Detecting Phishing Attacks Using Analytics and Machine Learning

1. E.Anbalagan 2. Charan.V 3.Abisath.AB 4. Thulasiram.B 5.Devasastha.P

1.Associate professor of dept of Information Technology at Jeppiaar Institute of Technology

2,3,4,5 Students of Dept of Information Technology at Jeppiaar Institute of Technology

Abstract

Malicious websites have accelerated the growth of Internet crime and restricted the creation of Web services. Therefore, there is an incentive to come up with effective solutions to block users from accessing these websites. We propose a study based on categorizing websites into 3 categories: harmless, spam, and malicious. Our strategy focuses only on the Uniform Resource Locator (URL) itself, without accessing the site's content. This eliminates latency and potentially exposes users to browser-based vulnerabilities. By using a learning algorithm, our plan outperforms blacklist services in terms of scope and coverage.

Keywords: phishing, vulnerability, Pretending, dataset, parameters, URL, analysis.

1.Introduction

The Internet has made it easier than ever for many to manage their finances and investments, while also creating more opportunities for fraud at lower cost. Less and less fraud Scams control more users than hardware/software affecting the business. Phishing is one of the most common Internet scams. It is designed to steal personal information such as passwords and credit card information. There are two types of phishing attacks: tries to trick victims into revealing their passwords by pretending to be a trusted organization that wants real information. Attempts to collect passwords by placing malware on the victim's machine. The specific viruses used in phishing attacks are discussed by the virus and malware community and beyond this article. Phishing attacks that deceive users are the focus of this article, and the word "phishing" is the focus of this article. Security mechanisms that detect phishing URLs and prevent them from reaching users will be developed and awareness will be raised. focus of this article, and the term "phishing" is the focus of this article. To be used to send attack

1.1 Purpose

Main purpose of this document This is to identify the system's Genuine, Malware and Malware URLs. to learn .

1.2 Motivation

The reason behind this system is to remind users to visit the problematic site. will raise awareness and develop security mechanisms that can detect phishing URLs and prevent them from reaching users.

1.3 Summary:

malicious websites contributed to the growth of cybercrime and limited the development of Web services. Therefore, there is an incentive to find effective solutions to block users from accessing these websites. We present a study based on classifying websites into 3 categories: harmless, spam and malicious. Our Policy focuses only on the Uniform Resource Locator (URL) itself and does not access the content of the site. This eliminates lags and potentially exposes users to browser vulnerabilities. Using a learning algorithm, our strategy outperforms blacklist services in terms of detail and scope.

URLs for websites fall into 3 categories:

Benign: Safe websites with normal service

Spam: Active websites that try to flood users with advertisements, or websites such as fake research and live chat

Malware: Do not use the functionality of the website by attackers Write responsive, built to break. access to information or personal computer systems.

2. EXISTING SYSTEM

A poorly structured NN model may cause the model to underfit the training dataset. On the other hand, exaggeration in restructuring the system to suit every single item in the training dataset may cause the system to be over fitted.

One possible solution to avoid the Over fitting problem is by restructuring the NN model in terms of tuning some parameters, adding new neurons to the hidden layer or sometimes adding a new layer to the network.

For instance, the model designer may set the acceptable error rate to a value that is unreachable which causes the model to stick in local minima or sometimes the model designer may set the acceptable error rate to a value that can further be improved.

2.1 Disadvantages of Existing system

-> It will take time to load all the dataset.

-> Process is not accurate.

-> It will Analyse slowly

3. PROPOSED SYSTEM

Analyzing lexical features enables us to capture the property for classification purposes. We first distinguish the two parts of a URL: the host name and the path, from which we extract bag-of-words (strings delimited by ‘/’, ‘?’, ‘.’, ‘=’, ‘-’ and ‘’).

We find that phishing website prefers to have longer URL

Here, we performed the experiments using two different classifiers: Random Forest and Support vector machine

3.1 Advantages of proposed system

-> All of URLs in the dataset are labelled.

-> We used two supervised learning algorithms random forest and support vector machine to train using scikit-learn library

4 Literature Survey

Title: Large-Scale Automatic Distribution of Phishing Pages.

Author: Colin Whittaker, Brian Ryner, Marria Nazif.

