**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

Wishvajith B.L.D.V.

IT18032666

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.

|  |  |  |
| --- | --- | --- |
| Name | Student ID | Signature |
| Wishvajith B.L.D.V. | IT18032666 |  |

The supervisor/s should certify the proposal report with the following declaration. The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor: Date

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

# Abstract

Cyber-attacks are types of attack often carried out in order to obtain sensitive information or cause disruptions to the services carried over internet. The recent events both globally and locally happening have increased these attacks exponentially spreading through the form of malicious URLs. With traditional methods like blacklisting malicious URLs, it is imperative to respond to such attacks in time efficiently. Most of the existing solutions lack in scalability and proactively protecting the users in a situation where the sprung up URLs connect with an event like recent Covid-19 related frauds. The proposed solution is presented with the main objective of tackling the issues of traditional systems and to provide an interface for users to protect by detecting phishing / malicious URLs in real time efficiently. We like to focus on letting the user input keywords and populate keywords with NLP algorithms and match them with the related URL token data to determine if the URLs are malicious or benign.

Key words - Malicious URL Detection, Machine Learning, NLP

# Table of Contents

[Declaration 1](#_Toc67513128)

[Abstract 2](#_Toc67513129)

[Table of Contents 3](#_Toc67513130)

[List of figures 4](#_Toc67513131)

[List of tables 5](#_Toc67513132)

[1 Introduction 6](#_Toc67513133)

[**1.1** Background & Literature survey 6](#_Toc67513134)

[**1.2** Research Gap 9](#_Toc67513135)

[**1.3** Research Problem 11](#_Toc67513136)

[2 Objectives 12](#_Toc67513137)

[**2.1** Main Objectives 12](#_Toc67513138)

[**2.2** Specific Objectives 13](#_Toc67513139)

[3 Methodology 15](#_Toc67513140)

[**3.1** System Architecture 15](#_Toc67513141)

[**3.2** Keyword Populator 16](#_Toc67513142)

[**3.3** Related URLs Finder 16](#_Toc67513143)

[**3.4** Comparison Engine 17](#_Toc67513144)

[3.4.1 Entropy Calculation 17](#_Toc67513145)

[3.4.2 Finding Obfuscated Keyword Variants 17](#_Toc67513146)

[4 Gantt chart 19](#_Toc67513147)

[5 Budget and Budget Justification 20](#_Toc67513148)

[6 Commercialization 21](#_Toc67513149)

[Reference List 22](#_Toc67513150)

[**Appendix A – Plagiarism Report** 23](#_Toc67513151)

# List of figures

Figure 1.1- Phishing incidents in 2020 compared to prior years 8

Figure 3.1- System Architecture 15

Figure 3.3- Related URLs Finder Architecture 16

Figure 4.1- Gantt Chart 19

# List of tables

[Table 1.1- Comparison of proposed solution with former researches](#_bookmark10) 10

[Table 5.1-Budget Chart](#_bookmark37) 20

Table 6.1- Commercialization Versions 21

# Introduction

## Background & Literature survey

The rise of new correspondence innovations has had a huge effect in the development and advancement of organizations crossing across numerous applications including internet banking, web based business, furthermore, digital communication. Therefore, the significance of the World Wide Web has been consistently expanding. Lamentably, these advancements come combined with new strategies for attackers to assault and trick clients. There are a wide assortment of procedures to execute such assaults, for example, drive-by download, social designing, phishing, and numerous others. As with the technological advancements, the restrictions of conventional security strategies are getting increasingly more serious given this dramatic development of new security threats, quick changes of new IT innovations, and the lack of security experts. The focused part of this research is focused on preventing these attacks. For example phishing attacks, which is a type of attack carried out in order to obtain sensitive information by tricking people into providing them usually by presenting themselves as a legitimate entity. The greater part of these attack methods are acknowledged through spreading traded off URLs (or the spreading of such URLs forms.) Depending on the type of the phishing attack performed, it can either be highly targeted to some organization or a person or it can be random. These highly targeted ones are called ‘Spear Phishing’ and often carried out as an initiative step for a massive cyber-attack.

