Sri Lanka Institute of Information Technology

PROJECT REGISTRATION FORM

The purpose of this form is to allow final year students of the B.Sc. (Hon) degree program to enlist in the final year project group. Enlisting in a project entails specifying the project title and the details of four members in the group, the internal supervisor (compulsory), external supervisor (may be from the industry) and indicating a brief description of the project. The description of the project entered on this form will not be considered as the formal project proposal. It should however indicate the scope of the project and provide the main potential outcome.

PROJECT TITLE (As per the accepted topic assessment form)	Event-Driven Phishing URL Extractor
RESEARCH GROUP **(as per the Topic assessment Form)**	Information Security
PROJECT NUMBER	(will be assigned by the lecture in charge)

PROJECT GROUP MEMBER DETAILS: (Please start with group leader's details)

	STUDENT NAME	STUDENT NO.	CONTACT NO.	EMAIL ADDRESS
1	S.W Sajeth Jonathan	IT18071412	0775575495	it18071412@my.sliit.lk
2	Renu Harshatha A.	IT18034400	0762814977	it18034400@my.sliit.lk
3	Wishvajith B.L.D.V	IT18032666	0716209643	it18032666@my.sliit.lk
4	Ramanayake A.M	IT18021912	0763597251	it18021912@my.sliit.lj

SUPERVISOR, CO_ SUPERVISOR Details

SUPERVISOR Name	CO-SUPERVISOR Name
Mr Amila Senarathne	**Mr. Kanishka Yapa**
Signature	**Signature**
Attach the email as Appendix 1	**Attach the email as Appendix 2**
15th February 2021	**15th February 2021**
Date	**Date**

EXTERNAL SUPERVISOR Details (if any, may be from the industry)

				Attach the email as Appendix 3
Name	Affiliation	Contact Address	Contact Numbers	Signature/Date

ACCEPTANCE BY CDAP MEMBER (This part will be filled by the RP team)

Name	Signature	Date

PROJECT DETAILS

Brief Description of your Research Problem: (extract from the topic assessment form)

Cyber-attacks have been leveraging as the world digitally transforms. COVID-19 has affected government, corporate industries, educational organisations, non-profit organisations where they had to shift to online platforms from their traditional on premises system. Organisation started working remotely to continue their daily business routine. This challenge became an opportunity for cyber-attackers to lure sensitive information through Phishing attack, which is done through social engineering. Phishing is also considered as one of the common entry points of a cyber-attack.[1]

Using the pandemic COVID -19 situation, attackers targeted their victims. Their victims are, users who are unaware of Phishing attacks, this unawareness is dangerous for an individual but an opportunity for the attacker. [4] Numerous Phishing URLs are spawned at an instant. [2] Cybercriminals come up with different ways to make the site look legit as possible. One way is making the domain name have words familiar to the user so that they would visit the website. The COVID-19 phishing URLs had keywords like covid-19, WHO, vaccine and so on. These keywords change with the trending words related to COVID-19 as the global environment changes. Similar to this, any event would relate to spawning of malicious URLs to increase the probability of the user trusting the site. Identifying and categorizing URLs related to events as soon as they occur at a local/global scale is crucial for organizations and researchers.

In addition, Machine Learning [ML] approaches are used find malicious URL and gives better solutions. But current models do not incorporate scalability dominantly. This prevents a system to scale according to the processing power given to it. A malicious URL is identified with a delay due to lack of scalability. Additionally, the existing solutions for detecting phishing URLs takes extra time which reduces the speed and efficiency of identifying them. [3]

[1]Hulten, G. J., Rehfuss, P. S., Rounthwaite, R., Goodman, J. T., Seshadrinathan, G., Penta, A. P., Mishra, M., Deyo, R. C., Haber, E. J., & Snelling, D. A. W. et al. Finding phishing sites, 2014, US Patent 8,839,418.
[2] 2020 APWG trend report. https://docs.apwg.org/reports/apwg_trends_report_q3_2020.pdf Accessed from 23rd January 2021
[3] Basit A., Zafar M., Xuan L., Javed A.R., Jalil Z., Kifayat K . A comprehensive survey of AI-enabled phishing attacks detection techniques,2020 doi: 10.1007/s11235-020-00733-2
[4] Fbi warns of dramatic increase in business e-mail compromise (bec) schemes—fbi. https://www.fbi.gov/contactus/field-offices/memphis/news/press-releases/fbi-warns-ofdramatic-increase-in-business-e-mail-compromise-bec-schemes Accessed from 23rd January 2021

Description of the Solution: (extract from the topic assessment form)

To mitigate the problems, we propose a system to filter malicious URLs based on global or local events. This approach will assist researchers and organizations to proactively block malicious URLs based on an event. An event-driven system will take an event in the form of a keyword from the user and generate related keyword tokens by corresponding with trending information through Online sources. Once the keyword tokens are identifies, they can be matched with a dataset of Phishing URLs to categorize them.

