by employing mutation techniques. ML approaches are used to deal with this. Using area ability to separate lexical properties from URLs has

turned into the most widely recognized technique as of late, trailed by Ml models. The most notable method for feature planning is bag-of-words (BoW), while the most broadly perceived ML model is support vector machine (SVM).

Since the quantity of malignant URL dispersions has expanded after some time, it is important to research and execute techniques or systems to recognize and forestall these URLs. There are at present two fundamental patterns with respect to the distinguishing proof of noxious URLs: malicious URL identification in view of conduct examination approaches and signals or rules-based recognition.

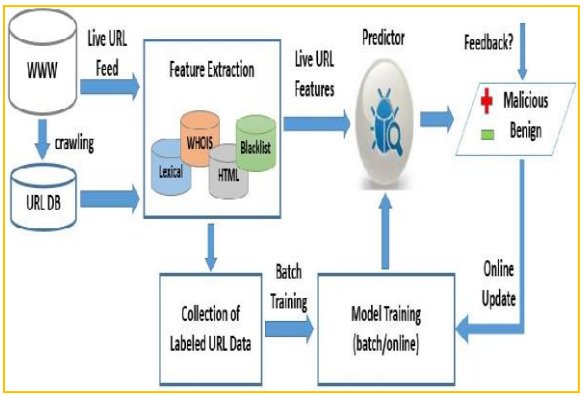

Figure 1. Example Figure

A method that uses a collection of markers or criteria to identify malicious URLs may be able to detect harmful URLs quickly and reliably. In any case, new perilous URLs that don't match any of the predetermined pointers or rules can't be distinguished utilizing this technique. During the time spent distinguishing pernicious URLs in light of conduct examination techniques, ML or deep learning calculations are utilized to characterize URLs as per their activities. This work utilizes ML strategies to characterize URLs as indicated by their properties. Moreover, an original strategy for extricating URL credits is remembered for the review. AI strategies are utilized in our research to classify URLs in light of their qualities and conduct. The traits, which are previously unknown, are gleaned from both static and dynamic URL behaviors [6]. The newly suggested characteristics are the research's main contribution. ML strategies are integrated into the malicious URL identification framework. Support vector machine (SVM) and random forest (RF) are two frequently utilized administered ML strategies.

Literature Survey

Towards a feature rich model for predicting spam emails containing malicious attachments and URLs [1]

Spam emails are increasingly containing malicious content, like URLs and attachments. URLs and attachments that are malicious attempt to spread malware that could compromise a computer's security. In order to avoid being detected by virus scanners, which are utilized by the majority of email systems to check for such threats, these dangerous attachments also aim to imitate their content. We report early exploration on recognizing the sort of spam email that is probably going to incorporate these very destructive spam messages in this review, which depends on two true informational collections. We give a far reaching put of qualities for the substance of messages together to recognize designs in messages containing hurtful substance. We show the way that these qualities can anticipate noxious associations with an AUC-PR of up to 95.2% and URLs with an AUC-PR of up to 68.1%. Our strategy might assist with decreasing dependence on URL boycotts and infection scanners, which much of the time don't refresh as fast as the hurtful substance that is being recognized. The quantity of assets presently expected to recognize perilous data might be diminished by these innovations.

Malicious spam emails developments and authorship attribution [3]

