**EVENT-DRIVEN MALICIOUS URL EXTRACTOR**

S.W. Sajeth Jonathan

(IT18071412)

B.Sc. (Hons) Degree in Information Technology Specializing in Cyber Security

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

October 2021

**EVENT-DRIVEN MALICIOUS URL EXTRACTOR**

S.W. Sajeth Jonathan

(IT18071412)

Dissertation submitted in partial fulfilment of the requirements for the B.Sc. (Hons) Degree in Information Technology Specializing in Cyber Security

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

October 2021

# Declaration

We declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

|  |  |  |  |
| --- | --- | --- | --- |
| Student Number | Student Name | Student Email | Signature |
| IT18071412 | S. W. S. Jonathan | It18071412@my.sliit.lk | Shape  Description automatically generated with medium confidence |

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor Date

...................................

# Abstract

The Internet has seen a tremendous boom in the past decade. Along with it, attackers have come up with various ways to infiltrate networks. Successful attacks do not necessarily require 0-days or complex attack vector, but a mere manipulation of the human factor. Social Engineering is an effective way to trick users to perform a malicious actor’s bid. A prominent social engineering method is using malicious URLs to conduct phishing, drive by downloads and spam. With the increased variation of cyberattacks, URL classifiers models must also adapt to the requirements. The prevention method evolves by not only using blocklists, but also by leveraging Machine Learning/ Deep Learning Capabilities. The current research space has provided various outputs of having classifiers in both the Deep Learning and Machine Learning Domain to increase accuracy. Our research identifies the current gaps in detecting these malicious URLs by proposing an ensemble model and using feature reduction to build a accurate and robust model. In addition, we propose an event-driven system to classify malicious URLs and provide lists to researchers/organizations that can be used in their existing blacklists. This system encompasses a modular approach since various models will be used. Working with modules, we can also work on improving the scalability of the system. At its core, the system would depend on the event provided by the user to classify URLs accordingly. Scalability should be integrated to each of the modules depending on the event.

**Keywords – Ensemble Models, Malicious URLs, Machine Learning, Deep Learning, Event-driven**

Table of Contents

[Declaration iii](#_Toc85058389)

[Abstract iv](#_Toc85058390)

[List of Figures vi](#_Toc85058391)

[List of Tables vii](#_Toc85058392)

[List of Abbreviations viii](#_Toc85058393)

[1 Introduction 1](#_Toc85058394)

[1.1 Introduction 1](#_Toc85058395)

[1.2 Background & Literature Survey 2](#_Toc85058396)

[1.3 Research Gap 5](#_Toc85058397)

[1.4 Research Problem 7](#_Toc85058398)

[2 Methodology 8](#_Toc85058399)

[2.1 System Architecture 11](#_Toc85058400)

[2.2 Commercialization 13](#_Toc85058401)

[3 Budget and Budget Justification 15](#_Toc85058402)

[Reference List 16](#_Toc85058403)

# List of Figures

[Figure 1.1: Phishing Activity Trend [3] 3](file:///C:\Users\nightcrawler\Downloads\IT18071412-2021-085.docx#_Toc85058466)

[Figure 3.1: Ensemble Modeling 8](#_Toc85058467)

[Figure 3.2: Deep Learning for Phishing Attack Detection [10] 9](#_Toc85058468)

[Figure 3.3: Activity diagram for the Web Interface - Extractor 10](#_Toc85058469)

[Figure 3.4: Overall System Architecture 11](#_Toc85058470)

[Figure 3.5: Ensemble Model + Web Interface Architecture 12](#_Toc85058471)

[Figure 4.1: Homage - Extractor 14](#_Toc85058472)

[Figure 4.2: Results Page - Extractor 14](#_Toc85058473)

[Figure 4.3: Visuals Page – Extractor 15](#_Toc85058474)

# List of Tables

[Table 1.1: Comparisons with previous works 6](#_Toc85058460)

[Table 3.1: Required Technologies, Techniques and Algorithms 8](#_Toc85058461)

[Table 3.2: Pricing Models 13](#_Toc85058462)

