DETECTING PHISHING WEBSITES USING MACHINE LEARNING

Mrs.V.Sunnetha [1], Buddha Sowjanya[2], Gembali Mounika[3], Allu Anusri[4], Bavirisetti Devika[5]
[1]Assistant Professor, Department of Computer Science and Engineering, Vignan 's Institute of Engineering for Women, Visakhapatnam, India
[2-5]Department of Computer Science and Engineering, Vignan 's Institute of Engineering for Women, Visakhapatnam, India

ABSTRACT

One of the most frequent threats to internet consumers is phishing via URLs. In this kind of assault, as opposed to software bugs, the attacker takes advantage of human weakness. The easiest method of obtaining sensitive information from unwitting users is through a phishing attack. The goal of phishers is to obtain crucial data, such as login, password, and bank account information. It preys on both private individuals and corporate entities, tricking them into clicking on URLs that appear safe in order to steal sensitive data or introduce malware into our system. Researchers are always looking for ways to make current models work better and be more accurate. The machine learning approach has been used for this, along with the datasets and URL features that have been used to train the machine learning models.

INTORDUCTION

Internet consumers lose billions of dollars each year due to phishing. In order to fish for personal information in a pool of unwary Internet users, identity thieves use luring strategies. To acquire usernames and passwords for financial accounts as well as personal information, phishers employ faked emails and phishing software. The topic of this study is how to use machine learning techniques to analyse different characteristics of legitimate and phishing URLs in order to identify phishing Web sites.

We go over techniques for phishing site detection based on lexical traits, host qualities, and page importance properties. In order to better comprehend the structure of URLs that spread phishing, we take into consideration a variety of machine learning methods for the evaluation of the features. The refined parameters are helpful in choosing the best machine learning algorithm for differentiating between legitimate and fraudulent websites. In order to collect sensitive information, the crooks first make unofficial copies of legitimate websites and emails, typically from banking institutions or other businesses that deal with financial information. The email will be made utilising the slogans and logos of an authentic company.

The "spoofed" emails are then distributed to as many people as possible in an effort to deceive them. Customers are routed to a fake website that appears to be from the real company when they receive these emails or click a link in them.

LITERATURE SURVEY

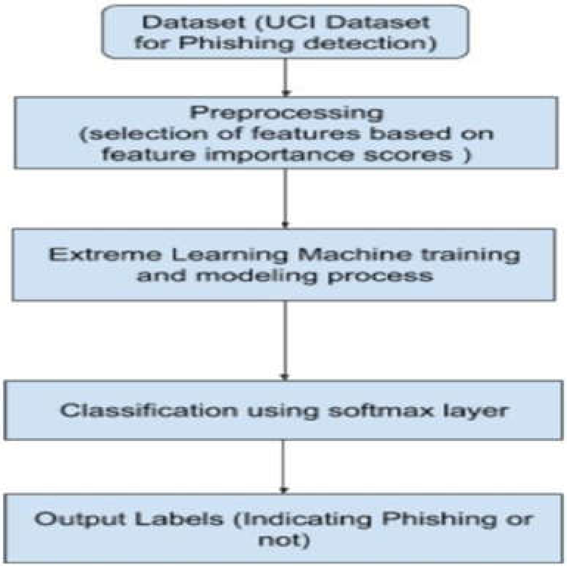
[1] The classification techniques chosen by the authors **Rishikesh Mahajan and IrfanSiddavatam** were Decision Tree, Random Forest, and Support Vector Machine. Their dataset includes 17,058 benign URLs and 19,653 phishing URLs, each with 16 features, that were gathered from the Alexa website and PhishTank, respectively. The training and testing datasets were split 50:50, 70:30, and 90:10, respectively, into the training and testing sets. As performance evaluation metrics, the accuracy score, false negative rate, and false positive rate were taken into consideration. They used the Random Forest method, which had the lowest false negative rate, and obtained 97.14% accuracy. The study found that using additional data during training improved accuracy