The decentralized structure of the Internet makes it possible to communicate rapidly, communicate with people located on opposite sides of the world, and maintain one's anonymity, all of which are essential for conducting criminal activities. Cybercrime has transformed from a typically low-volume wrongdoing to a typical high-volume wrongdoing almost immediately as the Web has grown. One normal illustration of this sort of wrongdoing is the circulation of spam email, in which the message attempts to get the beneficiary to download a hazardous connection or snap on a URL that focuses to a noxious site. Investigators endeavoring to give data on spam tasks rapidly find a gigantic everyday spam stream; thusly, any information acquired addresses just a little model, rather than the whole picture. Our underlying examination inspects the convenience of utilizing creation based models for this reason, though past investigations analyzed the utilization of point based models to mechanize portions of these investigations, for example, partitioning email bunches into bunches with tantamount subjects. We bunched a gathering of spam messages utilizing origin based grouping in the initial step. In the subsequent stage, we analyzed those bunches by utilizing an assortment of etymological, underlying, and syntactic qualities. Notwithstanding the way that it is impossible that we had the option to bunch all spam made by each gather, these outcomes demonstrate that messages inside each group were doubtlessly composed by a similar creator. Origin concentrates on in the past have resolved this issue of high virtue and low review. Despite the fact that this is an imperfection in our examination, the actual groups are as yet valuable for computerizing examination since they require little exertion. Our subsequent stage revealed significant data about the association that could be utilized in resulting examination to find extra connections behind spam tasks or other comparable gatherings.

An analysis of the nature of groups engaged in cybercrime [4]

The idea of cybercrime groups is the subject of this review. It provides a concise summary of the definition and scope of cybercrime, as well as theoretical and empirical concerns regarding the known characteristics of cybercriminals and the anticipated involvement of organized crime organizations. The study demonstrates individual and collective behavior, as well as the motives of typical offenders, including governmental actors, through the use of actual incidents. The McGuire typology is used to describe a variety of cybercrime and criminal organizations. It is unmistakably clear that a wide assortment of hierarchical designs are engaged with cybercrime. Administration, association, and specialization are commonly expected for big business or benefit looking for activities, especially cybercrime perpetrated by state entertainers. Then again, fight activities are regularly less organized and miss the mark on clear line of order.

Methodology

Malicious Uniform Resource Locator (URL), otherwise called a "pernicious site," is quite possibly of the most widely recognized way that ongoing frameworks store destructive information, like spam, vindictive ads, phishing, and drive-by weaknesses. Perceiving rebel URLs straightaway is basic. Boycotting, standard articulation, and mark matching have all been utilized in past examinations. For identifying new URLs or variants of existing malicious URLs, these methods are completely useless. Providing a solution that is based on machine learning may solve this issue.
This strategy fails because it is ineffective.

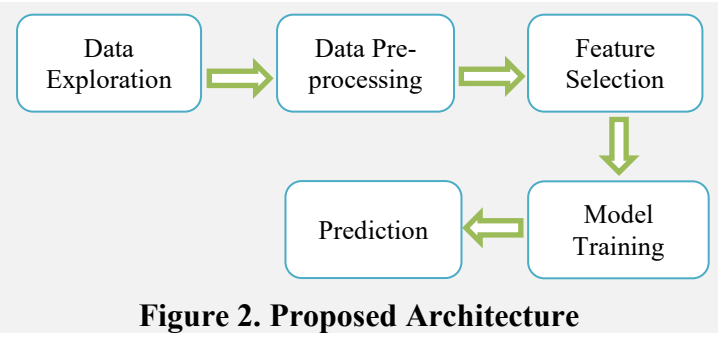

Figure 2. Proposed Architecture

In this research, assets for highlight designing and portrayal should be continually refreshed to oblige new URLs or varieties of existing URLs. Lately, frameworks in light of deep learning, ML, and artificial intelligence (AI) have arrived at human-level execution in various regions and, surprisingly, outperformed human vision in some PC vision applications [6]. From crude information sources, they can remove the best component portrayal consequently. We propose a variety of deep learning and machine learning calculations, such as Decision Tree, KNN, SVM, Random Forest, Logistic Regression, Naive Bayes, and RNN-LSTM in which crude URLs are encoded using character level implanting, in order to utilize and enhance their display in the field of network security [7].

The advantages of this framework incorporate its productivity and astounding execution.

The model detects the URL as malicious or legitimate based on attributes like: token count, number of dots, length of the URL, length of the host, average domain token, and path. For example- Phishing URL: http://br-ofertasimperdiveis.epizy.com/produto.php?linkcompleto=iphone-6-plus-apple-64gb-cinza-espacial-tela-5-5-retina-4g-camera-8mp-frontal-ios-10-proc.-m8/p/2116558/te/ipho/&id=10

Legitimate URL: https://www.youtube.com/

Implementation

Decision Tree, KNN, SVM, Random Forest, Logistic Regression, Naive Bayes, and RNN- LSTM are two or three the deep learning and ML estimations we inspect for embedding rough URLs at the individual level.