With many events happenings in a local/global scale, this system will help users to map the phishing campaign URLs to these events and the output can be used to various purposes. This can be, but not limited to.

- Researchers who are investigating certain events can identify phishing URLs associated with it, to find relationships.
- Organisations can block these event-based phishing URLs in response to any change in their social/technical/political environment.

The core of this system will use an ensemble model using Deep Learning and Machine Learning algorithms to classify URLs. An ensemble model is considered with the purpose of increasing the accuracy of the system. In the Deep Learning space, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) are primarily used so far, to detect phishing URLS. On the other hand, in the Machine Learning space, C4.5, Support Vector Machines (SVM) and K-nearest neighbors' algorithm (k-NNs) are widely used. Also, this research will let us explore the current detection methods using Natural Language Processing (NLP). These models can be compared to find which combination would suit the ensemble model with the goal of increasing efficiency in this event-driven system.

Training and Testing data will be fed to the URL classifier after identifying a suitable feature reduction model to boost efficiency and speed since not using feature reduction was identified as one of the major gaps in this domain. This system will incorporate scalability in its design to facilitate the amount of processing power given to it and also adjust according to the volume of URLs generated by an event.

The system will suggest events for the user to search based on common phishing attack scenarios. This feature will increase the usability of the system and it should be easy to add new templates in the future. With the advancement of URL Phishing Detection methods, there is possibility for a better model or ensemble model to be identified. In a situation like that, the system should be able to facilitate to use such a model.

Main expected outcomes of the project: (extract from the topic assessment form)

Main Objective:

Identify a suitable Ensemble Model using Deep Learning and Machine Learning Algorithms for accurate analysis.

Sub Objective 1: Improving Scalability using proper scalable architecture.

Sub Objective 2: Identify a model to find related keywords for an event by aggregating with online sources.

Sub Objective 3: Matching the identified keyword with malicious URL.

Sub Objective 4: Create and incorporate predefined common event templates in the system for the user.

WORKLOAD ALLOCATION (**extract from the topic assessment form after correcting the suggestions given by the topic assessment panel.**)

(Please provide a brief description about the workload allocation)

MEMBER 1

- Identify a suitable ensemble model by comparing the current algorithms available in DL, ML and NLP domains based on their accuracy and false positive rates.
- Create a suitable system architecture with Python Flask as the front-end web application
- Collect Training and Test Data from PhishTank and Common Crawler to train the ensemble model.

MEMBER 2

- Identify a scalability model that is compatible with Ensemble Model.
- Compare available scalability model. In order to do this a deep analysis on how each model works need to be studied and their algorithms needs to be compared. If the model doesn't satisfy the requirement of the Ensemble Model a hybrid model needs to be created.
- Test the selected model to check its compatibility with Ensemble Model. The scalability model should be able to expand workload when the scope increases and shrink when the scope is small.

MEMBER 3

- Identify NLP strategies to populate related keyword token from the user-input event keyword.
- Co-relate the populated keyword tokens with the phishing URLs database to analyze them according to their scenarios.

MEMBER 4

- Identify a working methodology to incorporate predefined templates on common events.
- Categorization of scenarios for scenario based phishing detection.
- URL feature representation and feature analysis for increasing the speed and efficiency in detecting malicious URLs.
- Work with the system's speed and efficiency - Search about feature reduction methodology and identify how it can improve the Extractor's speed and efficiency.

