IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTING PHISHING WEBSITE USING ML

Mr.Pratharv Surve, Mr.Ayush Singh, Ms.Khushboo Yadav, Mr.Aman Yadav
Assistant Professor, Undergraduate Student, Undergraduate Student, Undergraduate Student
Department of Information Technology
University of Mumbai, Mumbai, India

Abstract: Phishing assaults, which include content injection phishing, social engineering, online social networks, and mobile applications, employ advanced tactics and technologies to obtain sensitive information. To stop and reduce the chances of these attacks, a number of phishing detection methods were developed; deep learning algorithms were one of them that showed potential. The application of deep learning algorithms to phishing detection lacks a comprehensive overview, and the findings and related lessons are scattered throughout multiple publications. We carried out a systematic literature review (SLR) to find, assess, and condense the research results on deep learning algorithms for phishing detection as offered by the selected academic publications.

The phishing website may be recognized in the final phishing detection rate depending on a number of important features, including the URL and Domain Identity, security and encryption specifications, and other elements. When a consumer completes an online transaction and pays through a website, our technology employs deep learning algorithms to identify whether or not the website is a phishing website.

***IndexTerms* -** Detect Fake Urls

I. INTRODUCTION

The most dangerous illegal activity in cyberspace is phishing. Since the majority of people use the internet to access services offered by banking and governmental institutions, phishing attempts have significantly increased over the past few years. Phishers began to make money, and they now run a profitable business doing this. Phishers attack susceptible people using a variety of techniques, including VOIP, SMS, spoof links, and fake websites. It is quite simple to make fake websites that, in terms of design and content, resemble authentic websites. These websites would even have the exact same material as their authentic counterparts. These websites were made in order to collect personal information from users, such as account numbers and login.

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

II.PURPOSE

The goal of employing machine learning to identify phishing websites is to improve cybersecurity and shield users from falling for phishing scams. Phishing is a type of cybercrime in which criminals design phony websites that resemble real ones in an effort to dupe users into disclosing sensitive information like login passwords, credit card information, or personal information. Usually, these attacks spread via email, social media, or some other channel.

III. SCOPE

To protect clients from phishing schemes, machine learning-based phishing website identification was developed. It can also help businesses avoid financial losses and data breaches. Machine learning has the ability to greatly improve phishing detection systems' efficiency. Even while machine learning has shown considerable potential in identifying phishing websites, there are still certain drawbacks to the technique. Attackers can, for instance, try to avoid being discovered by constantly advancing their methods.

IV. EXISTING ALGORITHM

The website's URL was analyzed by the current system, which has a 94% accuracy rate in phishing detection. Different URL addresses can be created using the parameters for domain, subdomain, top level domain, protocol, directory, file name, path, and query. The relevant fields on phishing URLs typically differ from those on authentic websites. The accurate features derived from the URL improve the classification's accuracy. Accuracy can also be increased by altering the site's layout, CSS, content, meta data, and other characteristics. On the other hand, these capabilities will lengthen the time it takes to classify newly created websites

- **Random Forest**

One machine learning method for solving regression and classification issues is the random forest. It makes use of ensemble learning, a method that solves complicated problems by combining a large number of classifiers.An algorithm called random forest is made up of several decision trees. Through bagging or bootstrap aggregating, the random forest algorithm trains its "forest." An ensemble meta-algorithm called bagging raises the precision of machine learning algorithms.

- **Decision Tree Classifier**

In machine learning, a decision tree is a flexible, comprehensible approach for predictive modeling. It is appropriate for both regression and classification tasks since it organizes judgments according to input data. This article explores decision trees' uses and learning algorithms while delving into their constituent parts, jargon, construction, and benefits.

One popular supervised learning technique in machine learning is the decision tree, which models and predicts outcomes based on input data. Each internal node in the structure resembles a branch that corresponds to an attribute value, and each leaf node reflects the ultimate conclusion or prediction. The structure is like to a tree. Under the heading of supervised learning is the decision tree algorithm. Both regression and classification issues can be resolved with them.

- **Support Vector Machine**

Such adjustable machine learning algorithm with applications in multiple fields is Support Vector Machine (SVM). SVMs provide a significant part in biomedical research by improving with tasks like expression analysis and protein structure prediction.They specialize in analysing high-dimensional data and can find patterns in even the most complicated biological datasets. Additionally, SVMs have significantly advanced the field of computer vision, especially in areas such as facial recognition, image sorting, and identifying objects. Image-based applications have widely adopted them due to their capacity to learn discriminative characteristics from images and efficiently classify them into a number of categories.SVMs are very useful in the finance sector for jobs like fraud detection and stock market forecasting. Their ability to examine past data and spot complex patterns allows.