Algorithms

KNN

Classification and regression problems can be addressed with the straightforward supervised machine learning approach known as the k-nearest neighbors (KNN) method.

In view of an assortment of info values, ML models expect yield values. One of the most key sorts of ML calculations is the KNN, which is normally used for grouping.

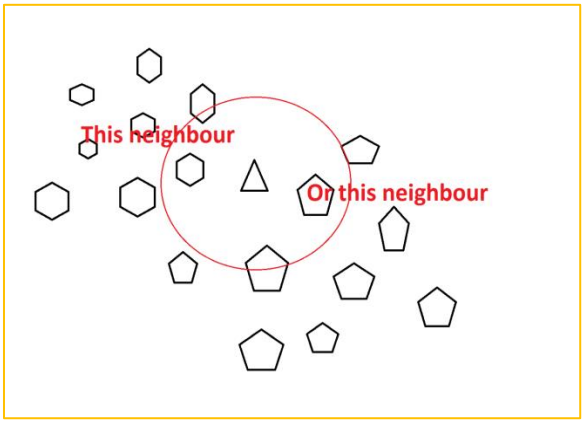

Figure 3. KNN

New information focuses are ordered by KNN in light of the fact that they are so like recently put away data of interest.

Random Forest

A kind of managed ML calculation known as random forest is regularly used in grouping and relapse issues. It utilizes the larger part vote in favor of arrangement and the normal for relapse from numerous examples to make choice trees. The random forest calculation's capacity to deal with

informational collections with both nonstop and absolute factors, as in relapse and order, is quite possibly of its most significant component. It beats various estimations in order endeavors.

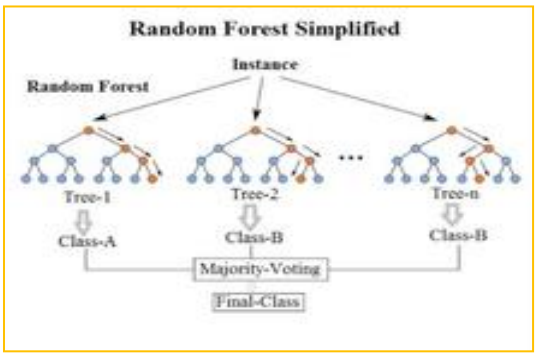

Figure 4. Random Forest

Decision Tree

The administered learning calculation family incorporates the Decision Tree calculation. As opposed to other regulated learning calculations, the choice tree approach can likewise be utilized to take care of relapse and grouping issues. A preparation model that is able to predict the class or value of an objective variable is created by incorporating fundamental decision rules from previous data into a decision tree.

In Decision Trees, determining a record's class name begins at the base of the tree. Consideration is given to the advantages of the record trait and the root trait.

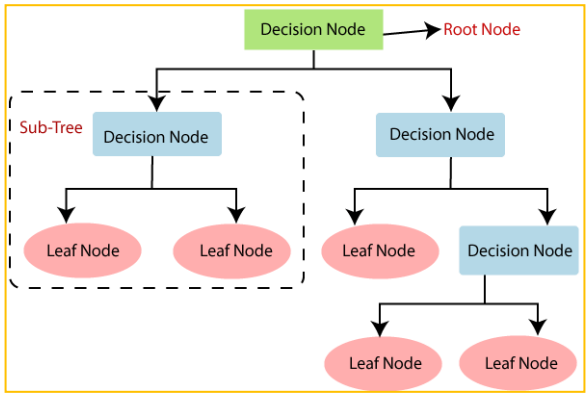

Figure 5. Decision Tree

RNN-LSTM

Long short-term memory (LSTM) is the name of a artificial recurrent neural network (RNN) design related to deep learning. LSTM networks are a type of RNN that utilization extraordinary units as well as customary units. A "memory cell" in LSTM devices allows for the long-term storage of data. Information's entry, production, and deletion are all controlled by a collection of gates.