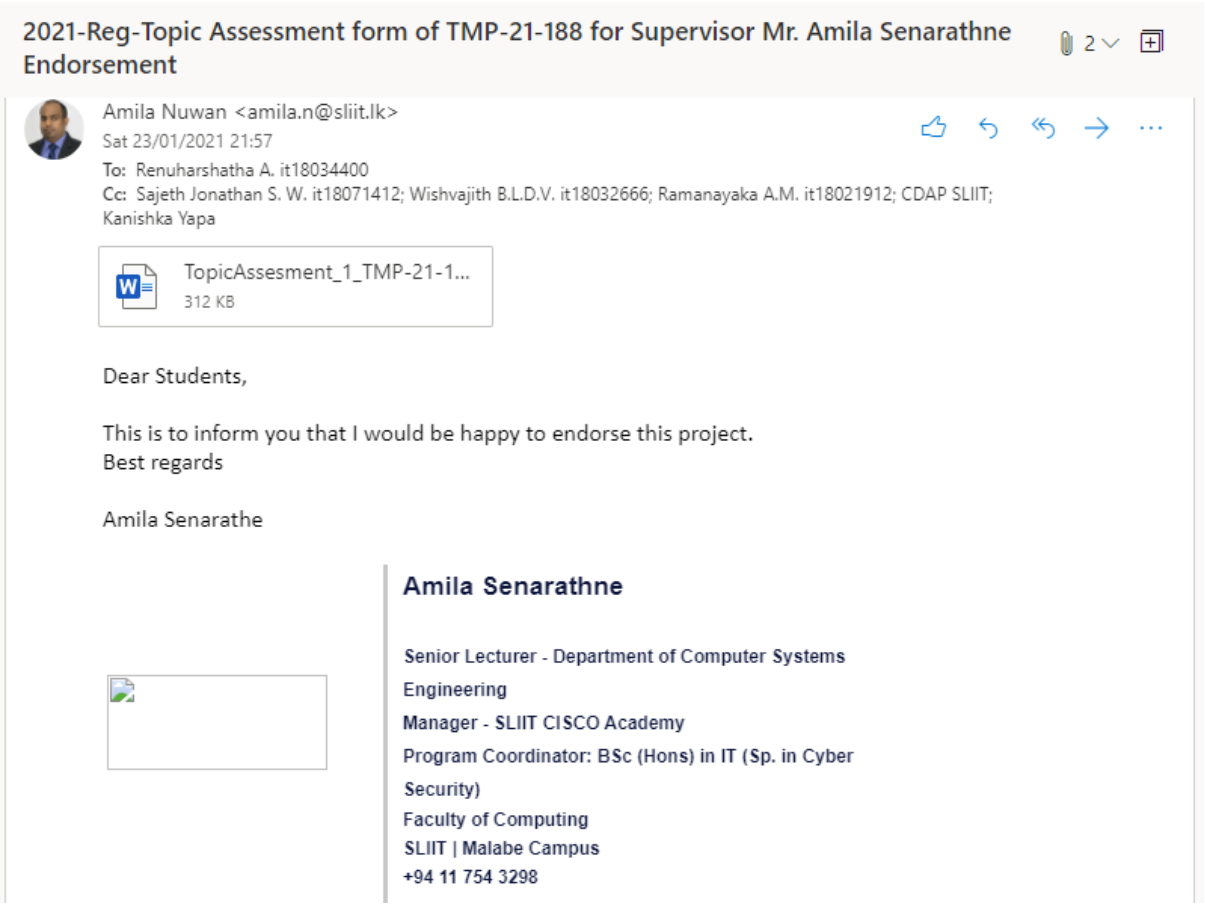
DECLARATION (Students should add the Digital Signature)

"We declare that the project would involve material prepared by the Group members and that it would not fully or partially incorporate any material prepared by other persons for a fee or free of charge or that it would include material previously submitted by a candidate for a Degree or Diploma in any other University or Institute of Higher Learning and that, to the best of our knowledge and belief, it would not incorporate any material previously published or written by another person in relation to another project except with prior written approval from the supervisor and/or the coordinator of such project and that such unauthorized reproductions will construe offences punishable under the SLIIT Regulations.

We are aware, that if we are found guilty for the above mentioned offences or any project related plagiarism, the SLIIT has right to suspend the project at any time and or to suspend us from the examination and or from the Institution for minimum period of one year".

	STUDENT NAME	**STUDENT NO.**	**Signature**
1	S.W Sajeth Jonathan	IT18071412	
2	Renu Harshatha A.	IT18034400	
3	Wishvajith B.L.D.V	IT18032666	
4	Ramanayake A.M	IT18021912	

APPENDIX 1: Supervisor Endorsement



2021-Reg-Topic Assessment form of TMP-21-188 for Supervisor Mr. Amila Senarathne Endorsement

2

Amila Nuwan <amila.n@sliit.lk>
Sat 23/01/2021 21:57
To: Renuharshatha A. it18034400
Cc: Sajeth Jonathan S. W. it18071412; Wishvajith B.L.D.V. it18032666; Ramanayaka A.M. it18021912; CDAP SLIIT; Kanishka Yapa

