International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Identifying the User As Genuine/Malign Based on Search Logs and Search History

D. Satya Bhavani, P. RajyaLakshmi Sobha Pavani

Abstract-One of the major challenges a developer may face is security issues/threats on the labelled data. The labelled data comprises of system logs, network traffic or any other enriched data with threat/not threat classification. . There were few studies which categorized the URLs to a specific category like Arts, Technology, etc. In this paper the main research is on the classification of users based on the search logs(URLs). Manually it is difficult to differentiate the user based on search logs. So, we train a machine learning model that takes raw data as input and classifies the user to genuine or malign. This model helps in intrusion detection/suspicious activity detection. For this first we gather data of past malicious URLS as training set for Naïve Bayes algorithm to detect the malicious users. By implementing KNN algorithm effectively we can detect the malign users up to an accuracy of 94.28%. With the help of Machine Learning algorithms like Naïve Bayes, KNN, Random Forest classifiers we can classify the malign and genuine users.

Keywords: URL, Malign, Naïve Bayes, KNN, Random Forest intrusion detection

I. INTRODUCTION

The main aim of the proposed work is to classify the users based on URLs, who can harm others socially or financially. In this context first we need to understand the web security and phishing. Phishing is a cybercrime in which the attacker try to extract information from the user. To solve that purpose the attacker creates fake and malicious webpages and

URLs. When the user frequently visits that page his data is extracted by the attacker and the attacker frequently accesses that page to collect the data.

In this proposed work a Machine Learning model is implemented to classify the User based on his search logs or URLs searched. With this trained model we can classify the user to genuine/malign. The classification is done effectively and efficiently to provide the most accurate results.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
D. Satya Bhavani, Assistant Professor Department of Computer Science and Engineering and Engineering Mahatma Gandhi Institute of Technology of TechnologyHyderabad, India
P. RajyaLakshmi Sobha Pavani, IV/IV B.Tech Department of Computer Science and Engineering and Engineering Mahatma Gandhi Institute of Technology of TechnologyHyderabad, India

For this we first gather the data of malicious and benign URLs and use it as the training set for machine learning algorithms like Naïve Bayes, KNN to classify the users. The algorithms used are able to classify the users with an accuracy of around 94.28%.

II. METHODOLOGY

A. Architecture

In the proposed work the model takes the input log data and it is pre-processed to find any inconsistent and null data.

Then it is split into training and testing to fit it into the classifier. The Naïve-bayes classifier is trained against training data and is tested using the test data and the accuracy is predicted. Therefore it classifies the user as genuine or malign based on the search logs and search history .The KNN classifier is also implemented for classification of the user to genuine or malign and later the results are compared. The input data is split into train and test using the modules present in the python sklearn module. The Fig.1 represents the system architecture

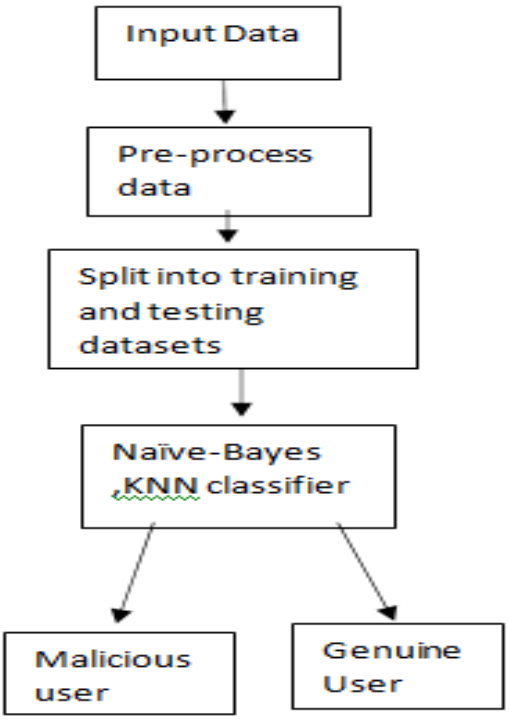

Fig.1:System Architecture

B.DATA ACQUISITION

Data acquisition is the process of importing raw data sets into your analytical platform. It can be acquired from traditional databases (SQL and query browsers), remote data (web services), text files (scripting languages), NoSQL storage (web services, programming interfaces), etc. The dataset used in the

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (1)$$

The μy and σy parameters were estimated using maximum likelihood..

