CSEC 620 CYBER ANALYTICS AND MACHINE LEARNING

PROJECT 3: MACHINE LEARNING FOR SECURITY - PHISHING

November 19, 2024

Nikhil S. Patil
Kamron J. Cole
Graham G. Rogozinski
Patrick T. Elser
Department of Cybersecurity
College of Computing and Information Sciences
Rochester Institute of Technology
nsp4746@rit.edu
kjc8084@rit.edu
ggr4460@rit.edu
pte6144@rit.edu

I. Abstract

This section provides on overview about the project. It should be completed at the very last stage of the writing, i.e., after you have completed all the other sections.

II. Introduction

In today's digital world, phishing remains one of the most prevalent cybersecurity threats, targeting individuals, organizations, and critical infrastructures. Phishing attacks employ deceptive tactics, such as fraudulent emails or malicious links, to manipulate victims into disclosing sensitive information or performing unauthorized actions. Consequences of these attacks can be devastating, regardless of victim. It ranges from financial loss to data breaches and reputational damage. The increasing sophistication of phishing techniques necessitates the development of robust detection mechanisms. Traditional approaches, such as rule-based filtering and signature matching, often struggle to keep pace with the evolving nature of these attacks. As a result, researchers and practitioners have turned to machine learning (ML) and deep learning (DL) techniques to enhance the detection capabilities against phishing attempts. This research explores the development of a neural network-based model for detecting phishing data. This model also includes other approaches to the dataset to evaluate the performance of different approaches. Leveraging a dataset of 111 features extracted from emails and links, the model aims to distinguish phishing attempts from legitimate communications with high accuracy. This paper provides an overview of phishing threats, discusses related work in phishing detection using machine learning, and details the proposed neural network architecture, its implementation,and performance evaluation. By advancing the application of deep learning in cybersecurity,this research seeks to contribute to the growing body of work aimed at mitigating phishing threats and securing the digital ecosystem.

III. Literature Review

- How, when the problem was raised by whom, what event, etc.
- Any attempts have been proposed to attack the problems
- Pros and Cons of each existing solution
- How does your proposed solution fit within here, i.e., your solution is improving an existing solution, just another solution, or revolutionizing the way of thinking?

Make sure to cite references. Here is how to cite the great textbook written by Knuth [1] in LaTeX. Here is the way to cite a journal article written by the father of computer science Alan Turing [2].

IV. Project Description

Phishing due to the introduction of large language models, has become more and more difficult to detect. Artifical Intelligence models such as OpenAI's ChatGPT, Google's Gemini, Meta's LLaMA can write highly detailed and highly personalized emails to phishing targets. As a result the necessity of a detection engine that is able to detect phishing emails has increased significantly. We have recognized this, and the goal of this project is to produce a model that is capable of detecting phishing emails with a high accuracy. In order to showcase the ability of machine learning in detection, we have also used K-nearest neighbors and Random Forest classification in order to see how different algorithms work on this dataset. The dataset was preassembled and prelabeled, and has many features. It examines the URL's parameters in the dataset, as well as the email headers to determine the legitmacy of the origin.

V. Project Implementation

In this section, describe details of your project design and implementation, document challenges and obstacles, and how you overcome them in design and implementation.

VI. Testing and Experiments

In this section, describe your testing and experiment design and setup, and conduct the testing and experiments, and generate data.

Need to explain why your experiment design will do what is supposed to do, and describe the expected result, and how the result may validate your ideas and/or support your project.

VII. Data Analysis

Based on the data generated from the testing and experiments, we derived the following results. The global temperature is increasing at an alarming rate as illustrated in Figure 4. It is not real data, it is for demonstration only.

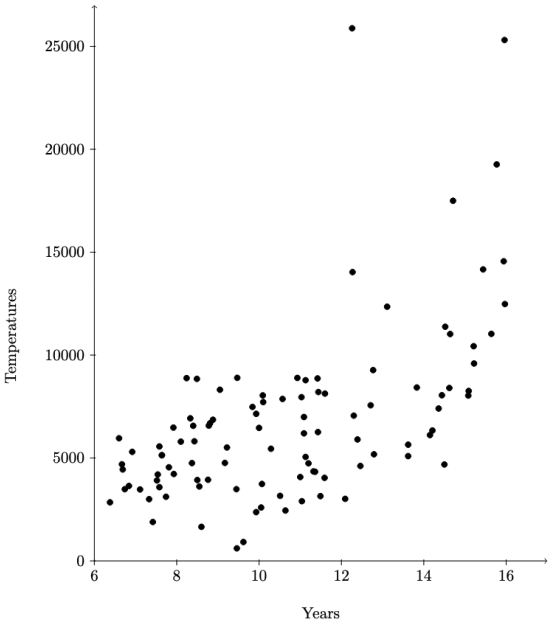

Figure 1. Recent Global Temperature Data

VIII. Conclusions

In this section, you should provide a concise summary of your work, and state the importance of your idea and your contribution towards solving the problem. Point out any improvement that could be done if you had more time. List some ideas for future work.

IX. Acknowledgment

I thank my advisor for the guidance and I appreciate the financial and emotion support from my family and sponsors, ...

References

[1] D. E. Knuth, *Art of Computer Programming, Volume 4, Fascicle 4, The: Generating All Trees–History of Combinatorial Generation*. Addison-Wesley Professional, 2006.

[2] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[3] S. Smallen, "Pseudo code." [Online]. Available: http://users.sdsc.edu/ ssmallen/latex/pseudocode.html

[4] K. M. Fauske, "Example: Simple flow chart." [Online]. Available: http://www.texample.net/tikz/examples/simple-flow-chart/

X. Appendices

This includes all source code, scripts, and any material that helps other people replicate the results.