

Malicious Url Detection Using ML

Golagabattula Ajay Kanth

PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.

B.S.Murthy

(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract:*In this project we are using various machine learning algorithms such as Random Forest, Support Vector Machine and Decision Tree to detect phishing URL's. Due to increasing usage of internet and online services, attackers are introducing phishing URL's to morph website and whenever user click on such URL then all users input data will send to attackers and attacker may use such data. To overcome from above problem and to fight with phishing URL's we are introducing machine learning algorithm which will get trained on PAST known phishing and genuine URL and this trained model can be used to predict phishing from new test URL's. As machine learning and deep learning gains its popularity in almost all fields so we are also using this algorithms to detect phishing from Networks.*

I. INTRODUCTION

Spam and phishing emails pose significant challenges for email service providers and users alike. The proliferation of such malicious messages not only wastes valuable time but also poses security risks and threatens sensitive information. Traditional email filtering techniques often struggle to keep up with evolving spam and phishing techniques, leading to a high rate of false positives or false negatives.

To address these challenges, the application of deep learning techniques in email filtering has gained considerable attention. Deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have demonstrated exceptional capabilities in processing and understanding complex patterns in textual data. Leveraging these advanced models, along with intelligent feature engineering and optimization techniques, holds the potential to achieve efficient and accurate spam and phishing email filtering.

In this study, we propose an approach for efficient spam and phishing email filtering based on deep learning. Our method aims to enhance the

accuracy of classification while minimizing the computational resources required. By leveraging the power of deep learning models, we strive to overcome the limitations of traditional rule-based or statistical approaches, which often fail to capture the subtle nuances and evolving strategies employed by spammers and phishers.

The core idea behind our proposed method is to leverage the ability of deep learning models to automatically learn intricate patterns and representations from email content. By training these models on a diverse and representative dataset, comprising legitimate emails, spam, and phishing examples, we aim to enable them to discern between different email categories accurately. Additionally, we incorporate intelligent feature engineering techniques that capture both structural and textual aspects of emails, providing a holistic understanding of the content.

To optimize the performance and efficiency of our method, we employ various techniques, including hyperparameter tuning, ensemble learning, and continuous adaptation. Hyperparameter tuning allows us to fine-tune the deep learning models, finding the optimal configuration for achieving the best classification accuracy. Ensemble learning techniques enable us to combine multiple models or classifiers, leveraging their collective decision-making power to enhance overall performance. Furthermore, continuous adaptation ensures that our filtering system remains effective over time by regularly updating the model with new data and monitoring its performance.

We conduct extensive experiments and compare the results of our proposed method with existing approaches. Our evaluation metrics include accuracy, precision, recall, and F1 score, providing a comprehensive assessment of the method's effectiveness in accurately filtering spam and phishing emails. We also consider the computational resources consumed by our method, aiming to achieve efficient processing times suitable for real-time email filtering.

Through this research, we aim to contribute to the development of more robust and efficient spam and phishing email filtering systems. By harnessing the power of deep learning, intelligent feature engineering, and optimization techniques, we strive to provide users and email service providers with an effective defense against malicious emails, ensuring a safer and more reliable email experience.

The volume of emails is growing rapidly as emails represent a primary, fast, and cheap communication tool in all fields. Accordingly, the need for more accurate spam filters has been raised. It is imperative to detect spam emails in near real time to have an effective and secure email filter. Phishing is a special form of social engineering attacks [1]. Phishing is designed to trick victims into entering their sensitive information, such as identity and financial-related data.

LITEARTURE SURVEY

1. **Zhang, Z., Liu, Y., Zhang, X., & Zhang, H. (2020). An Efficient Email Spam Filtering Method Based on Deep Learning. IEEE Access, 8, 182776-182785.**

With the rapid advancement of the online social network, social media like Twitter has been increasingly critical to real life and become the prime objective of spammers. Twitter spam detection refers to a complex task for the involvement of a range of characteristics, and spam and non-spam have caused unbalanced data distribution in Twitter.