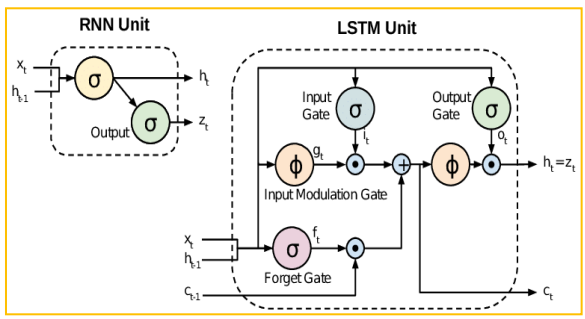

Figure 6. RNN-LSTM

Linear Regression

A technique for ML that depends on directed learning is linear regression. Regression testing is directed. Considering free factors, backslide models an objective assumption regard. More often than not, it's utilized to sort out what factors mean for expectations.

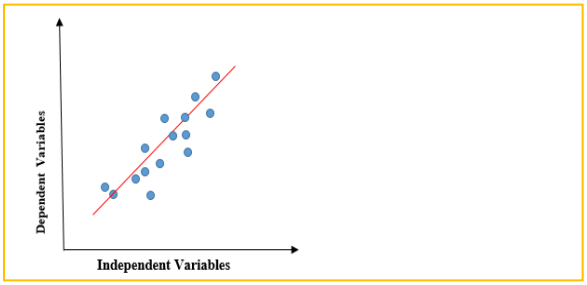

Figure 7. Linear Regression

SVM

SVM is a kind of regulated ML that can be utilized to take care of issues with relapse or grouping. It utilizes a strategy called the portion stunt to change your information and afterward utilizes these progressions to sort out a decent limit between the potential results.

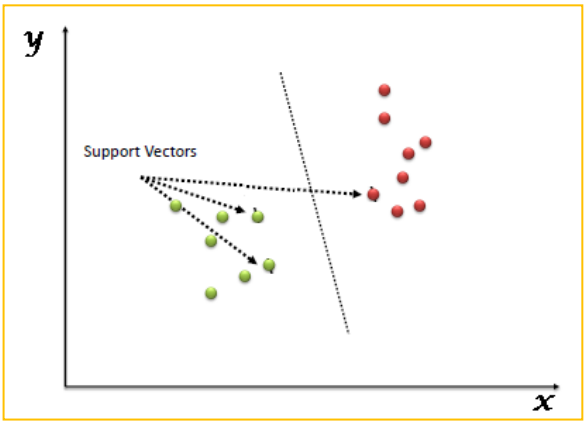

Figure 8. SVM

The data you give the kernel technique changes. It is comparable to unraveling a DNA strand. After using the kernel technique, you begin with a data vector that appears to be harmless, but it becomes a much larger collection of data that cannot be comprehended using a spreadsheet. However, the magic happens here: The SVM algorithm is able to generate a hyper plane that is much more optimal because the dataset has been expanded, resulting in more distinct boundaries between your classes.

Table 1. Model Accuracy Table

	Model	Test Score
4	Decision Tree	0.961104
1	KNN	0.941655
3	SVM	0.937585
5	Random Forest	0.936680
0	Logistic Regression	0.916780
2	Naïve Bayes	0.916780

The above tables are the accuracy or test score of the models we used in the project to detect phishing websites or URLs. Decision tree got the highest compared to all machine learning algorithms which is 96.1%. We also used deep learning model RNN-LSTM which got 99% accuracy.

