**EVENT-DRIVEN MALICIOUS URL EXTRACTOR**

2021-085

Project Proposal Report

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

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

February 2021

**EVENT-DRIVEN MALICIOUS URL EXTRACTOR**

2021-085

Project Proposal Report

Ramanayaka A. M

IT18021912

Amila Senarathne

Kanishka Yapa

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

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

February 2021

# **Declaration**

I declare that this is my 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 my 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 |
| IT18021912 | Ramanayaka A. M | it18021912@my.sliit.lk |  |

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

Signature of the supervisor Date

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

# **Abstract**

People have become more used to using the internet for various purposes at the present and this usage will increase in the future with the rapid development of technology. With this development in technology there will always be advantages as well as disadvantages. Advantages are that the communication and work becomes much easier and faster for people then more and more people start to rely on these services, due to that comes the internet fraud and internet based security concerns arise which becomes a disadvantage. With that being said, the need for web security increases and more and more people has to be aware as to not get caught for any malicious activities which makes them vulnerable to security issues. Malicious URLs is one of the main concerns that has caught the attention of people because it can approach people in any form such as a phishing email that redirects them to a malicious website or a voice call that seems to be from a legitimate person which will ask for private and sensitive information such as credit card numbers or passwords or drive by downloads which installs malicious software on the devices. Phishing/spam emails and phishing websites can be identified in various ways and they can be categorised into various formats. There are so many scenarios of which attackers use malicious URLs to target different people for different purposes but the end goal of the attackers will always be the same which is the personal and sensitive information of their victims. Categorization of phishing, spam and other malicious URL scenarios, URL feature representation and feature analysis are the main focus of my study and I will be studying more on how this can help with the speed and efficiency of the system for detecting phishing URLs. Lexical analysis and obfuscation techniques will be some of the areas that will be looked into in this study and more techniques will be studied with the existing techniques and I will also look into hybrid approaches to increase the speed and efficiency of the system.

**Table of Contents**

[**Declaration** 3](#_Toc67515771)

[**Abstract** 4](#_Toc67515772)

[**Table of Contents** 5](#_Toc67515773)

[**1 Introduction** 6](#_Toc67515774)

[1.1 Background & Literature Survey 6](#_Toc67515775)

[1.2 Research Gap 8](#_Toc67515776)

[1.3 Research Problem 10](#_Toc67515777)

[**2 Objectives** 11](#_Toc67515778)

[2.1 Main Objectives 11](#_Toc67515779)

[2.2 Specific Objectives 12](#_Toc67515780)

[**3 Methodology** 13](#_Toc67515781)

[**Feature Reduction Techniques** 14](#_Toc67515782)

[ **Missing Value Ratio** 14](#_Toc67515783)

[ **Low Variance Filter** 14](#_Toc67515784)

[ **Random Forest** 14](#_Toc67515785)

[ **Backward Feature Elimination** 14](#_Toc67515786)

[ **Forward Feature Selection** 14](#_Toc67515787)

[**4 Budget and Budget Justification** 15](#_Toc67515788)

[**5 Commercialization** 16](#_Toc67515789)

[**References** 17](#_Toc67515790)

# **1 Introduction**

## 1.1 Background & Literature Survey

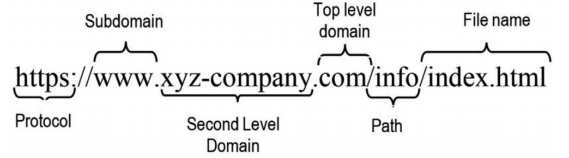
Internet is a boundless network that connects all the computers and smart devices in the world and people have become more used to it that they have arrived at a point where they cannot do any work without using the internet. Web services which are services provided through the internet are used in a wide variety of industries for many purposes. [1] People have started to rely on these web services more than ever. Online communication to online money transfer and many more activities have now been established which makes it more useful for the people while also encouraging more and more malicious web activities to occur therefore more and more security concerns arise with the increasing usage of web services among people. Attacker use this as an advantage to target their victims. Malicious URLs, that are used for many attacks such as phishing, drive by downloads or spam, has become one of the main security concerns because it manipulates people in many ways and people get caught for them due to various reasons. In this method attackers work using psychological tricks to get their targets to do what they want and unfortunately lot of people get caught for these scams and phishing emails. Besides normal people, even the security experts get caught for these scams therefore these phishing sites, emails and URLs should be filtered by any means so we have chosen this area of study to build a system to detect malicious URLs.