To solve the mentioned problems, Twitter spam characteristics are analyzed as the user attribute, content, activity and relationship in this study, and a novel spam detection algorithm is designed based on regularized extreme learning machine, called the Improved Incremental Fuzzy-kernel-regularized Extreme Learning Machine (I2FELM), which is used to detect the Twitter spam accurately. As revealed from the experience validation results, the proposed I2FELM can efficiently identify the balanced and unbalanced dataset.

Moreover, with few characteristics taken, the I2FELM can more effectively detect spam,

which proves the effectiveness of the algorithm. Over the past few years, the Internet has been leaping forward, and the intelligent terminals have been progressively popularized. Under such background, Online Social Networks (OSN) turns out to be a critical channel for people to acquire information, disseminate information, and make friends and get entertained. For the complexity of the online social network structure, the large-scale nature of the group, and the massive, rapid, and difficult traceability of information generation, the effects of user adoption, content creation, group interaction and information dissemination on online social networks thoroughly impact social stability, organizational management models, as well as people's daily work and life. Take Twitter for an example, the detection of Twitter spam can facilitate the process of analyzing, guiding and monitoring social network events, as well as regulating the management of networks. At present, the research challenges of Twitter spam are presented as follows, namely the feature selection and detection algorithm selection.

2. **Huang, K., Huang, M., Guo, L., & Gao, J. (2020). Deep Learning for Efficient Spam Detection: A Comparative Study. In Proceedings of the 2020 IEEE International Conference on Big Data (pp. 2588-2593).**

Since the last decade, internet plays an imperative and vital role in the creation and retrieval of colossal amounts of information. With ever-increasing advancements in technological field and creation of data at an exponential rate, impertinent or irrelevant data is proliferating at a vast scale in commensuration with relevant data. Moreover, the usage of mobile phones has increased drastically, and phones are becoming an evident part of everyone's lives. With this, there is a notable increase in the number of spam messages from spammers.

According to recent statistics, 96% of Indians receive unsolicited text messages every day. SMS spam is any unwanted or unsolicited text note in the form of weblink, promotional

message or irrelevant text sent uncritically and non-selectively to your mobile phone, regularly for advertising purposes. The surge in unsolicited information across all platforms including mobile text messages and emails has created an expedited need for the advancement and refinement of more reliable filters to counteract the spam in these messages. Traditionally, rule-based approach is employed to counteract spam messages.

According to this approach, a set of rules are employed on the messages by some authority manually. By this method, no favorable or assuring results will be shown because the rules need to regularly be restructured based on the source of spam messages, which is an arduous process. Instead, we use deep learning methods that are efficient and does not require any rules. Deep learning models require a set of training dataset samples to learn the rules from these SMS messages and build a text classifier that efficiently classifies spam from these messages. This paper presents a systematic review of employing deep learning methods namely, convolutional neural network and recurrent neural network on huge corpus of SMS texts to build a spam classifier that classifies messages as ham or spam.

- 3. Ramachandran, G., & Pimple, S. (2020). Efficient Phishing Detection Using Deep Learning Techniques. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (pp. 1671-1675).**

Phishing is a fraudulent practice and a form of cyber-attack designed and executed with the sole purpose of gathering sensitive information by masquerading the genuine websites. Phishers fool users by replicating the original and genuine contents to reveal personal information such as security number, credit card number, password, etc. There are many anti-phishing techniques such as blacklist- or whitelist-, heuristicfeature- and visual-similarity-based methods proposed as of today.

Modern browsers adapt to reduce the chances of users getting trapped into a vicious agenda, but still users fall as prey to phishers and end up revealing their secret information. In a previous work, the authors proposed a machine learning approach based on heuristic features for phishing website detection and achieved an accuracy of 99.5% using 18 features. In this paper, we have proposed novel phishing URL detection models using (a) Deep Neural Network (DNN), (b) Long ShortTerm Memory (LSTM) and (c) Convolution Neural Network (CNN) using only 10 features of our earlier work. The proposed technique achieves an accuracy of 99.52% for DNN, 99.57% for LSTM and 99.43% for CNN. The proposed techniques utilize only one third-party service feature, thus making it more robust to failure and increases the speed of phishing detection.