TopicAssesment_1_TMP-21-1...
312 KB

Dear Students,

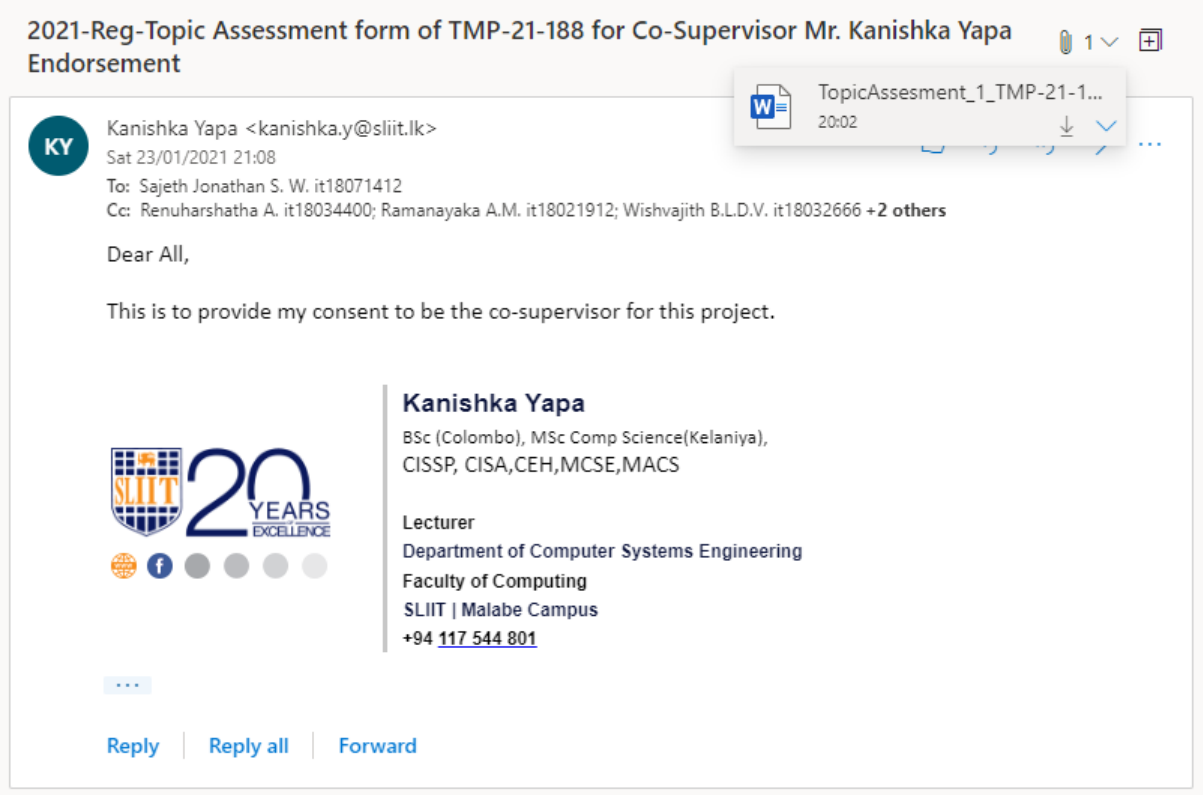
This is to inform you that I would be happy to endorse this project.
Best regards

Amila Senarathe

Amila Senarathne

Senior Lecturer - Department of Computer Systems Engineering
Manager - SLIIT CISCO Academy
Program Coordinator: BSc (Hons) in IT (Sp. in Cyber Security)
Faculty of Computing
SLIIT | Malabe Campus
+94 11 754 3298

APPENDIX 2: Co-Supervisor Endorsement



2021-Reg-Topic Assessment form of TMP-21-188 for Co-Supervisor Mr. Kanishka Yapa Endorsement

1

TopicAssesment_1_TMP-21-1...
20:02

KY

Kanishka Yapa <kanishka.y@sliit.lk>
Sat 23/01/2021 21:08
To: Sajeth Jonathan S. W. it18071412
Cc: Renuharshatha A. it18034400; Ramanayaka A.M. it18021912; Wishvajith B.L.D.V. it18032666 **+2 others**

Dear All,

This is to provide my consent to be the co-supervisor for this project.

Kanishka Yapa
BSc (Colombo), MSc Comp Science(Kelaniya),
CISSP, CISA,CEH,MCSE,MACS

Lecturer
Department of Computer Systems Engineering
Faculty of Computing
SLIIT | Malabe Campus
+94 117 544 801

Reply | Reply all | Forward