Recent COVID-19 pandemic situation has also increased these cyber-attacks which have been leveraging as the world digitally transforms affecting government, corporate industries, educational organizations, non-profit organizations because they had to shift to online platforms from their traditional on premises systems. [1] As depicted by the *Figure 1.1,* according to the F5 Labs Phishing and Frauds Report for year 2020 shows a massive 220% increase in phishing activities within the year compared to last 5 years prior to 2020. [2]

These phishing attacks can then often lead to massive cyber-outages or attacks. Also, cybercriminals since beginning have had come up with different techniques to make the phishing websites look as authentic as possible. As with such incidents like Covid – 19, attackers always try to exploit this event related data as an opportunity to trick users into landing on these malicious phishing pages in order to capture sensitive data. In doing so, one popular way is to make the domain name have words similar to a particular situation and user would want to visit the website due to this particular situation. A recent example would be the COVID-19 phishing URLs which had keywords targeting covid-19, WHO, vaccines in order to trick people into getting them directed to the phishing pages. The report by F5 Labs also showcases the usage of terms “covid” and “corona” in certificates and URLs peaking at nearly 14,940 in the month March. [2]

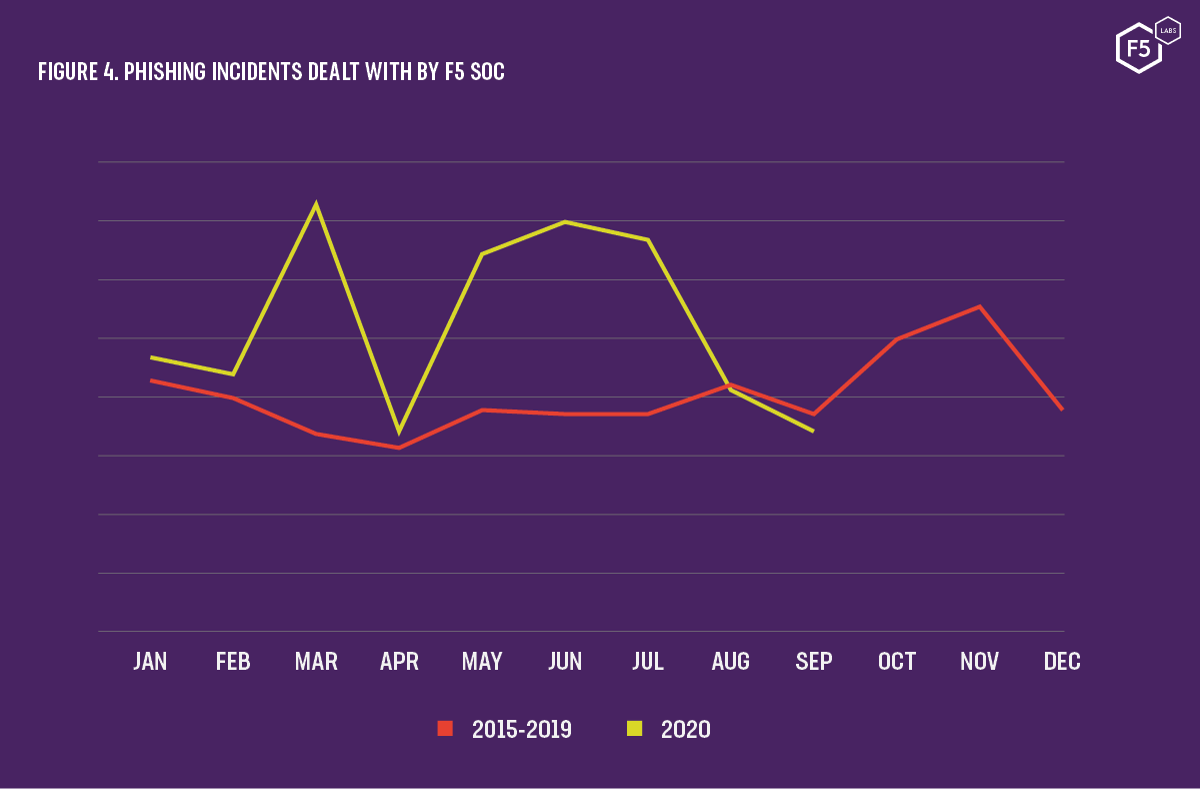
Many previous researches have been carried out on automation of detecting phishing URLs using both machine learning and non-machine learning approaches. Most of these approaches offer options for URL blacklisting, signature blacklisting. Traditional methods like blacklisting approaches are not feasible as maintaining such large databases of malicious URLs is not practical. In order to solve these problems, researches have looked into relying on machine learning algorithms to identify whether a URL is malicious or not.