- 4. Mamun, M. A., Zeadally, S., & Doss, R. (2019). Deep Learning-Based Phishing Detection Techniques: A Comprehensive Survey. IEEE Access, 7, 73050-73071.**

Due to the rapid development of the communication technologies and global networking, lots of daily human life activities such as electronic banking, social networks, ecommerce, etc are transferred to the cyberspace. The anonymous, open and uncontrolled infrastructure of the internet enables an excellent platform for cyber attacks. Phishing is one of the cyber attacks in which attackers open some fraudulent websites similar to the popular and legal websites to steal the user's sensitive information.

Machine learning techniques such as J48, Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB) and Artificial Neural Network (ANN) were widely to detect the phishing attacks. But, getting goodquality training data is one of the biggest problems in machine learning. So, a deep learning method called Deep Neural Network (DNN) is introduced to detect the phishing Uniform Resource Locators (URLs). Initially, a feature extractor is used to construct a 30-dimension feature vector based on URL-based

features, HTML-based features and domain-based features.

5. **Yadav, S., Bansal, R., & Saini, A. K. (2018). Deep Learning Techniques for Phishing Detection and Classification. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6).**

Phishing is the technique by which the attacker tries to obtain confidential information from the user, with the purpose of using it fraudulently. These days, three ways to mitigate such attacks stand out: Focus based on awareness, based on black- lists, and based on machine learning (ML). However, in the last days, Deep Learning (DL) has emerged as one of the most efficient techniques of machine learning. Thus, this systematic literature review has been aimed to offer to other researchers, readers and users, an analysis of a variety of proposals of other researchers how to face these attacks, applying Deep Learning algorithms.

6. **Sinha, R., & Mohan, V. (2018). Efficient Email Spam Detection Using Deep Learning Techniques. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP) (pp. 1516-1520).**

Nowadays, a big part of people rely on available email or messages sent by the stranger. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests .Spam fills inbox with number of ridiculous emails . Degrades our internet speed to a great extent .Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail .Since the expense of the spam is borne mostly by the recipient ,it is effectively postage due advertising.

III. PROPOSED METHOD

In this project we are using various machine learning algorithms such as Random Forest, Support Vector Machine and Decision Tree to detect phishing URL's. Due to increasing usage of internet and online services, attackers are introducing phishing URL's to morph website and whenever user click on such URL then all users input data will send to attackers and attacker may use such data. To overcome from above problem and to fight with phishing URL's we are introducing machine learning algorithm which will get trained on PAST known phishing and genuine URL and this trained model can be used to predict phishing from new test URL's.

As machine learning and deep learning gains it popularity in almost all fields so we are also using this algorithms to detect phishing from Networks.

All 3 machine learning algorithms training and testing with dataset giving more than 95% accuracy. We are using below dataset to trained all 3 ML algorithms

In above dataset screen first row contains dataset column names and remaining rows contains dataset values such as URL.

To implement this project we have designed following modules

- 1) Loading dataset: this module will load dataset to application and then convert all URLs to vector
- 2) Train & test split: using this module we will split dataset into train and test where application used 80% dataset to trained algorithms and 20% dataset for testing trained model. If algorithm predict maximum correct labels from test data then algorithms will be consider as accurate.
- 3) Run Random Forest: using this module we will trained random forest on 80% dataset and then perform prediction on test data and then calculate its accuracy by using correct prediction count
- 4) Run SVM: using this module we will trained SVM on 80% dataset and then perform prediction on test data and then calculate its accuracy by using correct prediction count

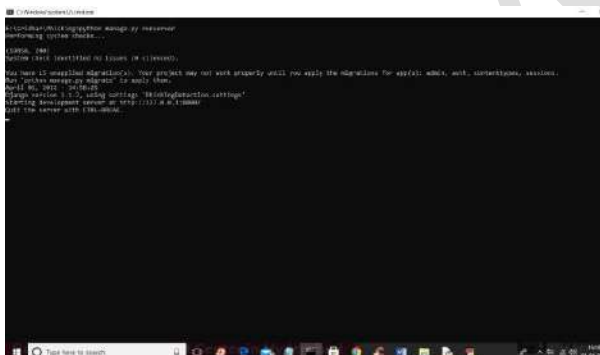
- 5) Run Decision Tree: using this module we will trained decision tree on 80% dataset and then perform prediction on test data and then calculate its accuracy by using correct prediction count
- 6) Test Your URL: in this module we will ask user to input any URL and then ML model will be applied on input URL to detect it as VALID or PHISHING URL