- **AdaBoost**

Adaptive Boosting, or AdaBoost, is a machine learning ensemble technique that builds a powerful classifier by aggregating the predictions of several weak learners. AdaBoost's main idea is to train a sequence of weak classifiers iteratively, with each new classifier placing more emphasis on the cases that the preceding classifiers misclassified. AdaBoost effectively makes the following weak learners pay greater attention to the misclassified examples by giving them higher weights during each iteration. Consequently, the ensemble gradually gains the ability to highlight the hard-to-classify cases, which results in an improved overall forecast.AdaBoost functions by using the training data to train a base classifier, which is typically a linear classifier or a decision tree with restricted depth. Based on the effectiveness of the current weak classifier, AdaBoost modifies the weights of the training instances at the end of each iteration. Erroneously classified instances are assigned a lower weight than correctly classified instances. Next, using the new dataset, the next weak classifier is trained, emphasizing the previously incorrectly categorized occurrences. This procedure repeats until a predetermined degree of accuracy is attained, or for a predetermined number of times.

- **Neural Network**

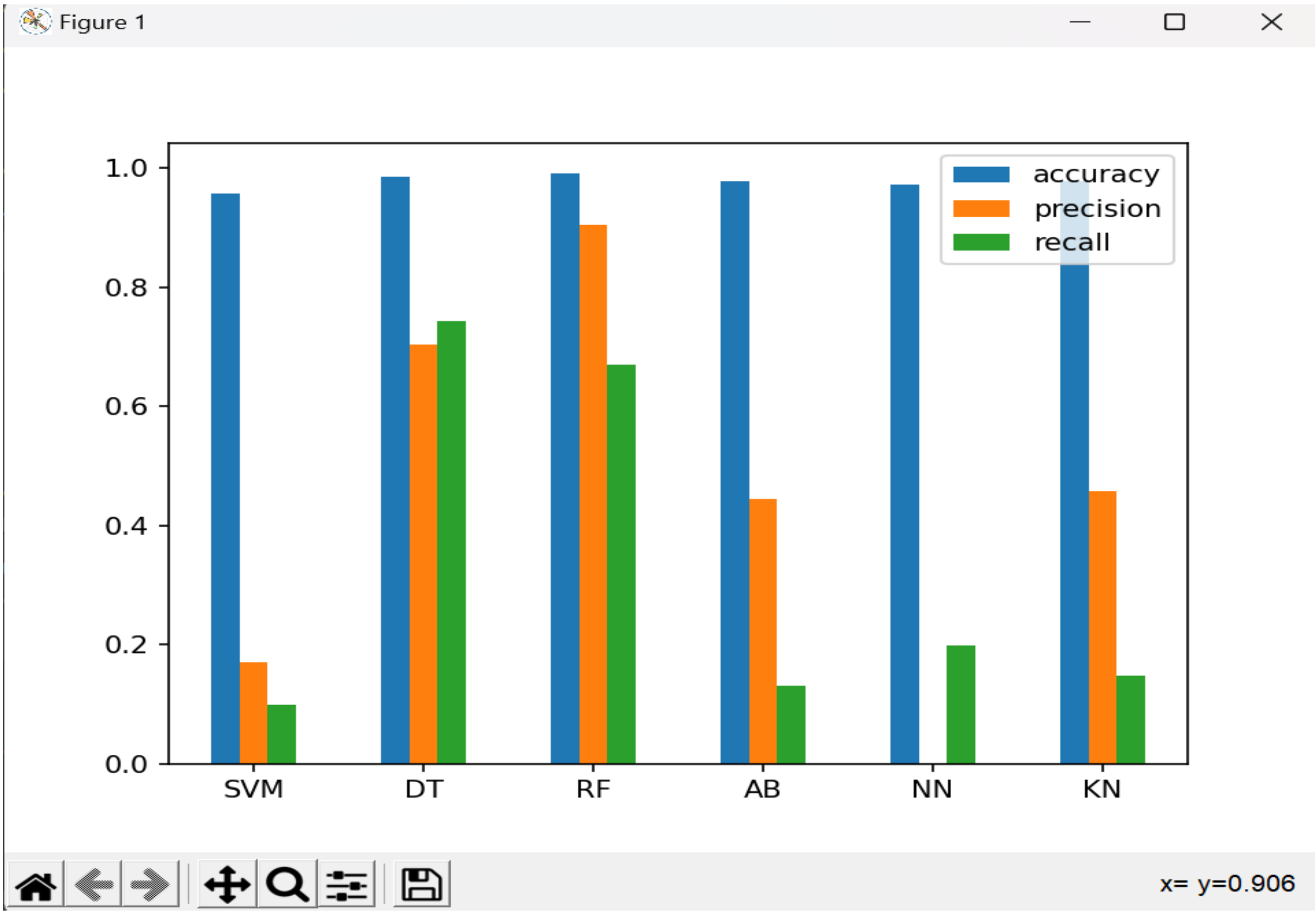
A class of machine learning models called neural networks is modeled after the composition and operations of the human brain. They are made up of interconnected nodes arranged in layers, with an output layer, an input layer, and one or more hidden layers. In a neural network, each node, or neuron, processes its incoming data using elementary mathematical operations before sending the output through weighted connections to the layer above.