[Table 6.1: Expenses for the proposed system 15](#_Toc85058463)

# List of Abbreviations

|  |  |
| --- | --- |
| Abbreviation | Description |
| CNN | Convolutional Neural Network |
| DBN | Deep Belief Network |
| DL | Deep Learning |
| kNN | k-nearest Neighbor |
| LSTM | Long-Short Term Memory |
| ML | Machine Learning |
| SVM | Support Vector Machines |
| URL | Uniform Resource Locator |

# Introduction

## **Introduction**

Internet adoption has surged over the past few years and the COVID-19 pandemic boosted a widespread adoption of digital-first lifestyle. Internet usage has increased between 50% to 70% [1] and over 4.66 billion active users have been recorded for January 2021 [2]. An average user utilizes 6h 55m of their time on the internet [3] browsing through social networks, digital media, educational platforms, and gaming.

Certain online applications have been rapidly adopted with the drastic shift caused by the pandemic such as a 31% increase in restaurant delivery services and 21% increase of health products purchases [4]. Along with this, the rate of cyberattacks have also increased with phishing being a major concern. 57% of respondents said their organization experienced a successful phishing attack in 2020 [5]. Digital transformation has been rapidly enforced to various industry sectors. The pace of this transformation made people overlook the security risks come with it. The need of awareness of cybersecurity hygiene is an important part to prevent cybersecurity attacks. Social engineering is a prominent way to lure victim in cybersecurity attacks. A report [6] found that 98% of cyberattacks rely on social engineering.

Chart, line chart

Description automatically generated

An event-driven malicious URL extractor is a software implementation to detect phishing at the earliest level possible by analyzing trending keywords in a global/national level by utilizing twitter trending data and comparing them with recent whois domain registrations to find malicious domains. This system will be using an ensemble model at its core to provide an efficient classification system. In addition, a chrome extension can be deployed to endpoints to take advantage of this ensemble model and protect users from visiting malicious websites.

## **Background & Literature Survey**

Badadhe et al. [5] suggest that a phishing attack consists of 4 phases. First, the attacker prepares a phishing webpage. Then, delivers it to the masses generally using emails or other messaging platforms. Thirdly, the victim is tempted to click the link and visit the malicious crafted page. Finally, the victim will be manipulated in disclosing their credentials or sensitive information. Malicious URLs are not only limited to phishing attacks as stated by Sadique et al. [6]. Other malicious URLs include drive by downloads URL and spam URLs.

Chart, line chart

Description automatically generatedThe Anti-Phishing Working Group’s (AWPG) report [7] for the last quarter of 2020 shows that the rise of phishing activity has been drastic [Figure 1].

Figure .: Phishing Activity Trend [3]

Authors in Benavides et al. [8] conducted a literature review on current Deep Learning (DL) solutions and concluded that most regularly utilized are Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs). Yi et al. [9] present a detection model based on Deep Belief Networks (DBNs) and achieve approximately 90% true positive rate and 0.6% false positive rate.

CNN networks deal better in URL classification more than a Long-Short Term Model (LSTM), also consuming lower amount of time for training [10]. The training of the CNN network lasted for three minutes while two milliseconds was used for the URL Checking. Authors in Yazhmozhi et al. [11] proposed an ensemble model using two DL networks, Recurrent Neural Networks (RNNs) and CNNs providing a precision of 97%. It is note-worthy to point out their model achieved this precision without using a feature extraction process thus saving time.

Unlike the Deep Leaning domain, the Machine Learning space has seen a lot of research done to detect malicious URLs. Authors in Isphany et al. [12] experimented with different ML classifiers such as Support Vector Machines (SVMs), kNNs, Naïve Bayes, Logistic regression and AdaBoostM1. They found that kNNs produced the highest accuracy (99.2%) and the lowest false positive rate (2%). Following that SVMs proved to have the second-highest results.