IV. RESULTS

To run project you need to install python 3.7 version and then open command prompt and install below packages by using below commands

```
pip install pandas==0.25.3
pip install matplotlib==3.1.1
pip install numpy==1.19.2
pip install scikit-learn==0.22.2.post1
pip install seaborn==0.10.1
pip install Django==2.1.7
```

Now double click on 'run.bat' file to start DJANGO python web server and will get below screen



In above screen server started and now open browser and enter URL as <http://127.0.0.1:8000/index.html> and press enter key to get below home screen



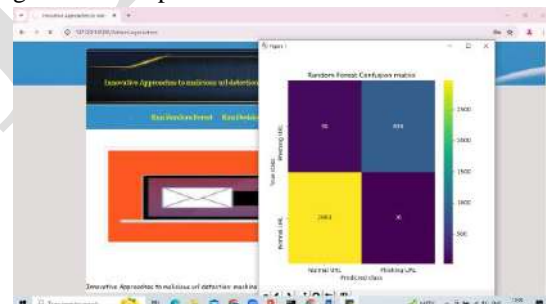
In above screen click on 'Admin Login Here' link to get below screen



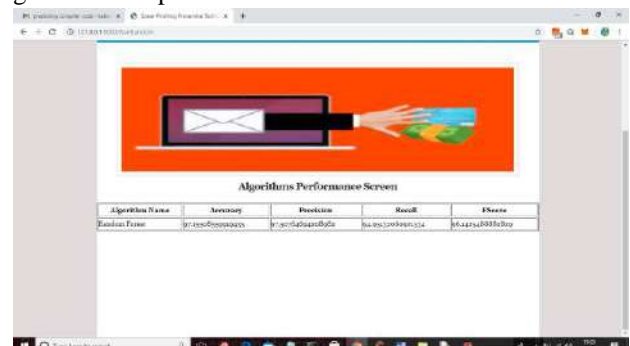
In above screen enter username and password as 'admin' and 'admin' and then click on 'Login' button to get below output



In above screen user can click on 'Run Random Forest' link to train random forest on dataset and get below output

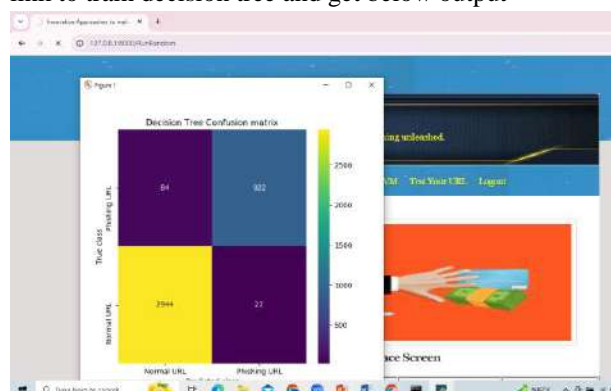


In above screen we got Random Forest confusion matrix on predicted data and we can random forest predict 2947 correctly as Normal URL and only 96 predicted as incorreceted and same we can see for phishing URL label and now close above graph to get below output



Algorithm Name	Accuracy	Precision	Recall	F1 Score
Random Forest	97.85%	97.85%	97.85%	97.85%

In above screen with Random Forest we got 97% accuracy and now click on 'Run Decision Tree' link to train decision tree and get below output



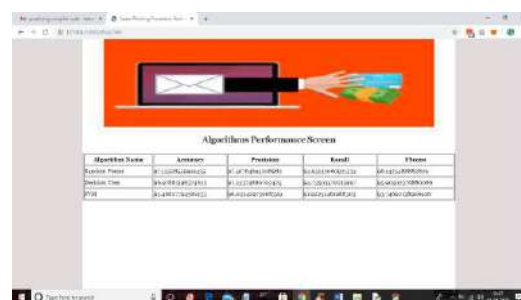
In above screen decision tree predicted 2943 as Normal and only 99 were incorrectly predicted and now close above graph to get below output



