DETECTING PHISHING UNIFORM RESOURCE LOCATOR(URL) BY USING MACHINE LEARNING TECHNIQUES

LIM CHIAN FANG

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: DETECTING PHISHING UNIFORM RESOURCE LOCATOR(URL) BY USING MACHINE LEARNING TECHNIQUES

SESI PENGAJIAN: 2020/ 2021

Saya: LIM CHIAN FANG

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan unituk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. * Sila tandakan (✓)

✓	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
✓	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan)
___________	TIDAK TERHAD	

(TANDATANGAN PELAJAR) Lim Chian Fang	(TANDATANGAN PENYELIA)
Alamat tetap: No66, JalanM 27, Taman Merdeka, Batu Berendam,75350 Melaka	Zakiah Binti Ayop
Tarikh: 05/09/2021	Tarikh: 05/09/2021

DETECTING PHISHING UNIFORM RESOURCE LOCATOR(URL) BY USING MACHINE LEARNING TECHNIQUES

LIM CHIAN FANG

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Security) with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2020/2021

DECLARATION

I hereby declare that this project report entitled

DETECTING PHISHING UNIFORM RESOURCE LOCATOR(URL) BY USING MACHINE LEARNING TECHNIQUE

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : ___________LIM CHIAN FANG__________ Date : 05/09/2021

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of Computer Science (Computer Security) with Honours.

SUPERVISOR : ___________ ZAKIAH BINTI AYOP ________ Date : 05/09/2021

DEDICATION

I would like to dedicate my work to my parents, who have always been a source of inspiration and determination for me, and who continue to support me morally, spiritually, and financially.

I also dedicate this dissertation to my brother, sister, mentor, and friends, who have always encouraged and supported me to finish my research.

ACKNOWLEDGEMENTS

First and foremost, I want to express my gratitude to Universiti Teknikal Malaysia Melaka (UTeM) and the Faculty of Information and Communication Technology for providing me with an excellent opportunity to apply what I had learned to develop a project.

In addition, I would like to express my great appreciation and unlimited thanks to my supervisor, Madam Zakiah Binti Ayop for taking part in the useful decision and giving necessary advice and guidance during the planning and implementation of this project. Despite being extremely busy with her duties, she took the time to share her knowledge, guide, and keep me on track throughout the project. I was lucky to get consistent encouragement, support, and advice, which enabled me to complete my project sarjana muda successfully.

I would like want to express my gratitude to my family and friends for their encouragement and support throughout the project's development. Hence, I would also like to thank everyone who involves directly or indirectly in the project. Without the support of the persons mentioned above, I may not have completed this project successfully.

ABSTRACT

As the internet has grown in popularity, phishing websites have become more common and caused significant harm to online financial services such as online shopping and data security. Phishing is a type of fraud whereby an attacker sends a fake message or creates a phishing website to mislead web users into sharing confidential information or allowing malicious software to be installed on the victim's device. Many attackers started creating phishing websites to misled web users into thinking it’s legitimate. So, web users may be exposed to common web attacks, which might result in the loss of money, personal information, and trust from online transactions. Hence, detecting phishing websites has become a critical task that requires more examination. The most commonly used blacklist- and whitelist-based methods have shown to be ineffective. Researchers have looked into using machine learning models to detect and prevent phishing attempts. The accuracy of the prediction can be increased using machine learning methods. CatBoost based URL classifiers for detecting phishing websites are proposed in this project. The first stage is dataset will be split to 80:20 ratio to train machine learning model. The second stage involves the comparison of 3 machine learning algorithms (Logistic Regression, Random Forest, and CatBoost) the third stage involves classification of the URL's legitimacy by using CatBoost. As a result, the URL will be classified as either a phishing or a legitimate URL.

ABSTRAK

Dengan pertumbuhan internet, pancingan data melalui laman web menjadi peristiwa yang biasa dan membawa kesan negatif kepada perkhidmatan kewangan dalam talian seperti membeli-belah dalam talian dan keselamatan data. Phishing adalah satu bentuk penipuan di mana penggodam menghantar mesej palsu atau membuat laman web untuk mengelirukan pengguna web agar berkongsi maklumat sulit atau membenarkan perisian jahat dipasang pada peranti mangsa. Banyak penggodam mula membuat laman web palsu untuk mengelirukan pengguna web sehingga pengguna web menganggap laman web sah. Oleh itu, pengguna web mungkin terdedah kepada serangan web biasa, yang mungkin mengakibatkan kehilangan wang, maklumat peribadi, dan kepercayaan dari transaksi dalam talian. Oleh itu, mengesan laman web pancingan data telah menjadi tugas penting yang memerlukan lebih banyak pemeriksaan. Kaedah berdasarkan senarai hitam dan senarai putih yang paling biasa terbukti tidak berkesan. Penyelidik mengaji kaedah pembelajaran mesin untuk meramalkan dan mencegah pancingan data. Ketepatan ramalan dapat ditingkatkan dengan menggunakan kaedah pembelajaran mesin. Pengasingan URL berasaskan CatBoost untuk mengesan laman web pancingan data dicadangkan dalam projek ini. Langkah pertama adalah set data akan berpecah kepada 80:20 nisbah untuk melatih model pembelajaran mesin. Tahap kedua melibatkan perbandingan 3 algoritma pembelajaran mesin (Logistic Regression, Random Forest, dan CatBoost) tahap ketiga melibatkan klasifikasi kesahan URL dengan menggunakan CatBoost. Akibatnya, URL akan diklasifikasikan sebagai phishing atau URL yang sah.