[2] Based on features retrieved from the lexical structure of the URL, **Jitendra Kumar** et alstudy .'s trained a variety of classifiers, including Logistic Regression, Naive Bayes Classifier, Random Forest, Decision Tree, and K-Nearest Neighbor. In order to address the problems of data imbalance, biassed training, variation, and overfitting, they designed the dataset of URLs. The dataset was further divided into training and testing in the ratios of 7:3 and comprised an equal number of labelled phishing and authentic URLs. The Naive Bayes Classifier came out to be more appropriate because it had the greatest AUC value, even though all the classifiers had about the same AUC (area under the ROC curve). With a precision of 1, Naive Bayes has the maximum accuracy of 98%.
[3] Using the UCI dataset, Abdulhamit Subasi et al. published an intelligent phishing detection method in [16]. Artificial Neural Networks (ANN), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), C4.5 Decision Tree, Random Forest (RF), and Rotation Forest (RoF) were some of the machine learning methods utilised as classifiers for phishing website identification. The proposedRF classifier outperformed the competition in terms of accuracy, F-measure, and AUC. Compared to the other classifiers, RF was more accurate, reliable, and rapid.
[4] Using Random Forest and Decision Tree, **Mohammad Nazmul Alam** et al. [15] devised a solution to identify phishing assaults. Together with feature selection approaches like principal component analysis, the 32-feature Kaggle dataset was employed (PCA). The dataset's redundant, irrelevant, or unneeded data is reduced using feature selection. Prior to applying PCA, the suggested model used the REF, Relief-F, IG, and GR algorithms for feature selection. 97% accuracy was attained using Random Forest. It handled the over-fitting issue and had reduced variance.
[5] A machine-learning based phishing detection system was proposed by **Mehmet Korkmaz** et al. in [14] employing 8 distinct algorithms on 3 different datasets. Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF), and Artificial Neural Network were the techniques employed (ANN). It was found that the accuracy rate of the models utilising LR, SVM, and NB is poor. The NB, DT, LR, and ANN algorithms produced better outcomes in terms of training time. They came to the conclusion that RF or ANN algorithms might be used since they require less training time and have a higher accuracy rate.

PROPOSED SYSTEM

By extracting useful feature representations of URLs and training a Malicious URL Detection Using Machine Learning prediction model using training data of both malicious and benign URLs, these approaches attempt to assess the information of a URL and its connected websites or webpages. Static features and dynamic features are the two sorts of features that can be used. With static analysis, we examine a webpage based on information that is already available, rather than running the URL. Lexical features from the URL string, host-related data, and occasionally even HTML and JavaScript content are among the aspects that can be extracted. These techniques are safer than dynamic alternatives because execution is not necessary.The fundamental presumption is that malicious and benign URLs have distinct distributions of these properties. This distribution data can be used to create a prediction model that can forecast new URLs. Static analysis techniques have been thoroughly investigated by applying machine learning techniques because to the comparatively safer environment for obtaining crucial information and the capacity to generalise to all forms of threats Using accurate feature representations of URLs, this approach attempts to assess the information contained in a URL and the websites or webpages it corresponds to.
•using training data of both harmful and safe URLs to develop a prediction model.
• A prediction model that can make predictions about new URLs can be created using this distributional information.

PROPOSED SYSTEM ARCHITECTURE


Dataset (UCI Dataset for Phishing detection)
Preprocessing (selection of features based on feature importance scores)
Extreme Learning Machine training and modeling process
Classification using softmax layer
Output Labels (Indicating Phishing or not)