In order to reduce the discrepancy between their expected outputs and the actual targets, neural networks acquire the ability to modify the weights of these connections throughout the training phase. Usually, to do this, an optimization approach such as gradient descent is used, in which the network's performance is iteratively enhanced by updating the weights in the direction that decreases the loss function.

- **K-Neighbours**

In supervised learning, the K-Nearest Neighbors (K-NN) algorithm is a straightforward but powerful technique for both regression and classification problems. The concept of similarity between data points is essential to its functionality. Based on a selected distance metric—typically Euclidean distance—K-NN determines the K data points (neighbors) that are closest to a newly provided data point for prediction. The training dataset is used to determine these neighbors.In classification tasks, K-NN gives the new data point the class label that is most common among its K nearest neighbors. To put it another way, it decides the class of the new instance by holding a majority vote among its closest neighbors.

TEST RESULT FOR CLASSIFIERS



V.Acknowledgment

The success and final outcome of any project requires a lot of guidance from many people and we are extremely privileged to have this all along with the completion of my project. We owe our deep gratitude to our project guide **Mr.Pratharv Surve**. who took interest in our project work and guided us all along till the completion of our project work by providing all the necessary information for developing a good system.I would like to extend my sincere and heartfelt thanks towards all those who have helped me in making this project. Without their active guidance, help, cooperation and encouragement, I would not have been able to present project in timeI also acknowledge with a deep sense of reverence, my gratitude towards **Mr.Pratharv Sir** for their valuable suggestions given to me in completing the project.

References

[1]. J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[2]. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[3]. T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[4]. M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[5]. [5] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[6]. Mohammad R., Thabtah F. McCluskey L. (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016.

[7]. A. Vazhayil, R. Vinayakumar, and K. Soman, “Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks,” in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 16.

[8]. W. Fadheel, M. Abusharkh, and I. Abdel-Qader, “On Feature Selection for the Prediction of Phishing Websites,” 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[9]. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, “Boosting the Phishing Detection Performance by Semantic Analysis,” 2017.

[10]. L. MacHado and J. Gadge, “Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

[11]. Sharma, Ushamary and Ghisingh, Seema and Ramdinmawii, Esther, “A Study on the Cyber - Crime and Cyber Criminals: A Global Problem,” International Journal of Web Technology, vol 03, pp. 172-179, June 2014.

[12]. Andrewa, “Cybercrime”, http://en.wikipedia.org/wiki/Computer_crime, October 15, 2003.

[13]. Vayansky, I. and Kumar, S., “Phishing – challenges and solutions.”, Computer Fraud & Security, vol 2018, no.
1, pp. 15-20, January 2018.

[14]. Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, “Phishing Detection Using Machine Learning Techniques,” unpublished.

[15]. Gokula Chandar ,Leeban Moses M; T. Perarasi M; Rajkumar; “Joint Energy and QoS-Aware Cross-layer Uplink resource allocation for M2M data aggregation over LTE-A Networks”, IEEE explore, doi:10.1109/ICAIS53314.2022.9742763.

[16]. Mustafa Alper Akkaş, Radosveta Sokullu, "An IoT-based greenhouse monitoring system with Micaz motes", https://doi.org/10.1016/j.procs.2017.08.300.

[17]. P. V. Vimal and K. S. Shivaprakasha, "IOT based greenhouse environment monitoring and controlling system using Arduino platform," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, 2017, pp. 1514-1519.

[18]. Dhuddu Haripriya, Venkatakiran S, Gokulachandar A, “UWB-Mimo antenna of high isolation two elements with wlan single band-notched behavior using roger material”, Vol 62, Part 4, 2022, Pg 1717-1721, https://doi.org/10.1016/j.matpr.2021.12.203.

[19]. Gokula Chandar A, Vijayabhasker R., and Palaniswami S, “MAMRN – MIMO antenna magnetic field”, Journal of Electrical Engineering, vol.19, 2019.

[20]. Rukkumani V , Moorthy V, Karthik M , Gokulachandar A, Saravanakumar M, Ananthi P, “Depiction of Structural Properties of Chromium Doped SnO2 Nano Particles for sram Cell Applications”, Journal of Materials Today: