MACHINE LEARNING APPROACH FOR REAL TIME BASED PHISHING WEBSITE DETECTION SYSTEM

By

Lekhani Agarwal (2014IPG-051)
Shivani Mandeliya (2014IPG-083)
Vaishnav Chandak (2014IPG-127)

ABV-Indian Institute of Information Technology & Management Gwalior (M.P.)

September 21, 2017

Outline

विश्वजीवनामृतं ज्ञानम्

Motivation

- As indicated by the measurements, one of 257.9 emails leads to phishing sites in oct 2012. The majority of them focused on the financial, payments and retail benefits.
- According to RSAs Fraud Report, phishing caused an estimated loss of 225 million dollars in 2013.
- MacEwan University lost 11.8 million dollars to scammers in phishing attack.

Phishing Life Cycle

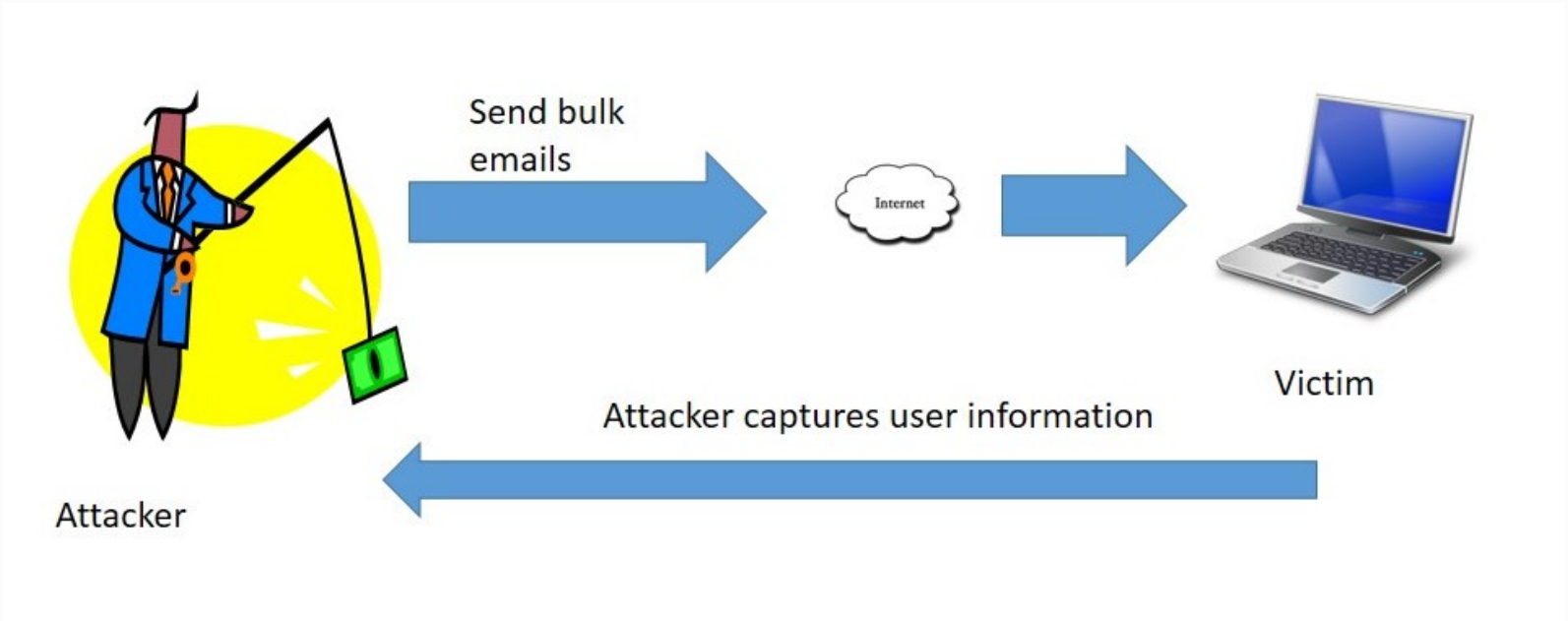


Figure: Phishing Life Cycle

Objective

- To develop a Real Time Phishing Website Detection System to curb the Phishing attacks.
- This system will analyze the Uniform Resource Locator (URL) of every web service.
- Before any exchange of activities happen, it will warn the user to not go through a website if it is found malicious.