Phishing incidents increase every year and as statistics show, the number of phishing incidents were known to have increased by 15% from 2019 to 2020. [2] While phishing being a popular form of cyber-attack, phishing techniques and the skills have improved over the years so it is becoming harder to identify them.

Cyber criminals uses trending topics for their phishing emails as well such as in 2020 there were many phishing and scam emails related to and themed by the title of the COVID-19 pandemic. These kind of trending scenarios are widely used by attackers to attract their targeted victims more easily. The scenarios need to be categorized with much more than trending topics and need to be looked into more deeply to carry out the categorization much precisely and efficiently to speed up the system because scenario examples like COVID-19, fortune telling or war will not help with capturing malicious URLs. Finding what is common among even these examples will help to categorize the scenarios properly.

The next area to be looked into is the phishing URLs which will again be studied on how to be detected using existing techniques and using URL feature based detection where URLs will be looked for special features that normal URLs does not contain such as special characters or having too many dots compared to a normal URL, etc. [1] Attackers create fraudulent websites in which the interface and the overall layout is similar to popular websites to make them look authentic enough for attracting their targeted victims. The problem is that the attackers cannot get the exact same URL of the original website so they use different URLs. If a person is keen enough to see the difference in the URLs of the original website and the malicious website/webpage then they will most probably not click on these links but in real life people do not spend time to look into these details and they are not careful enough so if a phishing email arrives with a message that gives a sense of emergency such as to login to a website to confirm the user’s account to stop their account from being deleted then they will definitely click the link to perform that process and the user will easily give out their personal information such as a username and password or credit card numbers to these attackers. Therefore phishing attacks have become more common and more people get caught in these attacks. [3]

Knowing URL components is one of the key factors of capturing malicious URLs. The components of the URL is shown in the figure below. [3]



**Figure 1 - URL Components**

Second Level Domain (SLD) and Top Level Domain (TLD) is together known as the domain name which is a unique and critical feature of a URL. The attacker is able to either find or buy the SLD for phishing but even without buying it the attacker can still use it to generate vast amount of URLs by adding path and file names to the SLD which will still make it look authentic for a normal person. [3]

According to many researches the most common way to detect malicious URLs is known as the blacklist method where it contains a database of URLs that were known to be malicious in the past and were added to the list after confirmation and this database is updated with new URLs whenever found. This method is not very efficient in the present because new URLs are generated in large amounts and it is difficult to keep track of them. Next method is a heuristic approach which is a similar to blacklisting but this time it is not blacklisting URLs, instead it will blacklist signatures where they identify common attack types and categorize them and find the signature features, then it will search for these specific signatures throughout a webpage for any suspicious behaviour to produce a warning flag that shows that this web page is suspicious. Both these methods are resource intensive and is hard to implement therefore another approach that is said to be more efficient than both these methods is the machine learning approach. In this approach the information of URLs will be analysed and compared with the original webpage URLs. The good features of the original webpage URLs will be extracted for training the system so that it will be easier to detect suspicious behaviours and features of phishing URLs when detected and being compared. [4]