Experimental Results

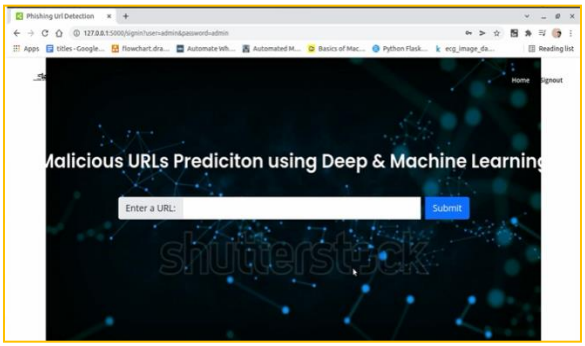

Figure 9. Main Page

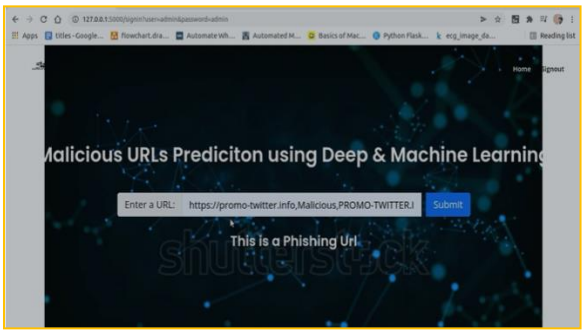

Figure 10. Prediction Result

Conclusion

A few deep learning-based person-level inserting models for identifying malicious URLs are examined in this paper. Every deep learning engineering is almost exactly the same. Manage harmful URL varieties with any deep learning character level installing based model. Despite the fact that deep learning has been effective, it is worthwhile to begin with customary strategies like standard articulation based boycotting, signature coordinating, and ML put together arrangements prior to moving with respect to deep learning-based character level installing models.

The data mentioned in Table-1 concludes that Decision Tree (96.1%) is more accurate of all machine learning models for the given dataset and RNN-LSTM (99%) is most accurate of deep learning models.

Future Scope

To make the Deep URL Detect (DUD) design more vigorous, assistant modules like enrollment administrations, site content, network notoriety, record ways, and vault keys can be added. One of the most urgent ways for improvement in what's to come is this.

Declarations

Acknowledgements: Not applicable.
Conflict of interest: Authors declare that there is no actual or potential conflict of interest in relation to this article.
Funding: Authors claim no funding received.
Author Contributions: All authors contributed equally.

References

1. Tran, K.N., Alazab, M. and Broadhurst, R. 2014. Towards a feature rich model for predicting spam emails containing malicious attachments and urls. Proceedings of Australasian Data Mining Conference, Conference paper. Retrieved from http://hdl.handle.net/1885/28534

2. Alazab, M. and Broadhurst, R. 2016. Spam and criminal activity. Trends and Issues in Crime and Criminal Justice, 526: 1-20.
3. Alazab, M., Layton, R., Broadhurst, R. and Bouhours, B. 2013. Malicious spam emails developments and authorship attribution. In Cybercrime and Trustworthy Computing Workshop (CTC), 2013 Fourth (pp. 58-68). IEEE.
4. Broadhurst, R., Grabosky, P., Alazab, M., Bouhours, B. and Chon, S. 2014. An analysis of the nature of groups engaged in cyber crime. An analysis of the nature of groups engaged in cyber crime. International Journal of Cyber Criminology, 8(1): 1-20.
5. Alazab, M., Venkatraman, S., Watters, P. and Alazab, M. 2011. Zero-day malware detection based on supervised learning algorithms of API call signatures. In: Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 171- 182). Australian Computer Society, Inc..
6. Vinayakumar, R., Alazab, M., Srinivasan, S., Pham, Q.V., Padannayil, S.K. and Simran, K. 2020. A visualized botnet detection system based deep learning for the internet of things networks of smart cities. IEEE Transactions on Industry Applications, 56(4): 4436-4456.
7. Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A. and Venkatraman, S. 2019. Deep learning approach for intelligent intrusion detection system. IEEE Access, 7: 41525-41550.

Citation: Manish Javvadi, Sai Suraj Mohan, M., Teja Naidu, S., Teja, G.V.S.S. and Prem Kumar Singh. 2023. Phishing Website Detection Using Deep Learning and Machine Learning. International Journal of Recent Innovations in Academic Research, 7(3): 28-35.