Related Works

AUTHOR	YEAR	CONTRIBUTION
W Liu, X Deng, G Huang, AY Fu	2006	Proposed an anti-phishing technique that makes use of visual features of a website. This approach analyses the visual similarity between the current website and the already present benign website.
Jianyi Zhang et al	2010	Content based approach for phishing-detecting. This approach pays attention particularly to the Google Page Rank.
J.H. Huh and H. Kim	2012	Search Engine Based techniques developed for Anti-Phishing(uses the information provided by search engine).
Sadia Afroz, and Rachel Greenstadt	2015	This proposed work, works on the basis of detecting the logo present in the webpage rather than executing full page matching.
Mona Ghotaish Alkhozae	2015	Proposed a phishing detection approach based on checking the webpage source code.
Ankit Jain and B.B. Gupta	2016	An approach to prevent against the Phishing attacks by the use of autoupdating white list.

Gap Analysis

AUTHOR	YEAR	GAP
W Liu, X Deng, G Huang, AY Fu	2006	Most image-based anti-phishing approaches are likely to be influenced by a tremendous variety in the appearance of phishing websites.
Jianyi Zhang et al	2010	In case of newly built websites which can be legitimate or malicious but google pagerank will always give a low value marking it as suspicious.
J.H. Huh and H. Kim	2012	The system fails if the search engine fails to deliver service and deceives in case.
Sadia Afroz, and Rachel Greenstadt	2015	Large storage is needed to stockpile the logos of the websites and it has high number of false negatives.
Mona Ghotaish Alkhozae	2015	This approach is time consuming as it analyze the whole source code of the website.
Ankit Jain and B.B. Gupta	2016	White list is fundamentally a better solution but the administrative effort involved is huge.

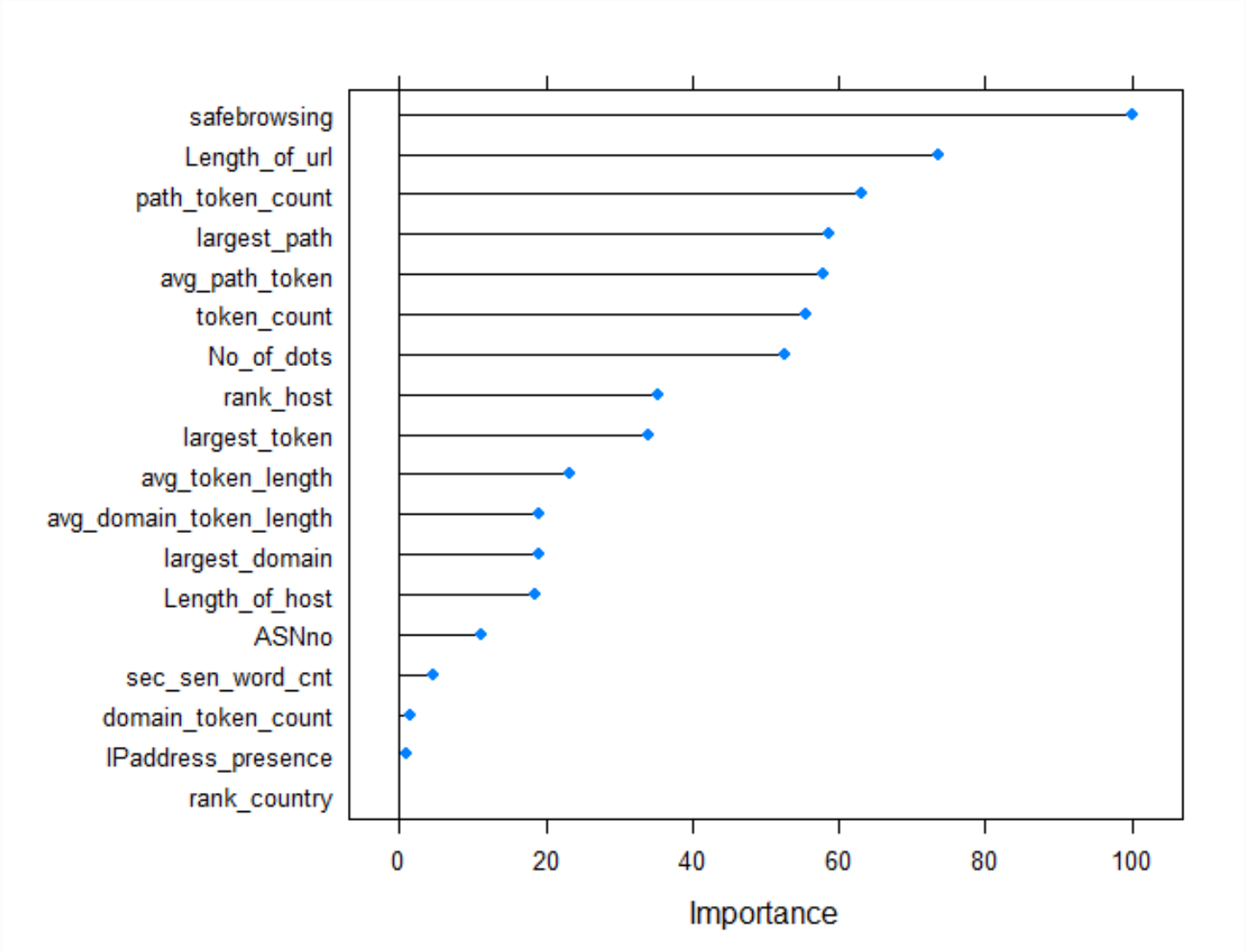
How to proceed?