The next area that I will be focusing on is the scenarios that the phishers/attackers use for phishing. It is true that they will be using scenarios like COVID-19 or any other health related scenario or else they will use a topic like giving out shopping deals. The topics vary time to time but for the categorization of scenarios it will not be helpful so what I did was find out what was more common for every scenario and I listed out a set of common scenarios that attackers use while masking them using trending topics so for every topic this scenario will be the same. While looking into few websites I found out that the categorization would go by; look-alike websites, deactivation scares, SEO Trojans, work mules, fake crisis notices, compromised credit card, and social media request. [5] [6]

Feature Reduction will be the next area that I will focus to improve the speed and efficiency of the system. When there is a large amount of features to analyse it is very difficult to keep a system efficient and on proper speed. It will also reduce the reliability of the system and will increase the use of computational power which is not recommended. We have to find methods to find patterns in a dataset that has a large amount of data which shows a large amount of features so that we can reduce the features to make the system work more efficiently. There are two types where one is feature selection and the other is dimensionality reduction. Feature reduction is where we keep the most relevant features and dimensionality reduction is where we try to find a small set of features that will be a combination of the basic features we find. [10]

## 1.2 Research Gap

There are many research papers that explains the suitable techniques to detect malicious URLs but there is not much detailed description on the classification and categorization of scenarios and how it would be detected when capturing them on phishing emails, scam emails or any malicious links that are sent to people. Relying on methods such as blacklisting will be efficient but it will not be able to detect future attacks sooner and sometimes a phishing webpage may not behave suspiciously in the beginning and may launch the attack later so with that problem we cannot be sure that the blacklist method would detect that URL and flag it. Most researches explain about the keywords being used by different emails or URL features that would help to identify it as a malicious URL but there are not many research papers about scenario classification or categorization which would help with a machine learning approach or natural language processing analysis approach. Feature reduction techniques are not very much discussed in most of the research papers and needs to be looked up on. Feature extraction and feature classification is discussed to some extent but feature reduction should be looked into and improved in order to speed up the system and make it work more efficiently. Below table shows how we have looked into researches and identified the components covered in them and it shows that our proposed model will be able to cover all the components listed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Papers​** | **Scalability​** | **Accuracy ​** | **Classification of URLs​** | **High rate of detection/efficiency​** | **Ability to sort through keywords / Events​**  **​** | **Classification of Scenarios​** |
| Research A [1]​ | ​ | ​ | ​ | ​ | ​ | ​ |
| Research B [2]​ | ​ | ​ | ​ | ​ | ​ | ​ |
| Research C [3]​ | ​ | ​ | ​ | ​ | ​ | ​ |
| **Proposed Solution**​ | ​ | ​ | ​ | ​ | ​ | ​ |

## 1.3 Research Problem

Scenario classification and URL classification as mentioned before should be looked up on many research studies but I could find only a very few research papers that actually talked about the problem of the classification of URLs and the classification of the scenarios and in one of the research papers, a few methods were explained on how it is done and what methods are suitable.

Classification of emails had been carried out in one study which they show that they have identified 15 features which are often used by the phishers when phishing and they explain all the features in detail. They have also explained that they achieved high accuracy with this method of classification and these features are accurate as well. A few features was noticed which might be of use for the system that we are building because we will not be analysing emails but the malicious URLs only so below are some of those features that were listed by the researchers [7]:

* URLs Containing IP Address [7]
* Disparities between “href ” Attribute and LINK Text [7]
* Number of Dots in the Domain Name [7]
* Number of Links Linked To Domain [7]

The feature extraction can be done in various methods but the feature reduction should be looked into more to increase the efficiency of the system. There are linear approaches and non-linear approaches as well. It can be further divided into feature extraction and feature selection. Principle component analysis is known as one of the main linear approaches and there are many more approaches that can be looked into but with the ensemble model that we will be building, we will have to choose the exact matching approach to the system which will be based on the feature extraction carried out. [11]

