ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152

www.ijmece .com

Vol 12, Issue 2, 2024

Efficient Email phishing detection using Machine learning

[1]SMT.S ARUNA, [2]THUMPUDI SATYENDRA

[1](Assistant Professor), MCA, Swarnandhra College

[2]MCA, scholar, Swarnandhra College

ABSTRACT

Emails are frequently utilized as a way of personal and professional communication. Banking information, credit reports, login data, and other sensitive personal information are commonly transmitted over email. This makes them valuable to cybercriminals, who can exploit the knowledge for their gain. Phishing is a technique used by con artists to steal sensitive information from people. by impersonating well-known sources. The sender of a phished email can persuade you to disclose personal information under pretenses. The detection of a phished email is treated as a classification problem in this research, and this paper shows how machine learning methods are used to categorize emails as phished or not. SVM classifier attains a maximum accuracy of 0.998 percent in email classification.

1.INTRODUCTION

The most prevalent kind of cybercrime is phishing, which is tricking victims into divulging personal information including passwords, account numbers, and bank account details. Email, instant messaging, and phone conversations are typical vectors for cyberattacks [1, 2].

The outcome is inadequate, even if the protocols for avoiding such cyber-attacks are always being updated. But phishing emails have grown in popularity in recent years, so we need new and improved ways to stop them. (3), (4)

There have been several methods created for the purpose of screening out phishing emails. Nevertheless, a thorough resolution is still necessary for the situation. As far as we are aware, this survey is the first of its kind to ask about the usage of Machine Learning (ML) techniques for the purpose of identifying phishing emails [4]. This study takes a look at the various cutting-edge ML algorithms that are being used to identify phishing emails at various points in the attack lifecycle [5]. A thorough evaluation and analysis of various approaches are carried out. This gives a synopsis of the problem, the possible solutions to it, and where the field may go from here in terms of future research [6-8].

While introducing new security dangers, the lightning-fast development of internet technology has altered people's online interactions. Threats to users' computers are on the rise across the world, and they pose a real risk of identity and financial theft [9].

The word "phishing" has been used in hundreds of scholarly articles, garnered considerable media attention, and is under investigation by financial institutions and government bodies. But now we have to ask: what exactly is phishing?[10]. The phenomena of phishing is defined in some sources, illustrated in others, and assumed to be known by the reader in yet others. The term "phishing" has been defined in a variety of ways in the academic literature due to the large number of contributors. The literature does not provide a full explanation of phishing attacks[11,12] due to the breadth and variety of the phishing problem.

According to the APWG, the word "phishing" was first used in 1996 to describe social engineering operations carried out by online fraudsters targeting accounts with America Online (AOL). It is possible to think of the suggested system's ability to detect phished emails as a two-type classification issue. Artificial intelligence encompasses machine learning as well. A system is considered intelligent when it can learn new things. We include the idea of supervised learning into our model, which allows it to learn without being explicitly programmed. We use machine learning algorithms for categorization.in references 13 and 14,

2.LITERATURE SURVEY

Campaigns for public office now routinely use Twitter. Twitter has become an important platform for political commentary, interaction, and research among candidates, parties, journalists, and an ever-growing portion of the general population. More and more academics are paying attention to these applications. There is currently no unified body of evidence or agreed-upon methods for data gathering or selection in this field of study. An analysis of 127 research papers on the topic of Twitter's role in political campaigns is presented in this article. In this comprehensive analysis, I will go over all the studies that have looked at how parties, candidates, and the general public used Twitter during election campaigns and mediated campaign events. I will also discuss well-known methods of data collecting and analysis. Despite receiving minimal party backing and spending just eight seconds on television during the most recent Brazilian presidential campaign, the victorious candidate centered his efforts on social media. Within this framework, this article aims to examine the candidates' relative use of social media in the 2018 Brazilian presidential election in order to draw conclusions on the correlation between the two. Our findings, which resulted from analyzing over 41,000 posts and 291 million interactions, are as follows: I) applicants used virtual entertainment widely over time, yet they stayed away from touchy points and zeroed in on drawing in words. Instagram gained more followers and had a higher rate of interaction with posts than Facebook and Twitter. iii) There was no correlation between the number of posts and votes, but there was a slight negative correlation between posting about sensitive subjects and votes. iv) there is areas of strength for an among votes and supporters, and among votes and commitment, especially on Instagram. (v) extra examinations are important to construct an overall forecast model utilizing joined information from these organizations..