- **DATA COLLECTION:-** Data consists of Malicious and Benign urls which were collected from www.PhishTank.com and www.Alexa.com respectively
  - MALICIOUS URLs – 1613 tuples
  - BENIGN URLs – 1450 tuples
- **Data Division:**
  Training dataset - 1113 malicious and 1000 benign Urls.
  Test dataset - 500 malicious and 450 benign Urls.
- **Labels:**
  label 1 is to signify Malicious Urls in classification approach.
  label 0 is to signify Benign Urls in classification approach.

Dataset Features

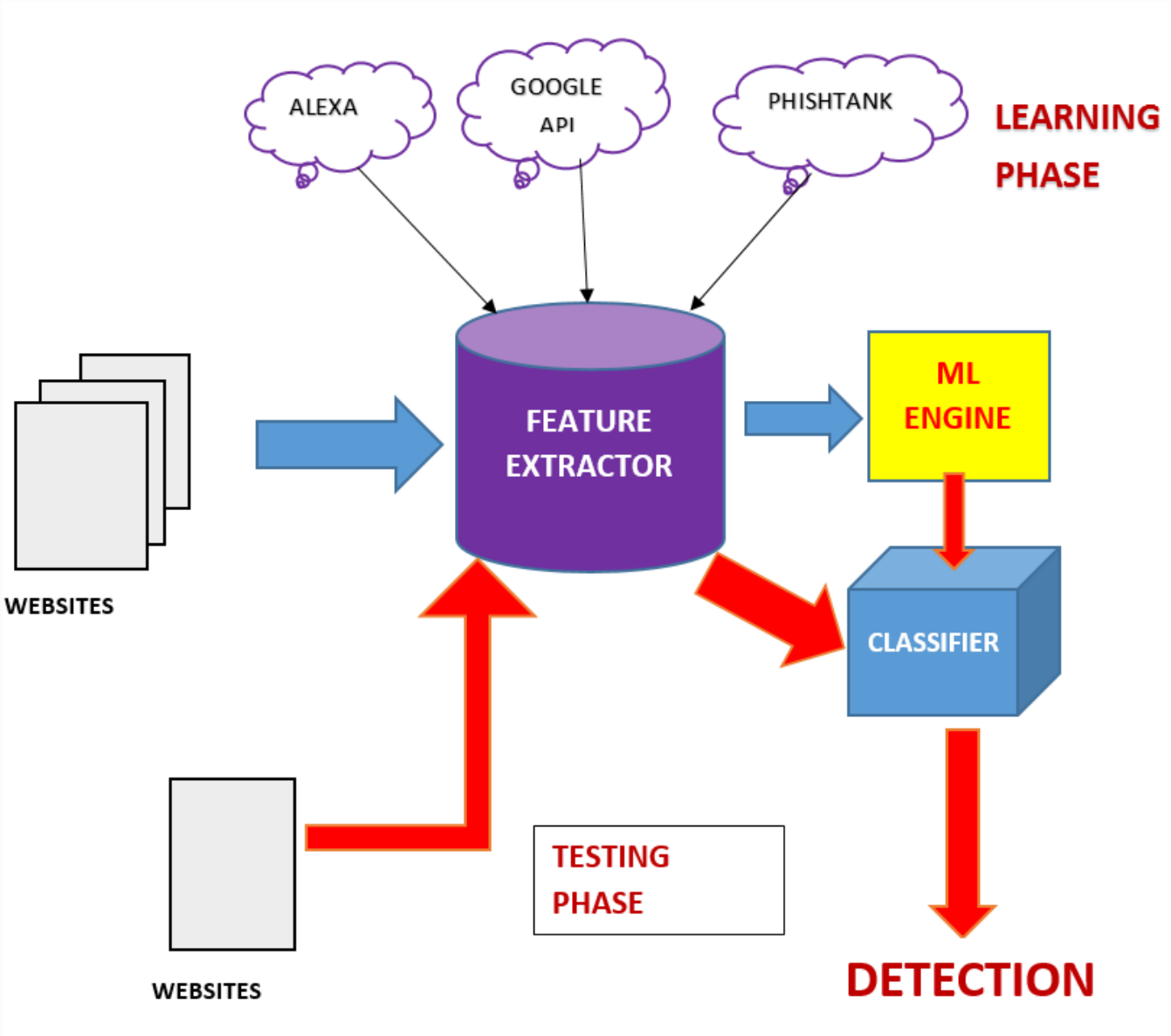
Lexical	Host based	Site popularity
token_count	ASN_no	rank_host
avg_token_length	IPaddress_presence	rank_country
no_of_dots	length_of_host	
length_of_url	host	
avg_path_token_length		
avg_domain_token		
path_token_count		
path		
largest_domain		
domain_token_count		
largest_path		
largest_token		
Sec_sen_word_count		