# **2 Objectives**

## 2.1 Main Objectives

The main objective of implementing the malicious URL detection system is to detect and block phishing sites and malicious URLs and to filter phishing and scam emails because at the present day, attacks like phishing, scamming and drive by downloads has become very common and the techniques and tactic used by the attackers are evolving with the development in technology and the researches done so far has not found a proper solution that will completely get rid of this problem. Our main objective is to make sure that we can at least create a more efficient, speedier and a much accurate system than the systems created previously while using ideas taken from the previously created detection systems and the studies done so far. Identifying a suitable Ensemble Model using deep learning and machine learning algorithms for efficient, speedier and accurate analysis is the goal of our research and to build this system we will use the existing models and modify or go towards a hybrid approach for betterment of this system.

The main reason for us to choose this area of study for the research is mainly because of the phishing, scamming and malvertising attacks that took place during this current period where the attackers targeted their victims using the pandemic situation which is COVID-19 that spread throughout the world. Then they used COVID-19 themed templates for their phishing sites and COVID-1 related phishing and scam emails were generated to attract more victims so due to this situation and due to people being at rest and mostly at home, it became easier to get people to fall for these attacks. People had more time to check on this kind of emails, malvertising related to the COVID-19 situation, and to spend time on social media and click on links that their friends might send them which could most probably be redirecting these people to a phishing website which appears to be legitimate. [9]

## 2.2 Specific Objectives

The main objectives are divided among the members to work better towards the goal so individually a lot of areas will be looked into more closely so that we can create a better model using machine learning and deep learning algorithms. I will be looking into the following areas as I have been explaining earlier.

* URL Feature extraction will be one of the objectives that will help us find specific features related to malicious URLs which then can be used to train the model for better detection of malicious URLs.
* URL Feature Analysis will be another objective which will be carried out after the feature extraction is done.
* URL Feature Classification will then be done after the analysis to understand how many features that we were able to find commonly among the malicious URLs.
* URL Feature Reduction can then be done to improve the efficiency of the system which will reduce the features that were extracted and classified.
* Scenario Categorization will be another objective to fulfill into finding out the differences that it would cause for the features as well.

# **3 Methodology**

The malicious URL extractor, which is proposed by this study, is capable of URL feature extraction and classifying the URLs to understand which URLs are malicious URLs and there will be 24 highly discriminating feature values that will be extracted and used for training the model. [10]

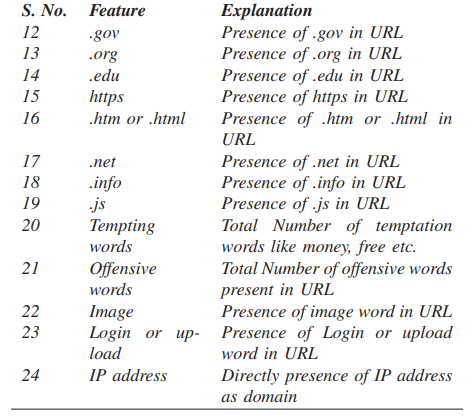
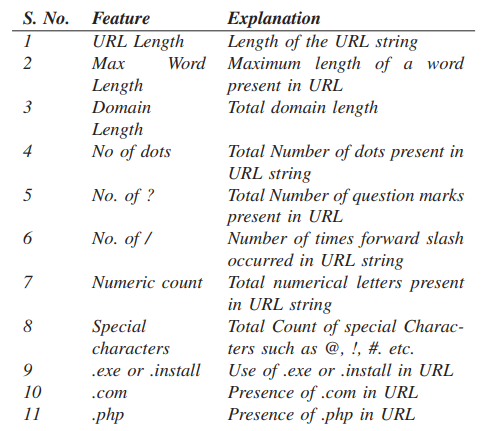


Figure 2 - URL Feature Extraction [10]

Feature extraction could be done in this way and by looking at the extracted features in the above figure we need to find a way to analyze and classify the URLs to carry out the feature reduction process. Feature reduction can be done in many ways. There are a few python based methods that are recommended to use when the reduction needs to be done for a large amount of features where a feature selection should be done. These methods are Missing Value Ratio, Low Variance Filter, Random Forest, Backward Feature Extraction and Forward Feature Selection. [10]