One of these researches use the lexical analysis of the URLs to analyze the hostname length, URL length and tokens in URL. The authors speculate that often in phishing URLs in order to make it appear authentic insertion of additional characters is used. This can be used to detect randomly generated malicious URLs using alphabet entropy. After extracting features from URL list the data will go through a classifier and gets grouped into malicious and non-malicious by analyzing their lexical features. [3]

In another research, researchers have introduced a framework named, “Monarch” which is propose a real-time web crawling solution to determine if the URLs are malicious or benign. It is estimated that “Monarch” could process up to 15 million URLs per day. The system performs exactly the same as before by collecting features from the URLs and feeding the data into classifier to build a model for identification of malicious URLs. What is different is the increased efficiency in memory capabilities and algorithmic updates, which is a linear classifier based on logistic regression along with the L1-regularizer to induce relatively sparse models. [4]

Paper [5] propose a solution for increasing the scalability of the keyword matching with URL tokens when larger scale of data is at analysis phase. This solution, “online learning” is much more efficient than the traditional batch processing methods. Authors categorize the online learning algorithms into two main categories depending on their application at malicious URL detection; (i) First-order online algorithms, and (ii) Second-order online algorithms.

As the above researches / case studies, the need of a system which is scalable and efficient at handling malicious URL detection is clear. However most of the papers do not focus on a proper mechanism to sort Malicious and benign URLs according to an event or provide users with such features for ease of use.



*Figure 1.1- Phishing Incidents in 2020 compared to prior years* [2]

## Research Gap

There are many established tools and technologies to determine if the website the clients/ users are trying to access is a fraudulent or phishing page based on the previous data fed through the deep learning models such as keyword extractors.

These researches conducted includes,

* Extracting keywords from webpages
* Monitoring for potential malicious matches
* Identifying and warning on malicious pages
* Storage of identified malicious webpages

If we were to compare the existing solutions [3] [6] [7] and research papers almost all the deep learning /ML systems have been developed on the concept of identifying and categorizing malicious pages by matching keywords related to malicious actions or using a categorization method to identify pages based on set of prior user feedbacks in order to warn the new users of the systems. Particularly in research (A) [6] focuses only on the classification aspect and the accuracy aspect of the data set keywords with the use of a Natural Language Processing (NLP) analysis method, “sentiment analysis” to classify into positive and negative categories given a text set to derive keywords. Paper [7] , research B has concerns set on the detection, feature rich classification and the increasing on detection speed/rate. Lastly, research C [3] focuses on accuracy and the classification of URLs only. As a conclusion there is not a solution which focuses on scalability, accuracy and efficiency along with the user friendliness by providing sorting features for users to select based on a particular event just by entering keywords related.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Paper** | **Scalability** | **Accuracy** | **Classification** | **High rate of detection/ Efficiency** | **Ability to sort through keywords / Events** |
| **Research A** | X |  |  | X | X |
| **Research B** |  | X |  |  | X |
| **Research C** | X |  |  | X | X |
| **Proposed Solution** |  |  |  |  |  |

*Table 1.1- Comparison of proposed solution with former researches*

The ensemble model which has been proposed, is trying to address the common issues of not being able to efficiently filter sets of URLs real-time and with added capacity for researchers and organizational clients to proactively filter malicious URLs based on an event.

## Research Problem

Digital transformation of the world has paved the way for criminals to launch cyber-attacks. World Wide Web has become a major platform for these attacks to be spread out globally. Due to the recent Covid-19 pandemic situation, almost all the traditional organizations had to shift to online mediums for various reasons. This situation became an opportunity for attackers as most of the shifting personnel lack the knowledge about mitigating phishing attacks as well as cyber-attacks in general. [1] Results of this situation was the exponential growth of phishing attack victims and spawn of phishing URLs in the year of 2020. [2]

Cyber criminals always come up with ways to lure the victims by pretending to be a legitimate source as much as possible. Several techniques like use of keywords embedding to the malicious URLs like for an example covid, WHO, vaccine which of all the words an user will be in search of during this pandemic and use of URL obfuscation techniques to mask the malicious URLs to evade the malicious URL detectors using lexical and static feature analysis was really popular recently.