RESULTS AND DISCUSSIONS

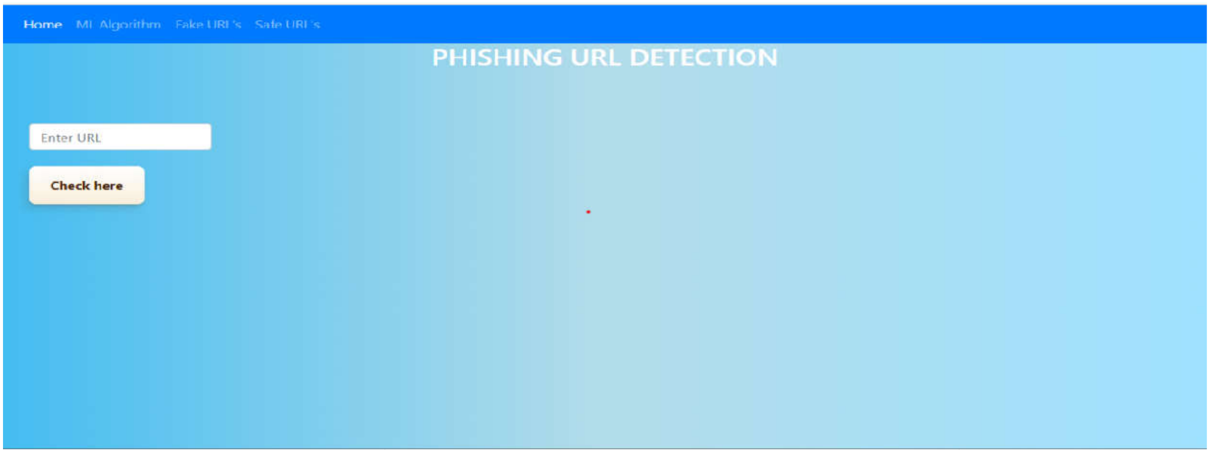
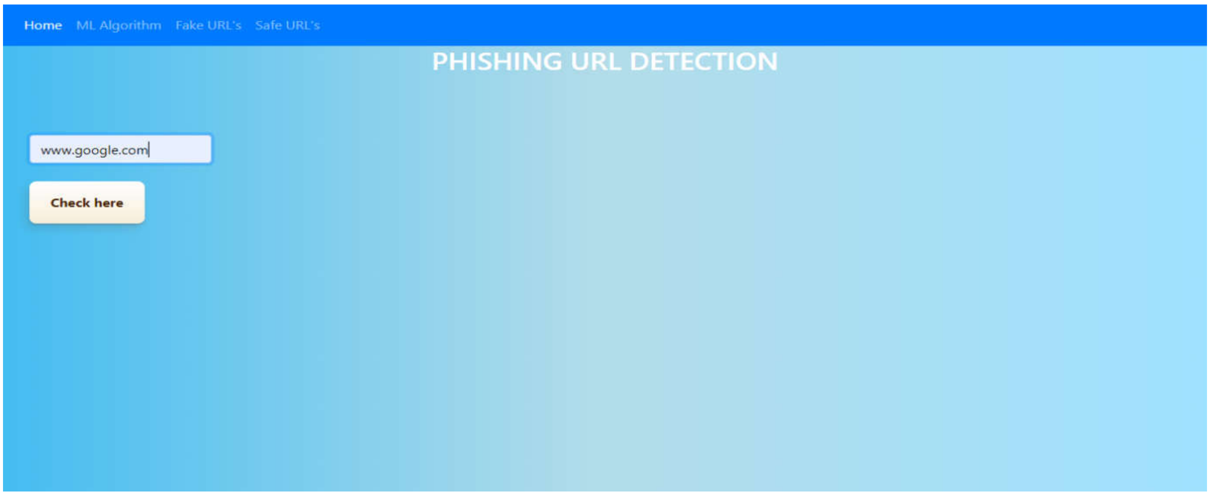
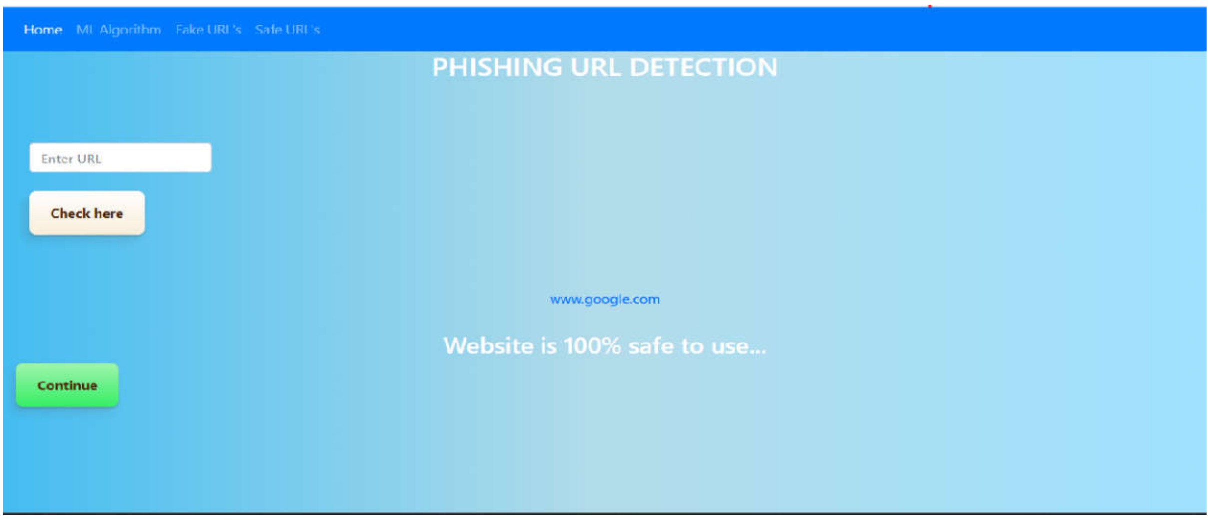