Summary:

Phishing sites, scam sites that trick trusted third parties to access personal information, continue to cost Internet users more than $1 billion each year. In this article, we describe the design and performance of a machine learning algorithm we developed to detect phishing websites.

We use this process to maintain Google's phishing blacklist. Our distributors analyze millions of pages every day by analyzing the URL and content of the page to determine if the page is phishing. Unlike previous work in this area, we train employees on noisy datasets with millions of samples from distributed data in real time. Despite the noise in the training data, our agent learned a good pattern for detecting phishing pages, which identifies over 90% of phishing pages after a few weeks of training.

SYSTEM REQUIREMENTS

Hardware:

1. Windows 7,8,10 64 bit
2. RAM 4GB

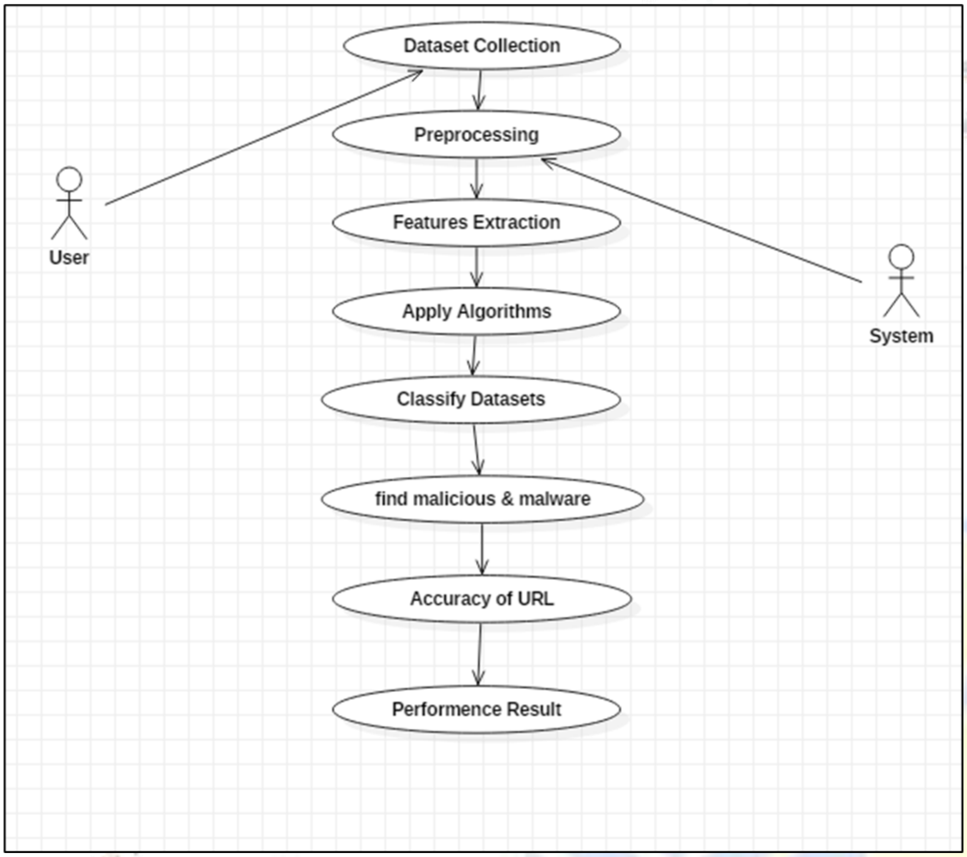
Software:

1. Data Set
2. Python
3. Python built-in modules
4. Anaconda navigator

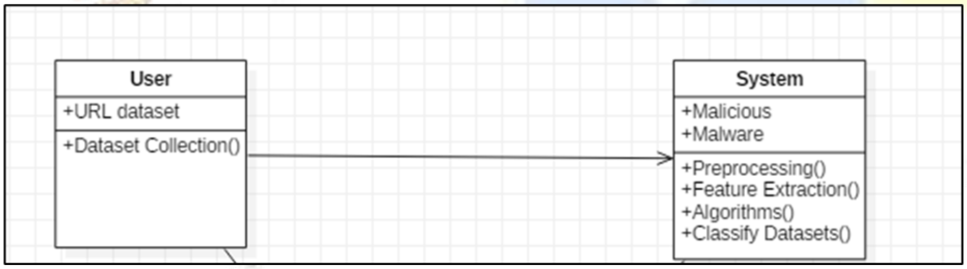
5. NumPy
6. Pandas
7. Matplotlib

5 UML DIAGRAMS

USE CASE DIAGRAM



5.1 CLASS DIAGRAM

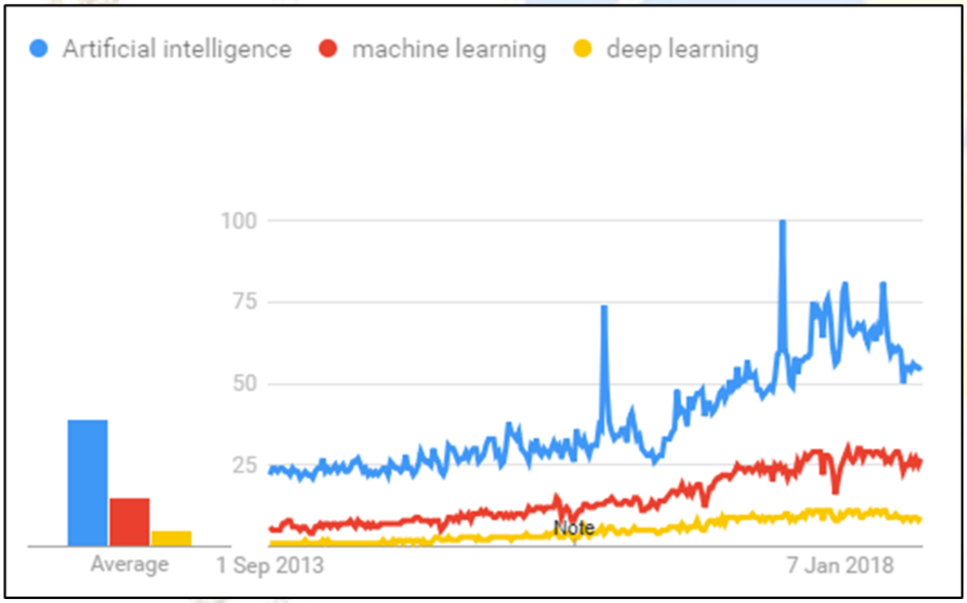


When to use ML or DL?

In the table below, we show the difference between machine learning and deep learning.

	Machine learning	Deep learning
Training dataset	Small	Large
Choose features	Yes	No
Number of algorithms	Many	Few
Training time	Short	Long

Using machine learning to teach algorithms requires less data than deep learning. Deep learning requires a lot of different data to identify underlying patterns. Also, machine learning is faster education model. Training deep learning programs can usually take from a few days to a week. The advantage of deep learning over machine learning is that it is very accurate. You do not need to know which features best represent the data; The neural network learns to select important features. In machine learning, you choose which features to include in your model.



CONCLUSION

In this we can describe our large-scale for automatically classifying phishing pages which maintains a false positive rate below0.1%. By automatically updating our backlist with our classifier, we minimize the amount of time that Phishing pages can remain active before we protect our users from them. Even with perfect classifier and rebuts system the backlist approach keeps us behind the phishers. We can only identify a phishing URL and normal URL using machine learning algorithm.

6.REFERENCES

[1] G. Aaron and R. Rasmussen, “Global phishing survey: Trends and domain name use in 2016,” 2016.

[2] B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, “Fighting against phishing attacks: state of the art and future challenges,” Neural Computing and Applications, vol. 28, no. 12, pp. 3629–3654, 2017.

[3] A. Aleroud and L. Zhou, “Phishing environments, techniques, survey,”countermeasures: Aand Security Computers & , vol. 68, pp. 160 – 196, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404817300810 G. Aaron and R. Rasmussen, “Phishing activity trends report: 4th

[4] quarter 2016,” 2014. R. Verma, N. Shashidhar, and N. Hossain, “Detecting phishing

[5] emails the natural language way,” in Computer Security–ESORICS 2012. Springer, 2012, pp. 824–841. M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature

[6] survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013. G. Park and J. M. Taylor, “Using syntactic features for phishing

[7] detection,” arXiv preprint arXiv:1506.00037, 2015. R. Dazeley, J. L. Yearwood, B. H. Kang, and A. V. Kelarev