These reasons have raised the need of a system where identifying and classification of potential malicious URLs related to each events as soon as possible to encounter the problem of being a victim of such cyber-attacks. In order to solve this problems , use of Machine Learning [ML] approaches are used but the models which are currently being used do not help in a situation like a massive emerge of URLs in a short span of time due to an event because most models do not incorporate scalability. This prevents the system to scale with respect to the processing power given to the system. This results in delayed identification of malicious URLs. [8] Another issue with event related emerging of malicious URLs is that the users do not have a simplified system to evade such in real time using a simplified system. [3]

# Objectives

## Main Objectives

The main objective of the system is to make use of trends / events to aid in identifying potential malicious websites and pages. Whenever an entity enters a keyword using our proposed system user interface, they will be presented with pre made lists/templates of event related phishing scenarios. User can either use the pre-defined scenarios or enter keywords in order to begin finding related malicious URLs. The system matches the entered keywords with URL lists tokens/ generated keywords and present those data as a list. Using these data either they can blacklist the malicious URLs or use the populated keyword data to research more on a specific event. These mentioned objective can be achieved easily through machine learning models, However as mentioned in the introduction one of the key problems with these systems is the efficiency or the lack of scalability to the keyword populate module. Therefore, we try to address the issues by having the main objective of increasing the efficiency, scalability and the accuracy of such solutions by introducing a new system.

The methodology of detecting and populating keywords for events real-time is explained further below in the document.

## Specific Objectives

In order for reaching the main objectives described, the following specific objectives needs to be attained,

* + 1. Identify NLP strategies to populate related keyword token from the user-input event keyword.

Whenever the user input a keyword using the web interface of the system, the keywords needs to be populated with related keywords.

* + 1. Identifying a model to find related keywords for an event by aggregating with online sources.

In order to find keywords related to the user input words, we need to find datasets to capture sufficient dataset to train the model in order to derive keyword that can be mapped to user inputs, this process can be achieved by collecting training and testing data from Phish Tank and Common Crawlers. After the training phase is completed the prediction phase will utilize the trained model for analysis of URL data.

* + 1. Co-relate the populated keyword tokens with the phishing URLs database

The system need to map the keyword populated from the user input with the keywords / token data from the URLs in order for the identification of the malicious and benign URLs.

* + 1. Creation and incorporation of predefined common event templates in the system for the ease of users

In order to achieve the ease of use and the objective of providing a user interface simplification for the system, we propose to generate common events related to a user inputted keywords. User can use these templates for time efficient searching of related URLs.

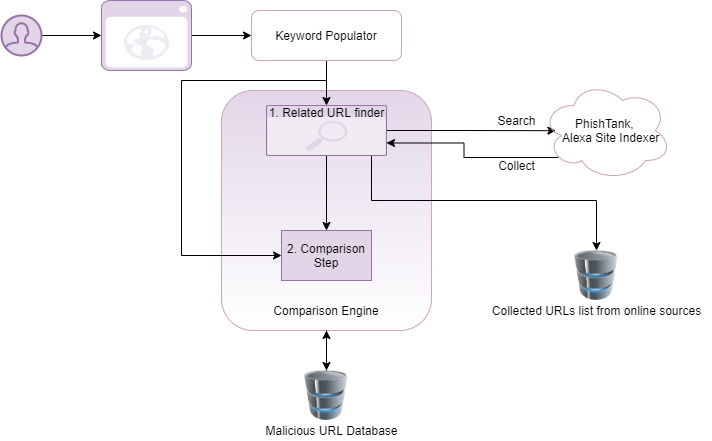
* + 1. Improving the scalability of the system

# Methodology

The proposed Event-Driven Phishing URL Extractor has the capability of,

* Generating keywords related to the user input keywords.
* Generating keywords/ tokens from URLs.
* Identifying and classification of matching URLs with keywords for specific events.
* Determining the matched URL is malicious or benign.