In another comparison done by A. Makkar et al. [13] , ensemble models such as Monotone Multi-Layer Perceptron Neural Network, Multi-Layer Perceptron, Neural Networks with Feature Extraction worked best in terms of accuracy. The results were improved by introducing Cross-Validation to the ensemble models

## **Research Gap**

A lot of research has been done to detect malicious URLs using Deep-learning and Machine Learning models to increase the accuracy [8, 9, 12]. However, the scope for incorporating ensemble models is yet to be explored thoroughly [14]. Ensemble models are techniques to create multiple simpler models and combine them to produce better and accurate results.

The need for an accurate model is justified with the advancements made in the artificial intelligence domain. The domain adopts new approaches every year and previous URL classifiers becomes less efficient as time goes on. The latest advancements can be utilized to build a modern and efficient model. On the other hand, reducing the false positive rate is a crucial feature of a robust model. False positive results can cause overhead in an organization since it would require extra steps of processes to whitelist the false positive URL. In a previous work for an Intrusion Detection System, ensemble model proves to have the highest classification accuracy and the lowest false positive rates [15]. Other works in the URL detection space, also have utilized some form of ensemble modeling aiming for the above-mentioned features [16] [11] [13].

A comprehensive literature survey strengthened our gap for the need of the ensemble model [14]. The same authors proposed an ensemble model using three machine-learning classifiers: Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Decision Tree (C4.5) to use in an ensemble method with Random Forest Classifier (RFC) [16].

This work could be extended by comparing with other models such as Bidirectional LSTM and CNNs, as mentioned by Yazhmozhi et al [11]. The models does not have to be limited to certain models, but we can also use various other combination of models and compare them with metrics to settle on a robust and suitable model for this event-driven system.

While many researchers have proposed novel ideas for malicious URL detection, even going for the extent of creating automated frameworks [6]. We had yet to come across a system designed to correlate malicious URLs with events. This system can be a practical approach for organizations and researchers to create malicious URLs lists with regards to any event. For example, the COVID-19 pandemic caused unprecedented disorder resulting in a drastic increase in phishing URLs [7]. They ranged from sites posing as World Health Organization to Video conferencing domains [17] . Attackers took advantage of the sudden remote working/learning culture. This also included tricking users who were looking for employment.

Such an event-driven system to output Malicious URL List can be used by organizations/researchers to build blocklist to stop a phishing attack in the early stage of its lifecycle.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Solution | Considers accuracy of the model | Considers  Efficiency of the model | Categories URLs based on events | Building a scalable system |
| Research A | ´ |  | ´ | ´ |
| Research B |  | ´ | ´ | ´ |
| Research C |  | ´ | ´ | ´ |
| Proposed system |  |  |  |  |

Table .: Comparisons with previous works

## **Research Problem**

A single event, COVID-19 Pandemic created a variety of pathways for attacker [12]. This gives some insight where a single event, can give arise to various related events in which attackers can leverage to craft malicious domains. We have to keep in mind that cyberattacks are also coming up with novel ideas to social engineer unwitting users.

Like this, there are lot of events happening in a global or local scale at any given time [18] . Even though, they might not have a higher entropy as the COVID-19 pandemic, they still do contribute to a portion of malicious URLs. Collecting, labelling, and categorizing these URLs by co-relating to an event can provide new insight for researchers and targeted protection for organizations. Our proposed system uses an event-based approach to label malicious URLs and categorize them. The system consists of a web-application for the user input events and outputs the labelled malicious URLs. The labelling is done by correlating the URL with keywords. The user interface has to be simple, clean and intuitive to make it welcoming for the user.

Combining the relevant research done in the DL and ML space to build a robust ensemble model can increase the accuracy of the system while having a better performance. In addition, ensemble models tend to showcase high accuracy along with a lower false positive rate. This model can fit into the event-driven system as a module. Whenever, a better ensemble model is introduced, this system can adapt to it. We must determine if the suitable ensemble model’s accuracy is adequate without trading too much in performance.