Fig 1 Home page



www.google.com

Fig 2 Input the url

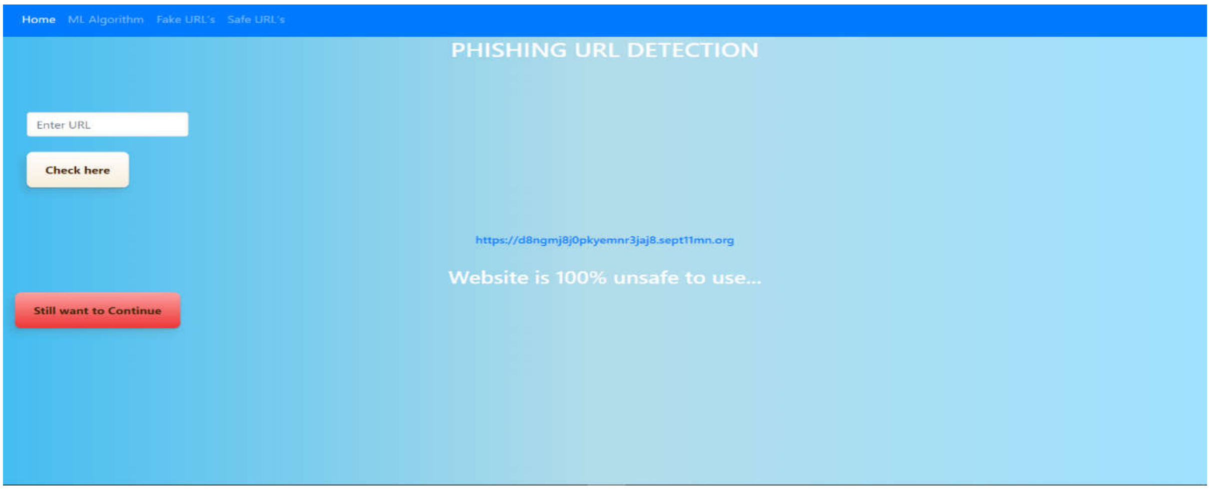


PHISHING URL DETECTION

www.google.com

Website is 100% safe to use...

Fig 3 predicting the url is safe



PHISHING URL DETECTION

https://d8ngmj8j0pkyemnr3jaj8.sept11mn.org

Website is 100% unsafe to use...

Fig 4 predicting the url is unsafe

CONCLUSION

It is remarkable that a good adversary of the phishing tool should be able to predict the phishing attacks in a reasonable amount of time. We acknowledge that in order to increase the size of the phishing site detection area, the availability of good anti-phishing tools at a reasonable time scale is also essential. Consistent retraining should be used to continuously enhance this device. We can actually retrain our model frequently and manage any changes in the highlights, which are important in determining the site class, thanks to the availability of clean and modern training datasets that can be acquired using our own device [30, 32]. Albeit neural system demonstrates its capacity to tackle a wide assortment of classification issues, the procedure of finding the ideal structure is very difficult, and much of the time, this structure is controlled by experimentation.Our model takes care of this issue via computerizing the way toward organizing a neural system conspire; hence, on the off chance that we construct an enemy of phishing model and for any reasons we have to refresh it, at that point our model will encourage this procedure, that is, since our model will mechanize the organizing procedure and will request scarcely any client defined parameters.

FUTURE SCOPE

Gradient Boosting Classifier currectly classify URL upto 97.4% respective classes and hence reduces the chance of malicious attachments.

REFERENCES

1. Liu J and Ye Y (2001) A brief introduction to e-business operators, including information on commercial centre arrangements, security concerns, and market demand. London, UK: E-business experts, commercial centre arrangements, security concerns, and market interest
2. Aaron G. Manning and the APWG (2013) APWG phishing reports. January 1, 2013, APWG.[Online].You can access it at http://www.antiphishing.org/assets/apwg-reports. . Accessed on 8 February 2013
3. Kaspersky Lab (2013) Spam in January 2012: love, politics, and games. [Online].Accessible at: http://www.kaspersky.com/about/news/spam/2012 Spam in January 2012 Love Politics and Sport. Accessed February 11, 2013
4. Seogod (2011) Black Hat SEO. Internet marketing instruments. [Online]. http://www.seobesttools.com/dark cap website optimization is available. . As of January 8, 2013, Dhamija R, Tygar JD, and Hearst M (2006) explain why phishing is effective. Human considerations in figure frameworks, Cosmopolitan Montré'al, Canada, Proceedings of the SIGCHImeetin
5. Dhamija R, Tygar JD, Hearst M (2006) Why phishing works. In: Proceedings of the SIGCHImeeting on human factors in figuring frameworks, Cosmopolitan Montre 'al, Canada
6. Miyamoto D, Hazeyama H, Kadobayashi Y (2008) An assessment of AI based techniquesfor recognition of phishing destinations. Aust J Intell Inf Process Syst 10(2):54–6