TABLE OF CONTENTS

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

LIST OF ABBREVIATIONS

AB	**AdaBoost**
CB	**CatBoost**
DS	**Decision Stump**
DT	**Decision Tree**
ELM	**Extreme Learning Machine**
HS	**Harmony Search**
HTTPS	**Hyper Text Transfer Protocol with Secure Socket Layer**
K*	**K Star**
KNN	**K-Nearest Neighbors**
LM	**Linear Model**
LR	**Logistic Regression**
NB	**Naïve Bayes**
NN	**Neural Network**
NR	**Nonlinear Regression**
RF	**Random Forest**
RT	**Random Tree**
SVM	**Support Vector Machine**
TWSVM	**Twin Support Vector Machine**
URL	**Uniform Resource Locator**
UTeM	**Universiti Teknikal Malaysia Melaka**
XGB	**XGBoost**

LIST OF ATTACHMENTS

PAGE

Chapter 1: INTRODUCTION

1.1 Introduction

Phishing is a type of web-based attack where attackers collect confidential data such as ID and passwords by sending a message that looks like it came from a trusted source, organization, or individual in other ways. The target of a phishing attack frequently receives e-mails that appear to be from a legitimate organization. The email usually includes web links that guide targets to the phishing web page to fool them into revealing personal or financial information such as ID, password, and card information. There are several reasons that people fall for phishing. First, a user unfamiliar with Uniform Resource Locators (URL) and URL usage. Second, users unable to differentiate between legitimate URLs and phishing URLs. Third, users unintentionally click on URLs or do not enough time to consult the URL. Forth, the user unable to access the target URL due to redirection or secret URLs. Last, users have no idea which of the URLs displayed can be trusted (Buber et al., 2017). Since phishing webpages focusing on banks, companies and web users are unavoidable, detecting web phishing attacks is critical. Because of various advanced techniques used by attackers to confuse web users, detecting a phishing website becomes difficult. To identify phishing webpages, several traditional strategies focused on set black and whitelisting databases. These methods, however, are ineffective since a new website can be created in a matter of seconds. As a result, most of these strategies are unable to determine if a new website is phishing or not in real-time (Ali, 2017). Machine learning is a multidisciplinary technique of learning that is mainly used in supervised learning to construct predictive models. Machine learning is suitable for detecting phishing webpage because machine learning can transform the problem into a typical classification task. Machine learning can create models based on previously labeled websites, which can then be incorporated into a browser to detect phishing attempts. The dataset that contains website features, as well as the availability of enough websites to build realistic predictive models, are critical to build an automated anti-phishing machine learning model (Abdelhamid et al., 2017).

1.2 Project Background

In recent years, the development of various websites including online banking, education, and social media has been driven due to the growth of the internet. Phishing attacks have increased significantly and are now widely regarded as a most serious new internet crime, potentially causing people to not be trusted in e-business. As a result, phishing has adverse effects on internet banking, e-business, organizational revenues, client partnerships, and overall market operations. The development of various phishing websites enables hackers to access confidential personal or financial data. Phishing URLs are used to collect password and login information, as well as other account information, by sending attackers to target users via e-mail or other communication networks as a recognized individual or entity. Phishing URLs host unsolicited and trick users and become scam victims and result in losses (Yi et al., 2018). Phishing URLs are designed to look legitimate to successfully confuse the user. Since humans are so easily duped, automated techniques to identify phishing websites and legitimate are required as a second of defense (Kulkarni et al., 2019). Phishing URL detection has traditionally been done primarily through the use of blacklists. Blacklists, on the other hand, aren't exhaustive and can't detect newly created malicious URLs. Machine learning algorithms have gained popularity in recent years as a way to improve the generality of malicious URL detectors. (Sahoo et al., 2019). This project aims to build an URL classifier that can classify URLs as malicious or legitimate based on the most accurate algorithm after comparison. Three machine learning algorithms which are Logistic Regression, Random Forest, and Catboost are evaluated and compared in terms of accuracy. This research introduces a phishing URL detection machine learning algorithm based on CatBoost. Machine learning algorithms for evaluating different features of URLs can help humans to distinguish between legitimate and phishing websites with high accuracy.