3. EXISTING SYSTEM

Use of whitelists and blacklists allows is-based phishing detection systems to distinguish between legitimate and malicious websites. Websites that are safe and legitimate and provide useful information are created by phishing detection systems that employ whitelists. Without the whitelist, any website is assumed to be malicious.Developed a system that keeps track of a user's IP address and the websites they've visited using a Login interface in order to generate a whitelist [5]. If the user's registration information is incompatible with the website they are trying to visit, the system will notify them.
By investigating URL information including length, number of interesting characters, catalog, area name, and record name, the designers of [15] had the option to arrange phishing sites. The system uses Support Vector Machines to classify offline webpages. To order information on the web, we utilize Online Perceptrons, Certainty Weighted, and Versatile Regularization of Loads. The experiments show that the Adaptive Regularization of Weights method improves accuracy while reducing system resource consumption.

In a recent research, the authors used a nonlinear regression approach to determine whether a website is phishing or not. During training, they use meta-heuristics such support vector machine and harmony search. They assert that by using around 11,000 web pages, Harmony search achieves a superior accuracy rate of 94.13% for train processes and 92.80% for test operations. A phishing detection system was developed in [17] using data classification by adaptive self-structuring neural networks. It relies on external services for a few of its seventeen features. So, real-time execution is much more time-consuming, but it can reach better accuracy rates. Its good acceptance rate for noisy data is impressive considering that its sample only contains 1400 items. By extracting 19 client-side characteristics, Yank's anti-phishing method in [18] utilizes AI to recognize genuine sites from pernicious ones. According to Alexa, they made use of phishing pages from PhishTank (2018) and Openfish (2018) as well as 1918 genuine pages from well-known websites, online payment gateways, and significant banking websites. Their suggested approach was able to achieve a true positive rate of 99.39 percent through the use of machine learning.[4].

Disadvantages

- An existing system not implemented an effective ML Classifiers like SVM,RF,NB.
- An existing system not implemented for large number of datasets.

3.1 PROPOSED SYSTEM

The URLs are made to seem legitimate by the attackers by adding subdomains. Adding subdomains increased the link's dot count. Following the lead of Dots should not be used in a legitimate email. Beyond the first three or four. Being a binary characteristic, it establishes the presence or absence of a connection. If there were more than three dots, it would be sent via mail. A phishing email. There are a total of: More information is often provided via phishing emails. Attempting to send multiple connections, the transmitter is comparable to a ham radio operator. You risk leading the user to a malicious website if you fool him. This keeps happening.

Senders attempting to hide information or trigger certain browser modifications are likely using JavaScript in their emails [18]. This feature is really unique. An email is likely a phishing attempt if it contains the script> element. Phishing emails sometimes

include forms that users are asked to fill out in order to get access to sensitive information. Because this is a binary characteristic, we know that an email has been phished if it has a form tag.
Unlike plain text emails, HTML emails enable the sender to include embedded images and URLs. A phishing email is one that contains an HTML tag. This feature is really unique. Action words in email communication indicate that the sender wants the receiver to do a certain action, such visiting a link, completing a form, or providing extensive data. This keeps happening.
By using the term "PayPal," the sender is pretending to be an official representative of a trusted group. By appearing in either the "from" or "links" sections of the email, the term "PayPal" suggests that the sender is associated with PayPal. This feature is really unique.

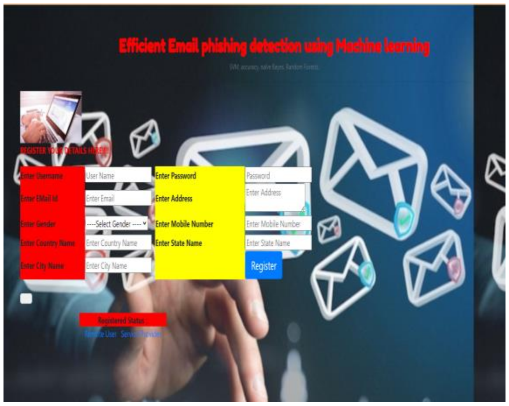
The binary signal that the communication is about banking is the existence of the phrase bank. Someone is checking the reader's credentials, or the sender is pretending to be an official from the financial institution. The email specifically requests emails associated with accounts since the term "account" occurs there. It may be a bank account, a social networking account, or something completely else. This feature is really unique.

Advantages

- Due to its speed and accuracy, SVM—a supervised technique—is often used for text classification. It uses the training data to create a hyper_plane, a two-dimensional line that effectively divides the categories. We call this hyper plane the decision boundary
- “The innocent” A probabilistic method for classifying data samples, the Bayes classifier[20] makes use of the Bayes theorem.

4. OUTPUT SCREENS

Registeration



Efficient Email phishing detection using Machine learning

SVM, accuracy, naive Bayes, Random Forests

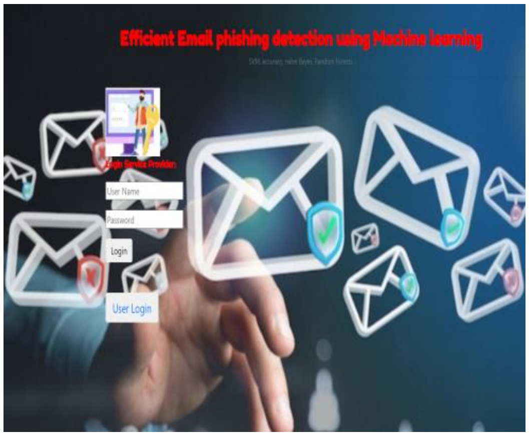
REGISTER YOUR DETAILS HERE!!

