# 

**Simple Phishing URL Detection Using Machine Learning and NLP Technology**

MD. Shariful Islam Bhuyan

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)



Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2023

Simple Phishing URL Detection Using Machine Learning and NLP Technology

MD. Shariful Islam Bhuyan

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof Dr. UtpalKanti Das**Chairman & Professor   
Dept. of Computer Science and Engineering  
IUBAT – International University of Business Agriculture and Technology

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Dr. Hasibur Rashid Chayon**Coordinator and Associated Professor  
Dept. of Computer Science and Engineering  
IUBAT – International University of Business Agriculture and Technology

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Aminun Nahar**Assistant Professor  
Dept. of Computer Science and Engineering  
IUBAT – International University of Business Agriculture and Technology

Spring 2023

## **LETTER OF TRANSMITTAL**

11th January, 2023

The Chairman

Thesis Defense Committee

College of Engineering and Technology - CEAT

IUBAT- International University of Business Agriculture and Technology 4

Embankment Drive Road, Sector- 10, Uttara Model Town

Dhaka-1230, Bangladesh

**Subject: Letter of Transmittal.**

Sir,

With due respect, we would like to inform you that it is a great pleasure and a great pleasure for us to submit this report entitled **“**Simple Phishing URL Detection Using Machine Learning and NLP Technology**”** to complete my Practicum course.

It was a great opportunity for us to work on this project to make my theoretical knowledge more realistic. We are now look forward to your kind commentary on this performance report.

We will always be very grateful to you if you kindly go through this report and check our performance.

Thanking you,

*\_\_\_\_\_\_*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MD. Shariful Islam Bhuyan

Student ID: 20103209

## **STUDENT’S DECLARATION**

I am MD. Shariful Islam Bhuyan, student of the BCSE-Bachelor of Computer Science and Engineering program, under the College of Engineering and Technology (CEAT) of the International University of Business Agriculture and Technology (IUBAT) announcing, this report entitled. ‘**Simple Phishing URL Detection Using Machine Learning and NLP Technology**‘ has been prepared for the completion of the CSC 488 course, which is part of the Bachelor of Computer Science and engineering degree.

The report and the project **"**Simple Phishing URL Detection Using Machine Learning and NLP Technology **"** was edited by me. All modules and procedures for this project are done after proper testing and online information.

It is not designed for other purposes, awards or presentations.

\_\_\_\_\_\_\_\_\_\_\_\_\_

MD. Shariful Islam Bhuyan

Student ID: 20103209

## **ACKNOWLEDGMENTS**

In the name of the Most Gracious and Most Merciful All-Powerful, Amen. We would like to express our gratitude to the late Professor Dr. Md. Alimullah Miyan, Vice-Chancellor of IUBAT— International University of Business, Agriculture, and Technology, who granted us ‘permission to enroll in this university—the most stunning and well-known non-government institution in this nation. We would like to express our gratitude and respect to Professor Dr. Abdur Rab, our current honorary a ‘chancellor at the International University of Business, Agriculture, and Technology. We are grateful to Prof. Dr. Utpal Kanti Das, Chairman and Professor, Department of Computer Science and Engineering, International University of Business Agriculture and Technology, for granting us permission to study in the Department of Computer Science and Engineering and enabling us to see the promising future in the new era of technology. I want to express my gratitude to Dr. Hasibur Rashid Chayon, the departments respected coordinator and associate professor, for providing improved guidance and support during the semester. We are extremely happy and proud to express our gratitude and deep respect to our esteemed faculty members, especially Aminun Nahar ma'am, Assistant Professor of Department of Computer Science and Engineering, IUBAT, for her academic guidance, supportive, and unwavering efforts to carry out this report. Finally, we would want to thank our parents and instructor for being such an example to us. Without their assistance, we would not have reached our current position

## **SUPERVISOR’S CERTIFICATION**

This is to ensure that the Thesis report on the **“Simple Phishing URL Detection Using Machine Learning and NLP Technology”** is compiled by MD. Shariful Islam Bhuyan, with ID # 20103209 of IUBAT– International University of Business Agriculture and Technology. The report has been prepared under my supervision and is a record of the work accomplished, successfully completed. To the best of our knowledge and as per her declaration, no portions of this report have been posted anywhere by any degree, diploma or certificate.