## **Feature Reduction Techniques**

* **Missing Value Ratio**

When there are missing values in a variable, if the threshold is more than 50% or a specific set percentage, then the variable should be dropped. This is how this approach will solve the problem. [10]

* **Low Variance Filter**

When there are variables which has similar values as another variable can be identified and dropped accordingly. [10]

* **Random Forest**

In random forest technique, we check for the most important features in the dataset. The importance of every feature is then checked to know which features to keep and use. The top most features will be kept. This is one of the most commonly used techniques. [10]

* **Backward Feature Elimination**

This is done by first taking all the variables and then the model is trained, then the performance of the model is checked. Then one at a time these variables are dropped and the performance is checked each time. The variables that cause the slightest or no change at all is dropped. [10]

* **Forward Feature Selection**

Each feature is checked with the model. If n number of features are there then the model is checked with the features separately n times. The feature that gives the best performance is used as the starting feature and then other features are added one at a time. This is done until there is no change in the performance. [10]

Both Backward Feature Elimination and Forward Feature Selection is time consuming and a lot of computational power is required to carry out these two techniques therefore it would be best to eliminate these techniques. We can look deeper into the other few techniques to see how compatible it would be with the ensemble model that will be built. [10]

# **4 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** |

# **5 Commercialization**

One part of the targeted user base for this system is small-medium enterprises (SME), which has higher usability and a lower cost. The system can be commercialized on the note of providing basic security.

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 |

# **References**

|  |  |
| --- | --- |
| [1] | Y. G. F. Z. Y. Y. W. W. a. T. Z. Ping Yi, “Web Phishing Detection Using a Deep Learning Framework,” *Hindawi,* vol. 2018, p. 9, 2018. |
| [2] | D. Warburton, “F5,” F5 Labs - Application Threat Intelligence, 11 November 2020. [Online]. Available: https://www.f5.com/labs/articles/threat-intelligence/2020-phishing-and-fraud-report#:~:text=The%20number%20of%20phishing%20incidents,at%20the%20access%20control%20layer.. |
| [3] | E. B. O. D. B. D. Ozgur Koray Sahingoz, “Machine learning based phishing detection from URLs,” *Elsevier,* p. 14, 2018. |
| [4] | C. L. S. C. H. H. Doyen Sahoo, “Malicious URL Detection using Machine Learning: A Survey,” 2019. |
| [5] | “19 Examples of Common Phishing Emails…And How to Avoid Them,” 2021 Terranova Worldwide Corporation, [Online]. Available: https://terranovasecurity.com/top-examples-of-phishing-emails/. |
| [6] | R. A. Grimes, “14 real-world phishing examples — and how to recognize them,” CSO Online, 9 April 2020. [Online]. Available: https://www.csoonline.com/article/3235520/15-real-world-phishing-examples-and-how-to-recognize-them.html#. |
| [7] | A. O. A. Andronicus A. Akinyelu, “Classification of Phishing Email Using Random Forest Machine,” *Hindawi,* vol. 2014, p. 7, 2014. |
|  |  |
| [8] | J. I. &. R. Islam, “Detecting Malicious Urls of COVID-19 Pandemic,” p. 17, 2020. |
| [9] | B. S. Yogendra Kumar, “A lightweight machine learning based security,” in *International Conference on Communication Systems & Networks (COMSNETS)*, 2021. |
| [10] | P. Sharma, “The Ultimate Guide to 12 Dimensionality Reduction Techniques (with Python codes),” Analytics Vidhya, 27 August 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/. |
| [11] | “Dimensionality reduction,” Wikipedia, 17 February 2021. [Online]. Available: https://en.wikipedia.org/wiki/Dimensionality\_reduction. |