Enter Username	User Name	Enter Password	Password
Enter EMail Id	Enter Email	Enter Address	Enter Address
Enter Gender	----Select Gender ----	Enter Mobile Number	Enter Mobile Number
Enter Country Name	Enter Country Name	Enter State Name	Enter State Name
Enter City Name	Enter City Name		Register

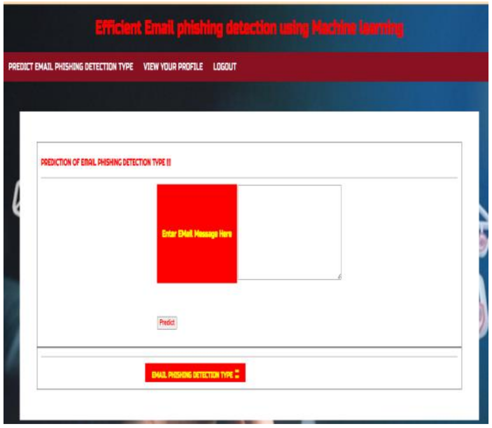
Registered Status :

Remote User Service Provider

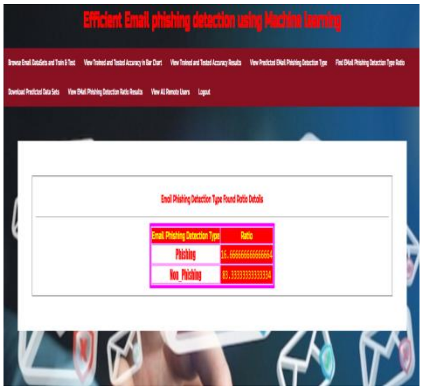
Login page


Efficient Email phishing detection using Machine learning
SVM, accuracy, naive Bayes, Random Forests
Login Service Provider
User Name
Password
Login
User Login

Upload Data


Efficient Email phishing detection using Machine learning
PREDICT EMAIL PHISHING DETECTION TYPE
VIEW YOUR PROFILE
LOGOUT
PREDICTION OF EMAIL PHISHING DETECTION TYPE !!!
Enter EMail Message Here
Predict
EMAIL PHISHING DETECTION TYPE :

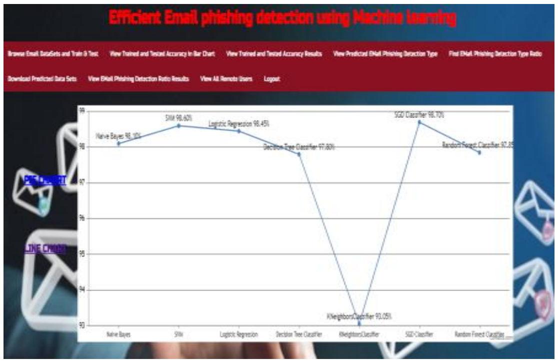
ACCURACY


Efficient Email phishing detection using Machine learning
Browse Email DataSets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Predicted EMail Phishing Detection Type
Find EMail Phishing Detection Type Ratio
Download Predicted Data Sets
View EMail Phishing Detection Ratio Results
View All Remote Users
Logout

Email Phishing Detection Type Found Ratio Details

Email Phishing Detection Type	Ratio
Phishing	16.666666666666664
Non_Phishing	83.33333333333334

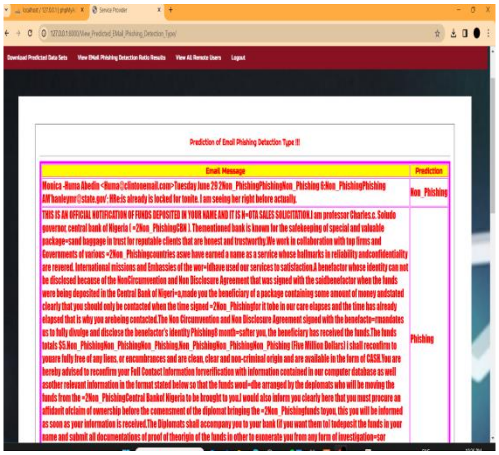
Line Chart


Efficient Email phishing detection using Machine learning
Browse Email DataSets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Predicted EMail Phishing Detection Type
Find EMail Phishing Detection Type Ratio
Download Predicted Data Sets
View EMail Phishing Detection Ratio Results
View All Remote Users
Logout
PIE CHART
LINE CHART
99
98
97
96
95
94
93
Naive Bayes 98.10%
SVM 98.60%
Logistic Regression 98.45%
Decision Tree Classifier 97.80%
KNeighborsClassifier 93.05%
SGD Classifier 98.70%
Random Forest Classifier 97.8
Naive Bayes
SVM
Logistic Regression
Decision Tree Classifier
KNeighborsClassifier
SGD Classifier
Random Forest Classifier