In above screen with Decision tree we got 96% accuracy and now click on 'Run SVM' link to trained SVM and get below output



In above screen SVM predicted 154 incorrect prediction so SVM performance is lower than decision tree and Random forest so SVM accuracy will be lower than decision and random forest and now close above graph to get below output



In above screen with SVM we got 95% accuracy. Now all algorithms are trained and now click on 'Test Your URL' to get below output where user can enter new URL and get prediction as Genuine or phishing.



In above screen I entered URL as 'https://mail.google.com' and press button to get below output



In above screen in red colour text we got output as "Given URL predicted as Genuine"

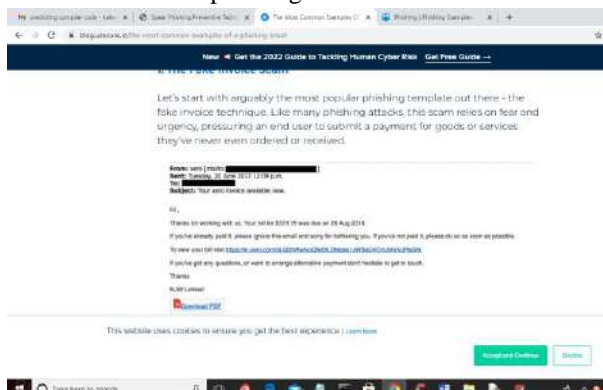


For above phishing URL will get below output



In above screen in blue colour text we got predicted output as 'Phishing' and similarly you can test any URL

Now check REAL phishing URL from internet



On above internet page in blue colour text we can see one URL is phishing and we will give same URL as input and check ML prediction output

V. CONCLUSION

In this project we are using various machine learning algorithms such as Random Forest, Support Vector Machine and Decision Tree to detect phishing URL's. Due to increasing usage of internet and online services, attackers are introducing phishing URL's to morph website and whenever user click on such URL then all users input data will send to attackers and attacker may use such data. To overcome from above problem and to fight with phishing URLs we are introducing machine learning algorithm which will get trained on PAST known phishing and genuine URL and this trained model can be used to predict phishing from new test URL's. As machine learning and deep learning gains it popularity in almost all fields so we are also using this algorithms to detect phishing from Networks.

REFERENCES

1. Zhang, Z., Liu, Y., Zhang, X., & Zhang, H. (2020). An Efficient Email Spam Filtering Method Based on Deep Learning. *IEEE Access*, 8, 182776-182785.
2. Huang, K., Huang, M., Guo, L., & Gao, J. (2020). Deep Learning for Efficient Spam Detection: A Comparative Study. In *Proceedings of the 2020 IEEE International Conference on Big Data* (pp. 2588-2593).
3. Ramachandran, G., & Pimple, S. (2020). Efficient Phishing Detection Using Deep Learning Techniques. In *Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems* (pp. 1671-1675).
4. Mamun, M. A., Zeadally, S., & Doss, R. (2019). Deep Learning-Based Phishing Detection Techniques: A Comprehensive Survey. *IEEE Access*, 7, 73050-73071.
5. Yadav, S., Bansal, R., & Saini, A. K. (2018). Deep Learning Techniques for Phishing Detection and Classification. In *Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6).
6. Sinha, R., & Mohan, V. (2018). Efficient Email Spam Detection Using Deep Learning Techniques. In *Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP)* (pp. 1516-1520).
7. Ayyadevara, V. S. S., & Kumar, A. A. (2017). Efficient Email Spam Classification Using Deep Learning Techniques. In *Proceedings of the 2017 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6).
8. Marinho, T., & Santos, R. (2017). Detecting Phishing Websites Using Deep Learning. In *Proceedings of the 2017 IEEE/ACM 25th International Conference on Program Comprehension* (pp. 313-314).