## **3.1** System Architecture



*Figure 3.1 – System Architecture*

## **3.2** Keyword Populator

In order to achieve the objective of populating user input keywords, when a user inputs keywords via the system’s interface those captured keywords will go through the keyword populator model. In the model mainly there are two mechanisms working together to achieve this objective,

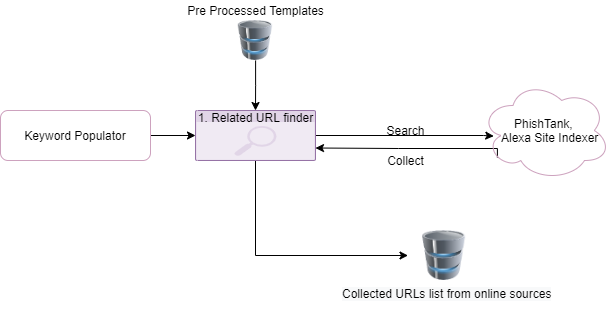
**1.** Use of pre-processed template keywords generated from the pre-process template creator model in the core system to aid the process.

**2.** Use of a Word level RNN (Recurrent Neural Network) to generate search queries related to user inputs.

To train this model, we need a sufficient amount of sample search queries created. In order to create this sample set, we are planning to use WooRank’s API.

## **3.3** Related URLs Finder

Related URL finder is a part of the comparison system. Here, results from the keyword populator will be compared with preprocessed scenarios and event templates in the core system to match the keywords with the events. In accordance with those keywords, a sample set of URLs containing the generated keywords will be collected from online sources like PhishTank and through Alexa’s site indexer, etc. These identified URLs will be categorized with respective keywords and stored in a database for future use/processing as well.



*Figure 3.3 – Related URLs Finder Architecture*

## **3.4** Comparison Engine

After the initial step of identifying related URLs matching with user inputted keywords, in order for the ML model to be able to process these inbound URLs, we need to standardize to extract features. Therefore, the collected URLs will be forwarded to processing from the database. After that these URLs will be initially checked if they were registered with a malicious intent. We plan to use several techniques to identify this,

### 3.4.1 Entropy Calculation

Shannon’s entropy calculation of the domain name is used for detecting URLs where the attackers try obfuscation techniques to appear as legitimate URLs. Basically attackers try to mimic a legitimate URL using simple character changes, spelling mistakes to mask the suspicious URLs. To mitigate this issue, randomness of a URL can be checked by the entropy calculation. Paper [10] shows the malicious URLs have a high entropy value than a benign one on average.

Entropy calculation can be performed by the following equation,

(𝑥) = −∑(𝑥𝑖) log𝑏 𝑝(𝑥𝑖)

Risk Rating Binary Value Calculation

For efficient storage of scanned URLs, we need to assign a binary value and label them accordingly.

* + - 0 – benign
    - 1- malicious

For determining what value to assign to the URLs, we need to take the existing datasets of malicious and benign URLs and train the model predictor.

### 3.4.2 Finding Obfuscated Keyword Variants

### 

Using this part of the model, will try to capture similar patterns of matching keywords in domain name.