# Research Methodology

## **Methodology**

As stated in the research gap, the goal of this research is to introduce a solution to list malicious URLs to the user by correlating with the keywords provided. To achieve this an ensemble model is built.

|  |  |
| --- | --- |
| Technologies | Azure Machine Learning Studio  SciKit Learn  TensorFlow  Keras  Flask |
| Techniques | Ensemble Modeling using ML and DL |
| Algorithms | LSTM CNN |

Table .: Required Technologies, Techniques and Algorithms

OpenPhish’s feed provides malicious URLs and tags them with the brand they targeted. But other feeds, only provide the malicious URL. Therefore, we have to label the collected URL dataset with the targeted keyword. This dataset will be used to train the keyword matcher module built by another team member

An ensemble model is formed by combining various simpler models (Figure 2). As stated in the research problem, this model will showcase a higher accuracy and false positive rate that solely using ML and DL models individually.

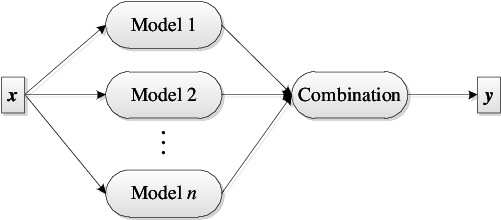


Figure .: Ensemble Modeling

Different ML and DL models will be examined considering their effective combination to prepare ensemble models. These ensemble models will be then compared with each other to determine the suitable model for this system. The conditions for benchmarking would be as follows.

1. The rate of False Negative (FN)
2. The rate of False Positive (FP)
3. The rate of True Positive (TP)
4. The rate of True Negative (TN)
5. The time taken for the model to classify the URLs

These five conditions can be used to determine metrics such as.

1. Precision
2. Recall
3. F-Measure
4. Accuracy

By gathering these metrics, we can compare the robustness of various ensemble models.

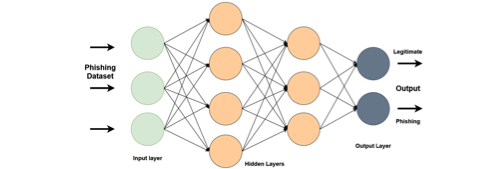


Figure .: Deep Learning for Phishing Attack Detection [10]

Bidirectional LSTM and CNNs will be given priority to build the ensemble model since it was suggested as a potential model by a previous literature [11]. LSTMs and CNNs are part of the Deep Learning Domain (Figure 3).

Figure 4.1 depicts the activity diagram for a user using the web interface to collect Malicious URL List.

Diagram

Description automatically generated

Figure .: Activity diagram for the Web Interface - Extractor

## **System Architecture**

Graphical user interface, application

Description automatically generated

Figure .: Overall System Architecture

Graphical user interface, application

Description automatically generated

Figure .: Ensemble Model + Web Interface Architecture

## **User Interfaces**

Graphical user interface

Description automatically generated

Figure 4.1: Homage - Extractor

Graphical user interface, table

Description automatically generated

Figure 4.2: Results Page - Extractor

Graphical user interface, application

Description automatically generated

Figure 4.3: Visuals Page – Extractor

## **Commercialization**

Since a part of the userbase for this system would be small-medium enterprises (SME), this system can be commercialized with the note of providing basic security. Even though, at this stage the system cannot compete with state-of-the-art endpoint management system, this will be suitable for SMEs due to the lower cost and higher usability.

Two versions of this system can be implemented.

* A free version that SMEs and Researchers can use to collect URL lists based on events with limited export capability.
* A paid version that will provide seamless export capability in addition to the basic endpoint protector using the browser plugin.

|  |  |
| --- | --- |
| Free Version | Rate Limit on event-based malicious URL list and restricted export capabilities. |
| Paid Version | * Browser plugin to protect users from malicious pages with basic reporting to the administrator * No limit on export and event-based malicious URLs |

Table .: Pricing Models

The low-cost barrier should be a key aspect to attract SMEs. Researchers/ Investigators investigating certain events can make use of the data to correlate with the incident.

## **Budget and Budget Justification**

|  |  |  |
| --- | --- | --- |
| Task | Cost($) | Cost(Rs.) |
| Azure Machine Learning Studio (Implementation and Cloud Deployment) | 100 (Monthly Fee) | **19,900** |
| Web Application Hosting | 72 | **14300** |
| Database – Mongo DB | 57 | **11400** |
| Plugin upload to Chrome Web store | 5 (One-time fee) | **995** |
|  | Total Cost | **46,595** |

Table .: Expenses for the proposed system

To cover these costs, we can use the pricing models mentioned in the commercialization section.