KNN Classifier

K-Nearest Neighbor is one of the Supervised Machine Learning technique. It assumes the similarity between the new datapoint and the available datapoints and places the new datapoint into the category that is most similar. This algorithm stores all the datapoints and classifies the new data point based on the similarity. K-NN algorithm can be used for both Regression and Classification.

D.DATA VISUALIZATION

Data Visualization is the graphical representation of data. This is done using graphs, charts, etc which are used to find the data patterns and outliers present in the data.

project is collected from Kaggle. This data was obtained from different sources of benign and malicious URL's. The final dataset which serves as the training set for machine learning model consists of around 1780 values.

C.CLASSIFICATION MODELS

Naïve Bayes Classifier

It is a classification technique which is based on Bayes theorem which assumes the independence of predictors. It assumes that the dependence of one feature in class is not related with the other feature. It can be extended to real-valued attributes by assuming a Gaussian distribution. This extension is called Gaussian Naive Bayes.

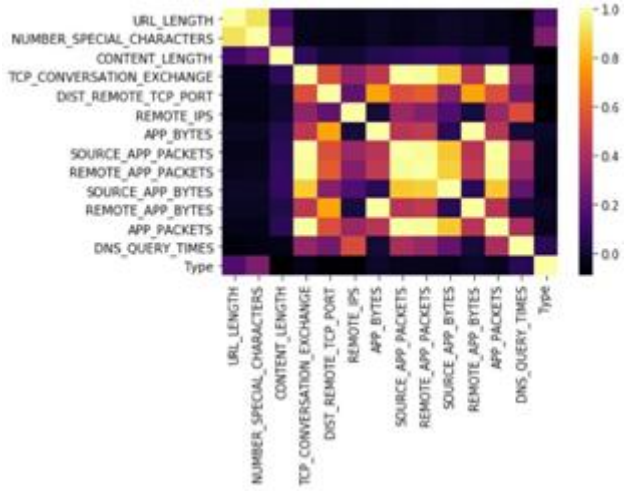

Fig.2:Heatmap constructed on the fields present in the data

The heatmap is used to show the correlation between different fields in the dataset i.e., the interdependence of the fields on each other.

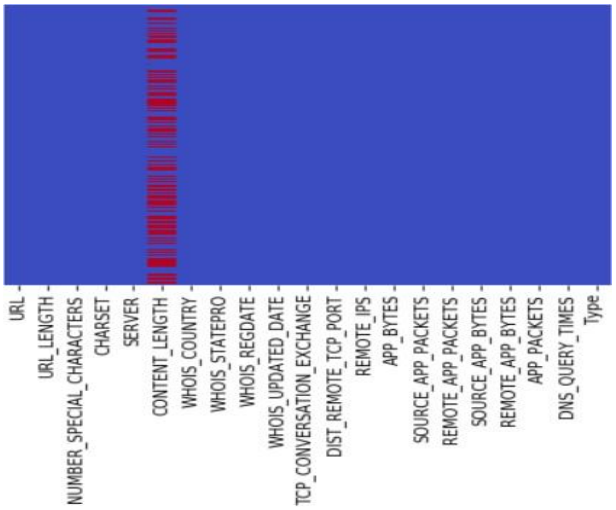

Fig.3:Plot that shows the null values present in the dataset

III. RESULTS

A. Testing

It is a process of executing a program to find the errors if present. If the testing process is done successfully it removes all the errors from the software.

Table.1.1: Black box Testing

Input	Actual Output	Predicted Output
[16,6,324,0,0,0,22, 0,0,0,0,0,0]	0	0
[16,7,263,7,0,2,700,9, 10,1153,832,9,2	1	1

B. Model Accuracy

The naïve bayes classifier and knn classifier is tested against several test datasets.. The models are trained to get a good accuracy. The models trained has the following specifications:

```
gnb = GaussianNB()
gnb.fit(train_img,train_labels)
```

```
GaussianNB(priors=None, var_smoothing=1e-09)
```

Fig.4:Naïve Bayes specifications

```
from sklearn.neighbors import KNeighborsClassifier

#Create KNN Classifier
knn = KNeighborsClassifier(n_neighbors=2)

#Train the model using the training set
knn.fit(train_img, train_labels)
```

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=2, p=2,
           weights='uniform')
```

Fig.5:KNN Classifier Specifications