Feature Importance



विश्वजीवनामृतं ज्ञानम्

Working model



Methodology

MODEL	FEATURES	DRAWBACK
Decision Tree Classifier	**Decision tree learning** uses a decision tree to go from observations about an item to conclusions about the item's target value .	Create over-complex trees that do not generalise the data well. Requires pruning of trees.
Tree Bagging Classifier	It combines the individual predictions from the subsets to form the final prediction for the dataset.	Bagging resamples the training set with replacement, some instance are represented multiple times while others are left out.
Random Forest Classifier	Random forest creates lots of decorrelated decision trees . classification is done by using all decision trees by analyzing the probability of the label class.	overfit for some dataset with noisy classification tasks

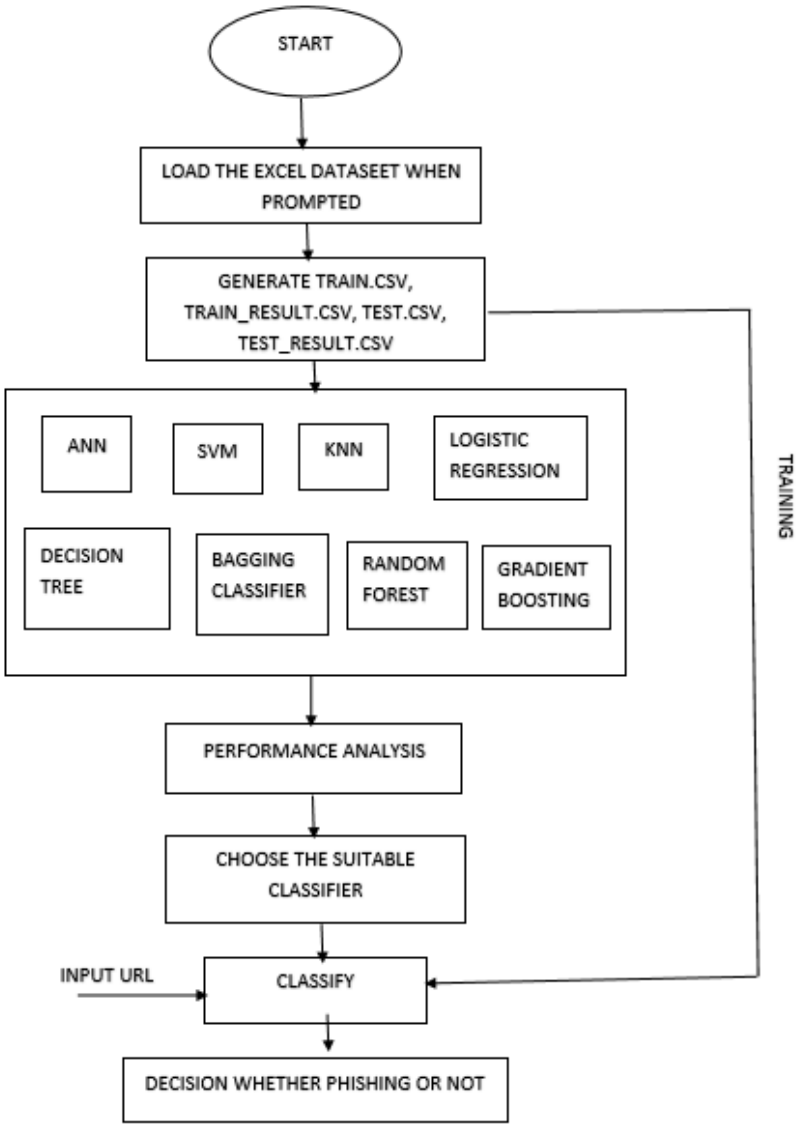
Methodology contd.

MODEL	FEATURES	DRAWBACK
Neural Network	The original goal of the neural network approach was to solve problems in the same way that a human brain would.	Difficult to design a perfect model with number of hidden layers and nodes which will best fit the data. & Overtraining.
Logistic Regression Classifier	Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function"	Problem of overfitting