## **Testing and Implementation**

# Results and Discussion

## Results

## Research Findings

## Discussion

# Conclusions

# Reference List

|  |  |
| --- | --- |
| [1] | M. Beech, "Forbes," 25 05 2020. [Online]. Available: https://www.forbes.com/sites/markbeech/2020/03/25/covid-19-pushes-up-internet-use-70-streaming-more-than-12-first-figures-reveal. |
| [2] | J. Johnson, "Statista," 10 09 2021. [Online]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide. |
| [3] | "Data Reportal," 04 08 2021. [Online]. Available: https://datareportal.com/global-digital-overview. |
| [4] | "Statista," 31 05 2021. [Online]. Available: https://www.statista.com/statistics/1107859/shifting-to-online-purchases-because-of-the-covid-19-pandemic-by-category/. |
| [5] | N.Badadhe, S.More and N. Puri, "An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms With Results.," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET,* vol. 3, no. 5, p. 1, 2014. |
| [6] | F. Sadique, R. Kaul, S. Badsha and S. Sengupta, "An Automated Framework for Real-time Phishing URL Detection," in *10th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2020. |
| [7] | "APWG Trend Report," 2020. [Online]. Available: https://docs.apwg.org/reports/apwg\_trends\_report\_q4\_2020.pdf. [Accessed 25 February 2020]. |
| [8] | E. Benavides, W. Fueters, S. Sanchez-Gordon and M. Sanchz, "Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review," Research Gate, 2020. |
| [9] | P. Yi, Y.Guan, F. Zou, Y. Yao, W. Wang and T. Zhu, "Web Phishing Detection Using a Deep Learning Framework," *Wireless Communications and Mobile Computing,* vol. 2018, 2018. |
| [10] | W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Schererd and M. Woźniake, "Accurate and fast URL phishing detector: A convolutional neural network approach," *Computer Networks,* vol. 178, 2020. |
| [11] | V. Yazhmozhi, B. Janet and S. Reddy, "Anti-phishing System using LSTM and CNN," in *IEEE International Conference for Innovation in Technology (INOCON)*, Bengaluru, 2020. |
| [12] | J. Ispahany and R. Islam, "Detecting Malicious Urls of COVID-19 Pandemic Using ML Techniques," 2020. |
| [13] | A. Makkar, N. Kumar, L. Sama, S. Mishra and Y. Samdani, "An Intelligent Phishing Detection Scheme Using Machine Learning," in *Sixth International Conference on Mathematics and Computing, Advances in Intelligent Systems and Computing*, Singapore, 2021. |
| [14] | A.Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Springer,* 2020. |
| [15] | D. Gaikwad and R. C.Thool, "Intrusion Detection System Using Bagging Ensemble Method of Machine Learning," in *International Conference on Computing Communication Control and Automation*, 2015. |
| [16] | A. Basit, M. Zafar, A. Javed and Z. Jalil, "A Novel Ensemble Machine Learning Method to Detect Phishing Attack," in *IEEE 23rd International Multitopic Conference (INMIC)*, 2020. |
| [17] | T. Kelly, "Security Magazine," [Online]. Available: https://www.securitymagazine.com/articles/92666-how-hackers-are-using-covid-19-to-find-new-phishing-victims. [Accessed 25 02 2021]. |
| [18] | L. O'Reilly, "Security Boulevard," 30 10 2020. [Online]. Available: https://securityboulevard.com/2020/10/four-days-to-the-election-dont-fall-for-these-phishing-scams/. [Accessed 25 02 2021]. |
| [19] | A. Petrakova, M. Affenzeller and G. Merkurjeva, "Heterogeneous versus Homogeneous Machine Learning Ensembles," *Information Technology and Management Science,* 2015. |
| [20] | M. Beech, "Forbes," 25 05 2020. [Online]. Available: https://www.forbes.com/sites/markbeech/2020/03/25/covid-19-pushes-up-internet-use-70-streaming-more-than-12-first-figures-reveal. |