Service Provider Login

Efficient Email phishing detection using Machine learning
SVM, accuracy, naive Bayes, Random Forests
Login Using Your Account
User Name
Password
sign_in
Login Using Your Account
SERVICE PROVIDER
REGISTER

Prediction Details


127.0.0.1:8000/View_Predicted_EMail_Phishing_Detection_Type/
Download Predicted Data Sets
View EMail Phishing Detection Ratio Results
View All Remote Users
Logout

Prediction of Email Phishing Detection Type !!!

Email Message	Prediction
Monica -Huma Abedin <Huma@clintonemail.com>Tuesday June 29 2Non_PhishingPhishingNon_Phishing 6:Non_PhishingPhishing AM'hanleymr@state.gov'; HRe:is already is locked for tonite. I am seeing her right before actually.	Non_Phishing
THIS IS AN OFFICIAL NOTIFICATION OF FUNDS DEPOSITED IN YOUR NAME AND IT IS N=OTA SALES SOLICITATION.I am professor Charles.c. Solude governor, central bank of Nigeria (=2Non_PhishingCBN). Thementioned bank is known for the safekeeping of special and valuable package=sand baggage in trust for reputable clients that are honest and trustworthy.We work in collaboration with top firms and Governments of various =2Non_Phishingcountries aswe have earned a name as a service whose hallmarks in reliability andconfidentiality are revered. International missions and Embassies of the wor=ldhave used our services to satisfaction.A benefactor whose identity can not be disclosed because of the NonCircumvention and Non Disclosure Agreement that was signed with the saidbenefactor when the funds were being deposited in the Central Bank of Nigeri=a,made you the beneficiary of a package containing some amount of money andstated clearly that you should only be contacted when the time signed =2Non_Phishingfor it tobe in our care elapses and the time has already elapsed that is why you arebeing contacted.The Non Circumvention and Non Disclosure Agreement signed with the benefacto=rmandates us to fully divulge and disclose the benefactor's identity Phishing6 month=safter you, the beneficiary has received the funds.The funds totals $5.Non_PhishingNon_PhishingNon_Phishing,Non_PhishingNon_PhishingNon_Phishing (Five Million Dollars) i shall reconfirm to youare fully free of any liens, or encumbrances and are clean, clear and non-criminal origin and are available in the form of CASH.You are hereby advised to reconfirm your Full Contact Information forverification with information contained in our computer database as well asother relevant information in the format stated below so that the funds woul=dbe arranged by the deplomats who will be moving the funds from the =2Non_PhishingCentral Bankof Nigeria to be brought to you.I would also inform you clearly here that you must procure an affidavit ofclaim of ownership before the comensment of the diplomat bringing the =2Non_Phishingfunds toyou, this you will be informed as soon as your information is received.The Diplomats shall accompany you to your bank (if you want them to) todeposit the funds in your name and submit all documentations of proof of theorigin of the funds in other to exonerate you from any form of investigation=sor	Phishing

5. CONCLUSION

This study suggests a smart method for efficiently identifying phishing emails. It compares and contrasts SVM, Random Forests, and Naive Bayes. When it comes to email phishing, finding the most intelligent classification model is the main objective. To assess how well each classifier performed, separate experiments were run on each of the three benchmarking testing levels. In the future, we want to evaluate SVM's efficacy using other benchmarking datasets. Also included is a comparison of SVM's performance with other kernels, including sigmoid and Gaussian kernels.

6.REFERENCES

- Aleroud and L.Zhou, "Phishing environments,techniques,and countermeasures: A survey," Computers & Security, vol. 68, pp. 160-196, 2017.

- Vayansky and S. Kumar, "Phishing–challenges and solutions," Computer Fraud & Security, vol. 2018, pp. 15-20, 2018.

- E. J. Williams, et al., "Exploring susceptibility to phishing in the workplace," International Journal of Human-Computer Studies, vol. 120, pp. 1-13, 2018.

- Odeh,etal.,"MachineLearningTechniquesfor Detection of Website Phishing: A Review for Promises and Challenges," in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0813-0818.

- Odeh, et al., "Efficient Detection of Phishing Websites Using Multilayer Perceptron," 2020.

- Odeh, et al., "PHIBOOST-a novel phishing detection model using Adaptive boosting approach," Jordanian Journal of Computers and Informati