For example

User input = Covid19

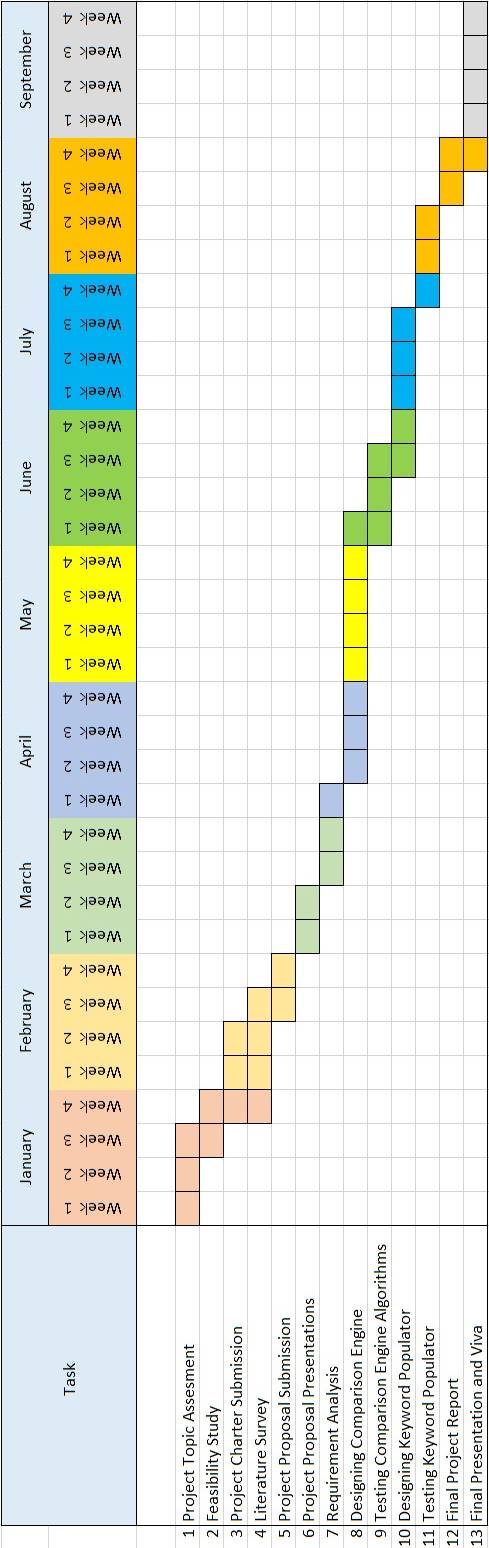
The model generates and match keyword variants like – C0vid19, COv1d

Utilization of this particular technique will be beneficial in delivering accurate results while determining URLs with malicious intents.

If a URL is found to be malicious, it will be directed to the Malicious URL Database and will get stored in the database.

# Gantt chart

The Gantt chart for development of the proposed system is depicted in Figure 5.1.



*Figure 4.1-Gantt Chart*

# 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 5.1-Budget Chart*

# 6 Commercialization

Targeted user base is small-medium enterprises (SME), due to the lower cost and higher usability.

Can be implemented in two versions.

* 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 plus 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 exporting event-based malicious URLs |

# *Table 6.1- Commercialization Versions*

# Reference List

|  |  |
| --- | --- |
| [1] | G. -. G. Technology, "2020: The Year the COVID-19 Crisis Brought a Cyber Pandemic," GT - Government Technology, [Online]. Available: https://www.govtech.com/blogs/lohrmann-on-cybersecurity/2020-the-year-the-covid-19-crisis-brought-a-cyber-pandemic.html. [Accessed 21 02 2021]. |
| [2] | D. Warburton, "2020 Phishing and Fraud Report," F5 Labs, [Online]. Available: https://www.f5.com/labs/articles/threat-intelligence/2020-phishing-and-fraud-report. [Accessed 21 02 2021]. |
| [3] | I. M. M. A. R. A. H. L. N. S. a. A. A. G. Mohammad Saiful, "Detecting Malicious URLs Using Lexical," in *University of New Brunswick, Fredericton, NB, Canada*, Fredericton, 2016. |
| [4] | C. G. J. M. V. P. a. D. S. Kurt Thomas, "Design and evaluation of a real-time url spam filtering service," in *IEEE Symposium on. IEEE*, 2011. |
| [5] | C. L. S. C. H. DOYEN SAHOO, "Malicious URL Detection using Machine Learning: A Survey," 2019. |
| [6] | S. S. Vishal A. Kharde, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Applications (0975 – 8887),* vol. Volume 139, no. 11, p. 15, 2016. |
| [7] | E. B. O. D. B. D. Ozgur Koray Sahingoz, "Machine learning based phishing detection from URLs," *Expert Systems With Applications,* 2019. |
| [8] | M. M. D. M. I. Vahid Shahrivari, "Phishing Detection Using Machine Learning Techniques," in *arXiv:2009.11116v1 [cs.CR] 20 Sep 2020*, 2009. |
| [9] | J. I. &. R. Islam, "Detecting Malicious Urls of COVID-19 Pandemic Using ML Techniques," Charles Sturt University, NSW, Australia. |
| [10] | D. R. &. P. J. B. Patil, "URLs detection using decision tree classifiers and majority voting technique," *Cybernetics and Information Technologies,* vol. 18, no. 1, pp. 11-29, 2018. |

## **Appendix A – Plagiarism Report**