K-nearest Neighbour	It is based on the clustering of the elements that have the same characteristics.	Finding a good similarity measure Complexity in searching the nearest neighbor for each sample.
Support Vector Machine	A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.	choice of the kernel, another is parameter tuning

Gradient Boosting Classifier

- It is an efficient algorithm to convert relatively weak learner into very strong learner.
- Training builds a series of small decision trees.
- Each tree attempts to correct errors from the previous stage.

Flowchart



विश्वजीवनामृतं ज्ञानम्

Confusion Matrix

	ACTUAL: NO	ACTUAL: YES	
PREDICTED : NO	TRUE NEGATIVE	FALSE NEGATIVE	TN+FN
PREDICTED : YES	FALSE POSITIVE	TRUE POSITIVE	FP+TP
	TN+FP	FN+TP	

Figure: Confusion matrix

$$\text{Accuracy} = (TN+TP)/(TN+TP+FN+FP)$$
$$\text{Precision}(P) = TP/(FP+TP)$$
$$\text{Recall}(R) = TP/(FN+TP)$$
$$\text{F1 Score} = 2PR/(P+R)$$

विश्वजीवनामृतं ज्ञानम्

Results

Model	Accuracy	F1 Score
Decision Tree Classifier	58.32%	70.66
Tree bagging Classifier	61.16%	72.60
Random Forest	66.32%	75.30
Neural Network	88.74%	88.35
Logistic Regression Classifier	90.63%	90.39
K-Nearest Neighbor	91.58%	91.90
Support Vector Machine	93.05%	93.15
Gradient Boosting Classifier	95.26 %	95.54