```
from sklearn.metrics import accuracy_score
preds = clf.predict(val_img)
print("Accuracy:", accuracy_score(val_labels,preds))
```

```
Accuracy: 0.9828571428571429
```

Fig.6: Accuracy of Random Forest Classifier

```
print("Accuracy:",accuracy_score(val_labels,y_pred))
```

```
Accuracy: 0.2914285714285714
```

Fig.7:Accuracy of Naïve Bayes classifier

```
print("Accuracy:",accuracy_score(val_labels,y_pred))
```

```
Accuracy: 0.9428571428571428
```

Fig.8:Accuracy of KNN classifier

From the two models trained it is observed that the KNN classifier performs better than the Gaussian Naïve Bayes classifier. The accuracy of KNN is 94% whereas that of for Naïve bayes is 29%.

Table.1.2:Positive Test Cases

Input id	Actual Output	Predicted Output	Remarks
test_sample	[1]	[1]	Success
test_sample1	[0]	[0]	Success

Table.1.3:Negative Test Cases

Input id	Actual output	Predicted output	Remarks
test_sample2	[0]	[1]	Failure
test_sample3	[1]	[0]	Failure

IV. CONCLUSION AND FUTURE SCOPE

A. Conclusion

performed to get even better results. The model needs to be updated in finite intervals to be more accurate. The more increase the dataset helps in faster suspicious or intrusion detection activity which further increases the security. This can further be extended to obtain the IP address of that user. Identifying the user as genuine/malign based on search logs project is completed successfully. The goal of the project is achieved. Gaussian Naïve Bayes Classifier and KNN Classifier are successfully developed to fit a training data . The trained Gaussian Naïve Bayes model now predicts the user up to an accuracy of 29.14%.The model trained using KNN classifier now predicts the user up to an accuracy of 94.28%.The same is also implemented using Random Forest Classifier which predicts up to an accuracy of 92%.Therefore KNN classifier identifies the user behavior more accurately.

B. Future Scope

A comparison between various machine learning algorithms such as CNN(Convolutional Neural Network),Ada Boost, Bagging, etc can be

(TIST), Vol, 4, 2013.

REFERENCES

1. J.R. Wen, J.Y. Nie, and H.J. Zhang, Query Clustering Using User Logs, ACM Trans. on Information Systems, Vol. 20, No. 1, 2002, pp. 59- 81.
2. J. Yi and F. Maghoul, Query Clustering Using Click-through Graph, In Proc. 18th Int'l Conf. World Wide Web (WWW), 2009.
3. Kenneth Wai-Ting Leung, and Dik Lun Lee. Deriving Concept-based User profiles from Search Engine Logs. IEEE Trans. Knowledge and Data Engineering, Vol. 22, No. 7, pp. 969- 982, July 2010.
4. Mazur, P. Serdyukov, and Y. Ustinovskiy, "Intent-Based Browse Activity Segmentation", IR, Russia, pp. 242-253, 2013. [the 35th European Conference on IR Research, Russia, 2013]
5. D. Jiang, H. Li, and J. Pei, "Mining Search and Browse Logs for Web Search: A Survey", ACM Transactions on Intelligent Systems and Technology
6. Veningston .k and dr. R. Shanmugalakshmi, "Personalized Grouping of User Search Histories for Efficient Web Search", Applied Computational Science, 2014.
7. I. Rish, "An Empirical Study of the Naïve Bayes Classifier", January 2014.
8. Pouria Kaviani, Mrs. Sunita Dhotre, "Short Survey on Naïve Bayes Algorithm", November 2017.
9. Baoli, L., Shiwen, Y., Qin, L. (2003) "An Improved k-Nearest Neighbor Algorithm for
10. Text Categorization, ArXiv Computer Science e-prints.
11. Sadegh Bafandeh Imandoust, Mohammad Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", September 2013.

AUTHORS PROFILE

Ms. D. Satya Bhavani, an Assistant Professor in the Department of Computer Science and Engineering at Mahatma Gandhi Institute of Technology, Hyderabad. She has a work experience of over 15years. Her areas of interest in research include Computer Networks, Network Security.

Ms. P. RajyaLakshmi Sobha Pavani, a Final year student of Bachelors in technology in the field of Computer Science at Mahatma Gandhi Institute of Technology, Hyderabad. She developed projects in Web technologies, Machine Learning, etc. Her areas of interest include Data Analytics, Machine Learning.