You are now allowed to submit a report. I wish him every success in their future endeavors.

Thesis Supervisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Aminun Nahar

Assistant Professor

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

## 

## DEPARTMENT’S CERTIFICATION

On behalf of the Department of Computer Science and Engineering, IUBAT-International University of Business Agriculture and Technology, I undersigned, confirm the performance report ‘Simple Phishing URL Detection Using Machine Learning and NLP Technology’ for Bachelor of Computer Science and Engineering (BCSE) degrees was duly presented by MD. Shariful Islam Bhuyan (ID No. 20103209) and approved by the department.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Aminun Nahar

Supervisor, Assistant professor

Department of Computer Science and Engineering

IUBAT- International University of Business Agriculture and Technology

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Hasibur Rashid Chayon

Coordinator and Associated Professor

Department of Computer Science and Engineering

IUBAT- International University of Business Agriculture and Technology

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. UtpalKantiDas

Chairman & Professor

Department of Computer Science and Engineering

IUBAT- International University of Business Agriculture and Technology

## **ABSTRACT**

Phishing is the most common and simplest type of cyberattack in the cyberspace. The phisher manipulates the user’s mind by sending phishing URL, which is much closer to the trusted site. Every year phishing causing a huge loss of internet users data and privacy. For security experts it always been a challenging task to detect a phishing URL and they are looking for a trustworthy and steady detection approach. This paper deals with machine learning algorithm and NLP technology for detecting phishing URL by extracting the key features of malicious and benign URL. The machine learning algorithms like KNN, Decision tree, MultinomialNB and Logistic regression is used for comparing the performance. The best performed model “Logistic regression” deployed for building the model. To make the model more beneficial for users the model is also deploy on to “Fast Api” for user interaction.

**TABLE OF CONTENTS**

Letter of Transmittal…………………………………………………………………………….III

Student’s Declaration……………………………………………………………………………IV

Supervisor Certification…………………………………………………………………………VI

Acknowledgement………………………………………………………………………………..V

Abstract………………………………………………………………………………………...VIII

List of Figures…………………………………………………………………………………….X

List of Tables……………………………………………………………………………………...X

Chapter I Introduction…………………………………………………………………………1-3

* 1. Problem Statement ..............................................................................1
  2. Existing Approaches…………………………………………………1

1.3 Objective of the Study ........................................................................2

1.4 The Resolving Technology..................................................................2

1.5 Research Question...............................................................................3

1.6 Hypothesis...........................................................................................3

Chapter II. Literature Review ......................................................................................................4-5

Chapter III. Research Methodology............................................................................................6-10

3.1 Research Methodology........................................................................6

3.2 Data Overview ....................................................................................6

3.3 Tokenization ........................................................................................7

3.4 Find Root words……………………………………………………7-8

3.5 Count Vectorizer……………………………………………………..8

3.6 Machine Learning Algorithm……………………………………..9-10

3.7 Work Flow………………………………………………………….10

Chapter IV. Result Discussion………………………………………………………………..11-14

4.1 Confusion Matrix…………………..……………………………11-13

4.2 Performance Comparison…………………………………………..14

Chapter V. Conclusion………………………………………………………………………….15

List of Figures

3.1.1 Proposed Methodology…………………………………………………………………..6

3.2.1 Dataset overview…………………………………………………………………………7

3.3.1 Tokenization………………………………………………………………………………8

3.4.1 Root words finding……………………………………………………………………….9

3.7.1 Work Flow diagram………………………………………………………………………10

4.1.1 Confusion matrix for KNN………………………………………………………………..11

4.1.2 Confusion matrix for MultinomialNB…………………………………………………….12

4.1.3 confusion matrix for Decision tree………………………………………………………..13

4.1.4: Confusion Matrix for LR…………………………………………………………………13

**List of Tables**

4.2.1: Comparison Table…………………………………………………………………………14