PHISHING URL DETECTION

PROJECT II (4IT32)

INFORMATION TECHNOLOGY DEPARTMENT

BIRLA VISHVAKARMA MAHAVIDYALAYA ENGINEERING COLLEGE

(AN AUTONOMOUS INSTITUTION)

ABSTRACT

Nowadays use of internet is increased drastically because most of the people have smart phone , by which they can easily connected with the network. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus web pages. These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc.In this project, we proposed a machine learning-based phishing url detection system by using sixteen different feature to analyze the URLs and we use Kaggle dataset to train our model.We make website and chrome extension to check the url.

INTRODUCTION

Phishing is the most commonly used social engineering and cyber attack. Through such attacks, the phisher targets naïve online users by tricking them into revealing confidential information, with the purpose of using it fraudulently. In order to avoid getting phished, users should have awareness of phishing websites.This System is machine learning based model.We use Kaggle as our dataset.We extract 16 different feature from those url and train our model.We use decision tree, svm and ann to train the model and we got maximum accuracy from ann. So, we decide to use ann as our machine learning algorithm,and make api of this model to use in our website. We make one website, with the help of that website user can check the url and get one alert box if url is phishing. On that web site, we also provoide different information about current cyber attacks, use of different tool and also different government website on which people report their complain about the cyber attack which they faced. We also make chrome extension.When user visit website at that time user can check website using that extension.

METHODOLOGY

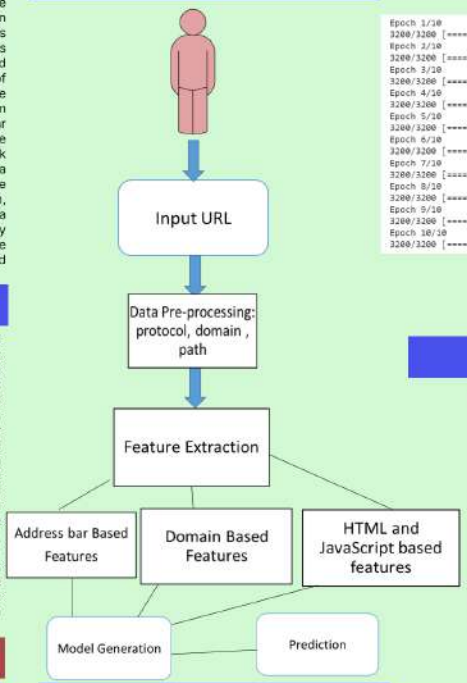
DATA PREPROCESSING

It is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.
Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.
To execute this, we will first get the dataset then import libraries and encode categorial data then split into training and test data.

1. Handling missing values in dataset.
2. Removing constant features, outliers, etc.
3. Min-Max Scaler shrinks the data within the given range, usually of 0 to 1.

It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

System Flow Diagram



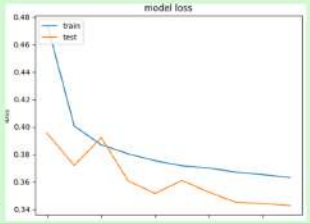
MODEL TRAINING

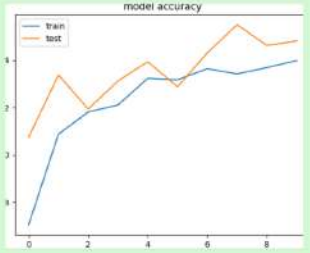
After building the model, we trained it.
This is the stage where the ML algorithm is trained by feeding datasets. This is the stage where the learning takes place. Consistent training can significantly improve the prediction rate of the ML model. The weights of the model must be initialized randomly. This way the algorithm will learn to adjust the weights accordingly.
So, to train the model first of all we splitt the data set into training and testing samples.
In this stage the most important term is epoch and batch size.
Epoch means when all the training data is used at once and is defined as the total number of iterations of all the training data in one cycle for training the machine learning model.
And batch size refers to the number of training examples utilized in one iteration.
We define epoch equals to 10 and batch size equals 0.2.

```
Epoch 1/10
3200/3200 [==============================] - 8s 2ms/step - loss: 0.4745 - accuracy: 0.7705 -
Epoch 2/10
3200/3200 [==============================] - 6s 2ms/step - loss: 0.4006 - accuracy: 0.8087 -
Epoch 3/10
3200/3200 [==============================] - 8s 2ms/step - loss: 0.3870 - accuracy: 0.8181 -
Epoch 4/10
3200/3200 [==============================] - 7s 2ms/step - loss: 0.3804 - accuracy: 0.8211 -
Epoch 5/10
3200/3200 [==============================] - 7s 2ms/step - loss: 0.3753 - accuracy: 0.8323 -
Epoch 6/10
3200/3200 [==============================] - 6s 2ms/step - loss: 0.3717 - accuracy: 0.8317 -
Epoch 7/10
3200/3200 [==============================] - 8s 2ms/step - loss: 0.3700 - accuracy: 0.8364 -
Epoch 8/10
3200/3200 [==============================] - 8s 3ms/step - loss: 0.3671 - accuracy: 0.8342 -
Epoch 9/10
3200/3200 [==============================] - 7s 2ms/step - loss: 0.3653 - accuracy: 0.8369 -
Epoch 10/10
3200/3200 [==============================] - 8s 2ms/step - loss: 0.3631 - accuracy: 0.8398 -
```

Fig 3: Model being Trained

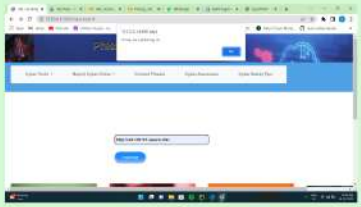
MODEL EVALUATION

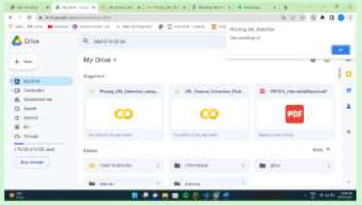




	precision	recall	f1-score	support
0	0.76	0.98	0.85	978
1	0.97	0.70	0.81	1022
accuracy			0.83	2000
macro avg	0.86	0.84	0.83	2000
weighted avg	0.86	0.83	0.83	2000

RESULTS





CONCLUSION

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing beside , but after our survey that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also we generated classifiers through several machine learning algorithms. We have found that our system provides us with 80.75 % of accuracy for Decision Tree Classifier, 79.60% accuracy for SVM Classifier, 81.35 % accuracy for Random Forest Classifier and finally 83.40% of accuracy when using ANN. Hence we found that the best among all the above classifiers is ANN which shows maximum accuracy. The proposed technique is much more secured as it detects new and previous phishing sites.

REFERENCES

[1]International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 8 Issue 11, November-2019. "Detection of URL based Phishing Attacks using Machine Learning".

[2]International Journal of Engineering Research & Technology (IJERT), ISSN: 2456-3307, Vol. 3 Issue, 2019. "Phishing Website Detection using Machine Learning : A Review"

[3]International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 9 Issue 04, April-2020. "Review on Phishing Sites Detection Techniques"

TEAM

Guide:
Prof. Priyank N Bhojak

Team Members:
1. Setu Patel (19IT404)
2. Kandarp Gandhi (19IT410)
3. Deep Panara (19IT412)