Figure: Accuracies of our applied algorithms

So,the best option is GRADIENT BOOSTING CLASSIFIER

विश्वजीवनामृतं ज्ञानम्

GUI

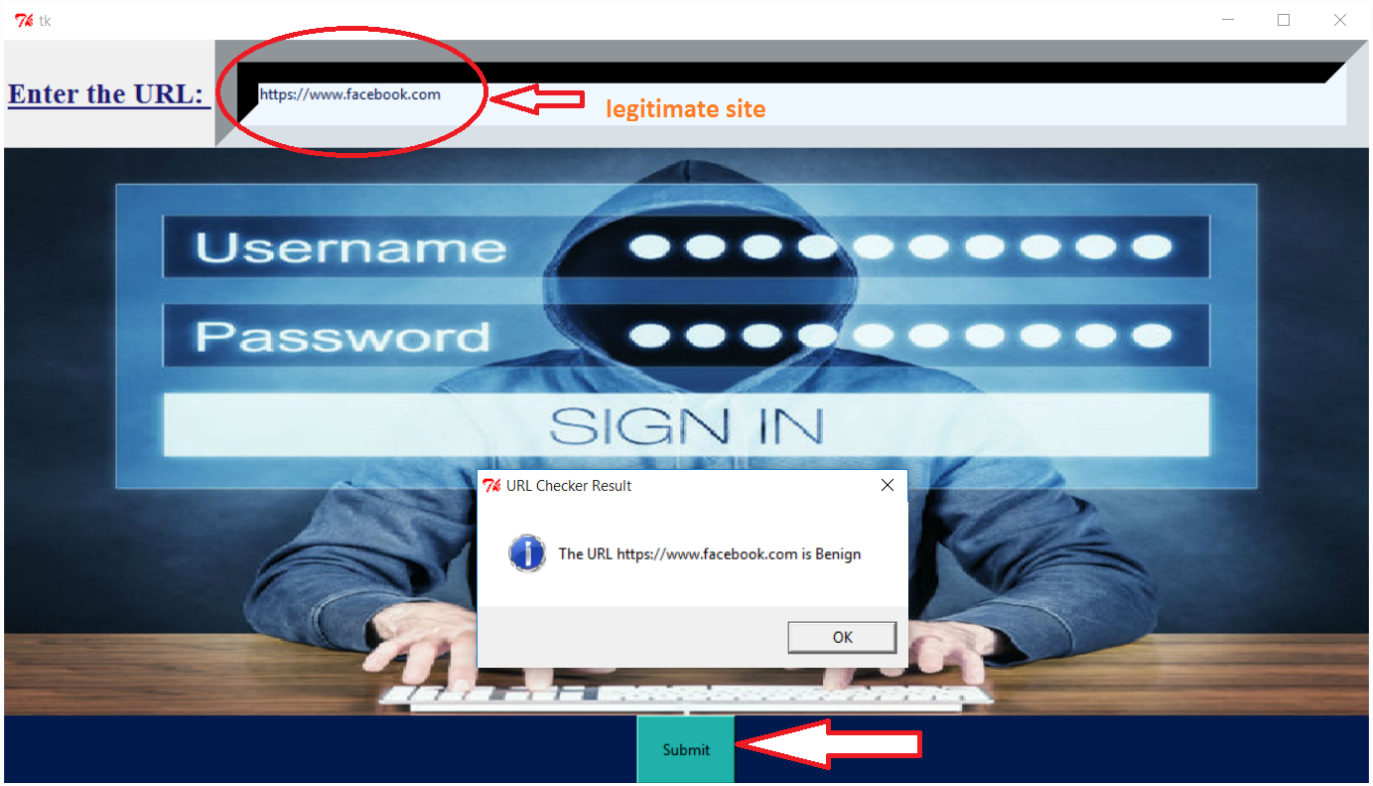


Figure: GUI output for benign sites

विश्वजीवनामृतं ज्ञानम्

GUI Contd.