1.3 Problem Statement

Phishing attack becomes a threat to web users, governments, and companies. In a phishing attack, the attackers use spam email or fake websites to obtain the client's sensitive data such as user account login information, credit/debit card numbers, and passwords. Attackers are also able to create web pages that look and feel like legitimate websites, such as banking, to trick victims into providing confidential information. Even though phishing attacks do not necessitate specialized technical expertise and those users are becoming more aware of these attack tactics, they continue to cause significant financial losses (Basit et al., 2020).

- The traditional blacklist approach will never be completed, resulting in a false positive.

In traditionally, the method to detect phishing websites is updating blacklisted URLs to the database. As a result, it can take hours or even months to be added to a blacklist, giving scammers sufficient time to target several users. However, the blacklisted method can never complete since malicious URLs are created regularly. Genuine websites are often blacklisted inadvertently, whether purposely or not, resulting in a false positive. Given that a single website can be blocked through multiple browsers, such a situation can cause as much trouble for the parties as a good fraud (Silva et al., 2019). To solve these issues, a machine-learning model based on training datasets of previous phishing sites is used to categorize new phishing sites. Security researchers focus on machine learning techniques which consist of algorithms that require data to decide on future data. Machine learning algorithms can study numerous blacklists and legal URLs and characteristics to correctly identify phishing URLs.

1.4 Project Question

Project question are written based on the problem statement above, as shown in Table 1.1.

Table 1.1: Project Question

No.	Project Question
1	Which machine learning algorithms are used for phishing detection with higher accuracy?

1.5 Objective

Three objectives are written based on the project question above, as shown in Table 1.2.

Table 1.2: Objectives

No.	Objectives
1	To investigate the suitable machine learning algorithms to detect phishing URLs.
2	To compare the accuracy of machine learning algorithms such as Logistic Regression, Random Forest, and CatBoost.
3	To implement a URL classifier based on the most accurate algorithm that can detect between phishing and legitimate URLs.

1.6 Scope

Scope of the Project:

- Understand the characteristic of phishing websites and distinguishing features from legitimate websites.

- Determine dataset for designing machine-learning-based approaches.

1.7 Project Contribution

In the research domain, a comparison among CatBoost, Random Forest, and Logistic Regression algorithms is conducted in this project to determine which algorithm has the best performance in terms of accuracy, precision, recall, and f1-score. In addition, this project would benefit web users by allowing them to identify and stay alert to phishing websites in real-time, resulting in a more secure network experience. It can also be used in the security domain whereby cybersecurity authorities can apply it to prevent users from visiting these phishing websites and develop powerful security mechanisms that can identify and avoid phishing domains from reaching the user.

1.8 Report Organization

Report Organization will explain the summary of each chapter. This report contains six chapters.

1.8.1 Chapter 1: Introduction

Chapter 1 explains detection of phishing URLs by using machine learning. Project background, problem statement, project question, objective, scope, project contribution and report organization are included in this chapter.

1.8.2 Chapter 2: Literature Review

Chapter 2 discusses and reviews several journals, articles and books regarding the detection of phishing URL by using machine learning. This chapter included related work or previous work, critical review of existing algorithms or techniques, and project solution.

1.8.3 Chapter 3: Project Methodology

Chapter 3 explains the methodology and process to complete this project. Research milestones, Gantt chart are included in this chapter to make sure that this project is completed on time.

1.8.4 Chapter 4: Implementation

Chapter 4 highlights the development of detecting phishing URLs by using machine learning such as coding. The project's expected outcome will also include in this chapter.

1.8.5 Chapter 5: Testing & Analysis

Chapter 5 examines and evaluates the results of detecting phishing URLs by using machine learning. In this chapter, a few tests and analyses will be carried out, and the results will be clarified.

1.8.6 Chapter 6: Project Conclusion

Research summarization, research contribution, and research limitation will be addressed in this chapter. At the same time, this chapter will discuss future works.

1.9 Conclusion

In conclusion, detecting phishing URLs by using machine learning can help people and companies identify phishing URLs and avoid information leakage even financial losses. As a consequence, this study expects to produce a detection tool that can help people in preventing any criminal acts from occurring. In the next chapter, the literature review will discuss the papers that have been studied and analyzed regarding the detection of phishing URLs by using machine learning.