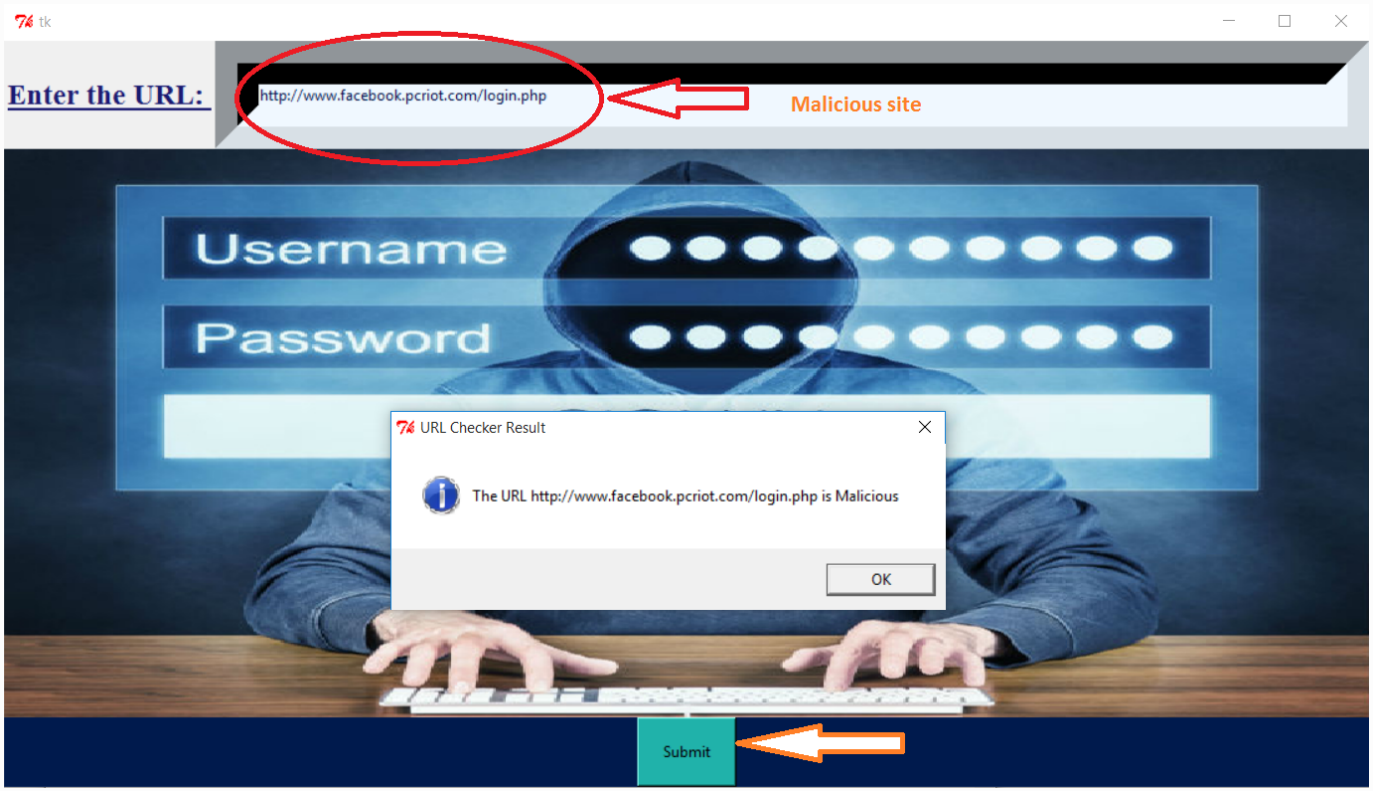


Figure: GUI output for malicious sites

Conclusion

We have implemented a Real-Time URL based phishing detection system for end user with the following characteristics.

- The GUI of our system engages end users.
- Provides user-friendly environment.
- Analyzes Uniform Resource Locator (URL) itself without accessing of Web sites.
- No runtime idleness.

Future Scope

- Develop a browser plug-in which can work online.
- Incorporate different parts of web based learning and assembling information to see the new patterns in phishing exercises.

REFERENCES

- W Liu, X Deng, G. H. A. F : 2006, An antiphishing strategy based on visual similarity assessment, IEEE Internet Comput..
- Hossain Kordestani, Mehdi Shajari: An Entice Resistant Automatic Phishing Detection, 2013 5th Conference on Information and Knowledge Technology (IKT), Department of IT and CE Amirkabir University of Technology Tehran, Iran
- Mona Ghotaish Alkhozae, O. A. B : 2015, Phishing websites detection based on phishing characteristics in the webpage source code, Department of Computer Sciences,FCIT King Abdulaziz University, Jeddah, KSA .
- Afroz, S. and Greenstadt, R : 2015, Detecting phishing websites by looking at them, IEEE Communications Society .

REFERENCES contd.

- Huh, J. and Kim, H : 2012, Phishing detection with popular search engines : Simple and eective, FPS11 Proceedings of the 4th Canada-France MITACS conference on Foundations and Practice of Security 8, 194 207.
- Jain, A. and Guptal, B: 2016, Eurasip journal on information security, EURASIP Journal pp. 426436.
- The Web information company : www.alexa.com .
- Phishtank: www.phishtank.com .

Thank you!

विश्वजीवनामृतं ज्ञानम्